universität wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Nutzen und Wert von Replikationen in der gegenwärtigen psychologischen Forschung: Theorie und Praxis des Reproducibility Project Psychology am Beispiel von Albarracín et al. (2008, JPSP)“

verfasst von / submitted by

Carina Sonnleitner

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, 2015 / Vienna, 2015

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 298
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Psychologie
Betreut von / Supervisor:	Assoz. Prof. DDDr. Martin Voracek

Danksagung

Größter Dank gebührt Assoz. Prof. DDDr. Martin Voracek. Er stand mir stets mit wertschätzender und kompetenter Betreuung zur Seite. Ihm verdanke ich die Möglichkeit, über ein besonders faszinierendes Thema zu schreiben.

Bei meinen Eltern möchte ich mich aus tiefstem Herzen bedanken – ohne eure Unterstützung und euren Ansporn wäre mir dieses Studium nicht möglich gewesen.

Meiner Kollegin Agnieszka Slowik danke ich herzlich für den regen Austausch sowie die gegenseitige Unterstützung im Entstehungsprozess dieser Arbeit und darüber hinaus.

Besonderer Dank gilt zudem allen Kolleginnen und Kollegen der Open Science Collaboration, welche die Entstehung des Reproducibility Project Psychology ermöglicht haben – insbesondere Mallory Kidwell und Johanna Cohoon danke ich für ihren durchgehenden und fachkundigen Einsatz.

Zum Schluss, aber nicht zuletzt danke ich Herrn Dipl.-Ing. Mag. Andreas Gartus für seinen wertvollen Rat und seine Unterstützung.

Inhaltsverzeichnis

1 Einleitung

Hypothesen gelten als Anbruch der wissenschaftlichen Forschung schlechthin. Begonnen bei den allerersten Fragen nach dem „Wie“ einer Sache, entwickelte sich zunehmend eine enge Umschreibung dafür, was unter einer Alltagshypothese zu verstehen, und welche Art von Hypothese der Wissenschaft zuzuschreiben sei. Demnach hat sich eine Reihe von Bedingungen hervorgetan, welche wissenschaftliche Hypothesen, die den meisten der heutigen Forschungsarbeiten zugrunde liegen, genauer charakterisieren. Als conditio sine qua non einer jeden wissenschaftlichen Hypothese gelten die Formulierbarkeit eines konditionellen Satzes, ihre Generalisierbarkeit, die Möglichkeit ihrer Widerlegung oder eben auch, nach geglückter experimenteller Durchführung, deren Wiederholbarkeit. Gerade diese Option der Wiederholbarkeit (= Replizierbarkeit) macht es uns möglich, die durch experimentelles Vorgehen sowie Beobachtung erlangte Erkenntnis zu Wissen zu aggregieren (Vaux, Fidler, & Cumming, 2012). Lange Zeit unterlag man der Annahme, der Wissenschaft zugrunde liegenden Untersuchungen würden sich auf kurz oder lang einer Selbstkorrektur unterziehen; eine Unterstellung, die sich leider nicht bewahrheitet hat (Ioannidis, 2012; Pashler & Harris, 2012; McBee & Matthews, 2014). So scheint es sich bei der Publikation von Replikationsstudien um eine Randerscheinung zu handeln, betrachtet man entsprechende Untersuchungen, z. B. im Bereich der Psychologie: Makel, Plucker und Hegarty (2012) beziffern den Anteil an Replikationen in psychologischen Fachzeitschriften mit 1.07%. Doch nicht nur die Psychologie, auch andere Disziplinen der Wissenschaft scheinen vom mangelhaften Willen der Replikation ihrer Studienergebnisse ergriffen zu sein. Die Pädagogik etwa offenbart mit 0.13% eine ähnlich niedrige Rate an publizierten Replikationen (Makel & Plucker, 2014). Diese Zahlen für sich genommen sagen selbstverständlich wenig über die tatsächliche Rate an „erfolgreichen“, d. h. die untersuchten Ergebnisse bestätigenden Replikationen aus. Jedoch scheint sich auch hier ein trübes Bild zu zeichnen. Man beachte die vielen fehlgeschlagenen Versuche einen jener Effekte zu bestätigen, der seit seiner Entdeckung als einer der gewichtigsten Bestandteile kognitionspsychologischer Forschung gilt: Priming (Spruyt, Hermans, Pandelaere, De Houwer, & Eelen, 2004; Doyen, Klein, Pichon, & Cleeremans, 2012; McCarthy, 2014). Ein weiteres Beispiel stellt die aufsehenerregende Studie von Bem (2011) dar, welche die Existenz übernatürlicher Kräfte, genauer der Präkognition, bestätigt sehen will. Etliche Forschende wollten, ob der schieren Unglaubwürdigkeit eines solchen Effekts, diesen repliziert wissen; erwartungsgemäß gelang das nicht (Yong, 2012). Besorgniserregend wiederum war die Tatsache, dass jene Forschende, welche Replikationen von Bem’s Studie durchgeführt hatten, diese Ergebnisse nur unter großen Schwierigkeiten publizieren konnten. Das Zersplittern eines der bedeutendsten Effekte der Psychologie, die zunehmende Kritik an

fragwürdigen aber gängigen Forschungspraktiken, sog. Questionable Research Practices (John, Loewenstein, & Prelec, 2012), sowie eine Reihe von Betrugsfällen – allen voran jener des niederländischen Sozialpsychologen Diederik Stapel (Callaway, 2011), infolge dessen es zum Widerruf von bis dato 55 Studienergebnissen kam (Palus, 2015) – leiteten ein, was unter dem Namen crisis of confidence Verbreitung in der psychologischen Forschung gefunden hat (Pashler & Wagenmakers, 2012). Neben nicht-replizierten Effekten der Psychologie wie auch anderer Disziplinen – ein groß angelegtes Replikationsprojekt im Bereich der Onkologie berichtet von 6 bestätigten aus insg. 53 Studien (11%) zur Medikamentenentwicklung (Begley & Ellis, 2012) – gibt es auch jene Studien, die ein optimistischeres Licht auf die Wiederholbarkeit von Ergebnissen werfen. Das Many Labs-Projekt bspw. berichtet 11 von 13 erfolgreich replizierte Effekte (Klein et al., 2014). Wie steht es nun um die Replizierbarkeit psychologischer Forschungsergebnisse im Ganzen? Diese Frage versucht die Open Science Collaboration (2015) zu beantworten: Das von Brian Nosek ins Leben gerufene Reproducibility Project Psychology nahm es sich zur Aufgabe über 100 Studien aus drei prominenten psychologischen Fachzeitschriften zu replizieren.

Ziel dieser Arbeit ist es, einen Überblick zum Thema Replikation unter Beachtung fragwürdiger Forschungspraktiken, der Nullhypothesen-Signifikanztestung, statistischer Teststärke und des Publikationsbias anzubieten, wobei die Schilderung des Reproducibility Project Psychology (Open Science Collaboration, 2015) sowie dessen exemplarischer Darstellung anhand der Replikation von Experiment 7 aus Albarracín et al. (2008) besonderem Fokus unterliegt. Zusätzlich werden Maßnahmen und Methoden vorgestellt, welche die Replikation von Forschungsergebnissen erleichtern sollen.

1.1 Replikation

Bereits bei der Definition wissenschaftlicher Hypothesen selbst kommt man um den Begriff der Replizierbarkeit nicht umher; es wird hier neben den Voraussetzungen der Widerlegbarkeit, der Formulierbarkeit eines Konditionalsatzes sowie der Generalisierbarkeit auch auf die Wiederholbarkeit, d. h. Replizierbarkeit des Hypotheseninhaltes verwiesen. Bei einer Replikation handelt sich somit schlicht gesagt um das Wiederholen von mit Hilfe wissenschaftlicher Forschung erlangten Ergebnissen (Sherman & Wood, 2014). Darüber hinaus lassen sich auch detailliertere Unterscheidungen des Begriffes vollziehen. So wird bei Schmidt (2009) etwa zwischen *exakter*, *direkter* und *konzeptueller* Replikation unterschieden; direkt meint hierbei das möglichst genaue Wiederholen einer Studie, d. h. inklusive Methoden, Ablauf, Stichprobe als auch Setting. Konzeptuell hingegen beschreibt das Überprüfen der Hypothese mit anderen Methoden und in einem anderen Setting; die Rahmenbedingungen einer Studie werden erweitert. Gegen exakte Replikationen sprechen sich sowohl Schmidt (2009) als auch Murray Lindsay und Ehrenberg (1993) aus. Zum einen

würden exakte Replikationen per definitionem ohnehin nicht existieren, da beispielsweise niemals ein Experiment zur exakt gleichen Zeit durchgeführt werden könne oder Testpersonen, sollte es sich auch um dieselben Personen handeln, nie mit den genau gleichen individuellen Voraussetzungen teilnehmen können würden. Neben deren praktischen Unmöglichkeit sei es ohnehin nicht wünschenswert eine exakte Replikation durchzuführen, da ein exaktes Experiment, sprich eine wahrhaft idente Replikation, folglich auch idente Ergebnisse und somit keinen Zugewinn an Informationen bringen würden. Der Mangel an informativem Nutzen einer exakten Replikation wird auch von Stroebe und Strack (2014) kritisiert. Diese vergleichen ebenso die exakte, direkte und konzeptuelle Replikation, begründen ihre Kritik der exakten Replikation im Gegensatz zu Schmidt (2009) jedoch in möglichen kulturellen und sozialen Unterschieden. Eine konzeptuelle Replikation sei überdies sehr wohl in der Lage den Mechanismus zu erfassen, welchen man dem untersuchten Effekt zugrundeliegend vermutet.

Rosenbaum (2001) stellt die *nicht-triviale* Replikation über die *triviale*, da es durch letztere nicht möglich sei zu sehen, ob ein Effekt auf die manipulierten Bedingungen des Experiments oder nicht beabsichtigte, etwaige andere Bedingungen zurückzuführen ist.

Eine weitere Unterscheidung behandelt Lykken (1968); er beschreibt die *wortwörtliche* Replikation, welche eine idente Version des Primärexperiments mitsamt all seiner Umstände darstellt (sinngleich der oben beschrieben exakten Replikation). Bei der *operationalen* Replikation werden Aufbau und Ablauf der Studie so konstruiert, wie sie von den Autor*innen der Primärstudie beschriebenen sind, und es wird auf die Einhaltung demografischer Besonderheiten der Stichprobe geachtet; so soll sich zeigen, ob die gegebenen Informationen für einen erneuten Nachweis des Effekts hinreichend sind. Die *konstruktive* Replikation orientiert sich lediglich an den Hypothesen des Primärexperiments, überlässt dann aber der Versuchsleitung die Entscheidung, wie diese zu ermitteln sind.

Der Kern einer Replikation liegt für Murray Lindsay und Ehrenberg (1993) nicht darin, ein Ergebnis bloß zu bestätigen, sondern zu überprüfen unter welchen anderen, von denen der Primärstudie abweichenden Bedingungen dieses Ergebnis bestätigt werden kann oder auch nicht – Schmidt (2009, S. 95) formuliert dies folgendermaßen: „Whereas a direct replication is able to produce *facts* a conceptual replication may produce *understanding.“* Er bemängelt jedoch das Fehlen einer klaren und praktikablen Abgrenzung zwischen direkter und konzeptueller Replikation. So würde etwa bei einer geglückten Replikation davon gesprochen, dass die Ergebnisse der Primärstudie hiermit bestätigen seien, bei einer nicht erfolgreichen, d. h. die Ergebnisse nicht wiederholende Replikation, lese man jedoch häufig, dies könne in der eventuellen Verschiedenheiten der beiden Studien begründet liegen.

Doyen, Klein, Simons und Cleeremans (2014) kritisieren auf gleiche Weise den begrenzten Nutzen von ausschließlich aus konzeptuellen Replikationen hervorgegangenen Ergebnissen, bei denen keine Möglichkeit bestehe von einer nicht gelungenen Replikation zu sprechen, da sich Unterschiede in den Ergebnissen stets mit den veränderten Bedingungen argumentieren ließen und nicht notwendigerweise auf den untersuchten Effekt zurückzuführen seien. Eine erfolgreiche konzeptuelle Replikation würde somit nie das vermutete Grundkonstrukt in Frage stellen oder gar falsifizieren können. Diese Meinung wird von Simons (2014) geteilt; er rät dazu, direkte Replikation in möglichst unterschiedlichen Laboratorien durchzuführen (vgl. Many Labs-Projekt von Klein et al., 2014), um zu überprüfen, wie weit ein Effekt auf Unterschiede im Setting und nicht auf die untersuchten Variablen zurückzuführen ist.

Brandt et al. (2014) bemühen sich die Begriffe exakt und direkt zu vermeiden, da es, wie bereits erwähnt, keine hundertprozentig exakten Replikationen geben kann. Jedoch sollte man stets bemüht sein, die entsprechende Studie so *nahe* (engl. close replication) wie nur möglich am Original zu halten. Für die möglichst direkte Replikation als Methode der Wahl spricht sich auch Lishner (2015) aus. Er gibt dieser Art der Replikation den Vorzug, da man aufgrund der geringeren Gefahr, falsch-positiven wie falsch-negativen Studienergebnissen Einzug in die Literatur zu gewähren und diese in weiterer Folge durch unterschiedliche Vorgehensweisen konzeptueller Replikationen zu wiederholen, eher zu verlässlichen Ergebnissen komme. Außerdem wäre es leichter, die Ergebnisse direkter Replikationen zusammenzuführen und zu analysieren, wodurch sich auch die Anzahl der den Ergebnissen zugrundeliegend vermuteten Erklärungen verringern ließe.

Auf eine weitere Problematik wissenschaftlicher Forschung, sowohl von Primär- als auch Replikationsstudien, verweisen Pashler und Harris (2012): Konzeptuelle Replikationen unterliegen eher einem Publikationsbias als direkte Replikationen. Infolge der geringen Publikationschance, nicht zuletzt aufgrund des Vorwurfs mangelnder Originalität, werden direkte Replikationen, trotz deren Zweckdienlichkeit, in der Praxis wenig und ungern durchgeführt (Schmidt, 2009). So wird etwa vom Journal of Personality and Social Psychology folgender Standpunkt vertreten: „The usual way of thinking would be that [a conceptual replication] is even stronger than an exact replication. It gives better evidence for the generalizability of the effect“ (Yong, 2012, S. 300).

Im Folgenden werden Verfahrensweisen näher beleuchtet, welche sich in den Forschungsalltag eingeschlichen haben bzw. schon seit jeher so gehandhabt werden, deren negativen Auswirkungen auf die Ergebnisse wissenschaftlicher Forschung jedoch zunehmend bewusst und diskutiert werden. Auf diesem Wissen aufbauend wird es

schließlich möglich sein, die Ergebnisse replizierter Studien im richtigen Licht zu interpretieren und aus ihnen Schlüsse für die Praxis zu ziehen.

1.2 Fragwürdige Forschungspraktiken

Unter fragwürdigen Forschungspraktiken (Questionable Research Practices) werden solche Maßnahmen verstanden, welche es Forscher*innen ermöglichen die Ergebnisse ihrer Studie in eine bestimmte, in der Regel von ihnen erwünschte Richtung zu lenken (John et al., 2012). Obgleich oft im selben Atemzug erwähnt, sei hier darauf hingewiesen, dass diese Praktiken von tatsächlichem Betrug (z. B. der Fälschung von Daten) abzugrenzen sind, wenn diese Abgrenzung auch nicht immer klar ersichtlich ist. Selbiger soll nicht weiter Gegenstand dieser Arbeit sein, mag er in vereinzelten Fällen auch die Ursache fälschlicherweise in die Literatur eingezogener Effekte darstellen.

In einer Studie von John et al. (2012) wurden über 2,000 Psycholog*innen anonymisiert und unter Zuhilfenahme erprobter Incentives dazu befragt, ob sie selbst je fragwürdige Forschungspraktiken angewandt hatten. Neben der Frage nach dem ob, wurde weiters erhoben wie hoch die Zahl jener Psycholog*innen eingeschätzt wird, die diese Praktiken ebenfalls angewandt hatten, als auch die Zahl jener, die solches auch zugeben würden. Die Ergebnisse der Studie zeigten, dass ca. 66% der Teilnehmer*innen bereits mindestens einmal darauf verzichteten, sämtliche von ihnen untersuchten Variablen zu berichten, etwa 57% setzten die Datenerhebung fort, nachdem sich gezeigt hatte, dass das Ergebnis ihrer Analyse nicht signifikant ausfiel, und ca. 22% beendeten die Datenerhebung, nachdem sich gezeigt hatte, dass die Ergebnisse signifikant ausfielen. Beinah die Hälfte der Teilnehmer*innen gab an, sie hätten bereits selektiv diejenigen Studien berichtet, die signifikante Ergebnisse hatten. Mehr als 40% der Teilnehmer*innen schloss bereits Daten aus, nachdem sich gezeigt hatte, welchen Einfluss dies auf das Ergebnis hatte.

Auch zu den von John et al. (2012) angeführten fragwürdigen Forschungspraktiken gehören das selektive Berichten von Bedingungen, das fälschliche Abrunden von *p*-Werten, die Behauptung ein unerwartetes Ergebnis vorhergesagt zu haben, die Negation demografischer Einflüsse auf das Ergebnis, und die Verfälschung von Daten (diese Vorgehensweise kann im Gegensatz zu den vorherigen jedoch unter keinen Umständen gerechtfertigt sein und fällt unter die oben genannte Art von Betrug).

Zusammenfassend lassen sich all jene Entscheidungen, die Forscher*innen im Laufe einer Studie treffen müssen, als Freiheitsgrade der Forschenden bezeichnen (Simmons et al., 2011). Zu diesen gehören u. a. die Beantwortung folgender Fragestellungen: Wann beende ich die Datenerhebung? Wenn ja, welche Daten werde ich ausschließen? Welche Bedingungen sollen verglichen werden? Ein Problem entsteht dann, wenn die Antworten auf

diese Fragen nicht zu Beginn, sondern während oder nach der Durchführung des Experiments bestimmt werden. All dies führt zu einer Zunahme falsch-positiver Studienergebnisse, welche Bestandteil der Forschungsliteratur werden.

Das starke Bedürfnis nach signifikanten Ergebnissen begünstigt wiederum data snooping (auch p-hacking), die negative Seite des data mining. Während man unter letzterem die legitime Art, große Datenpools strategisch auf Zusammenhänge hin abzusuchen versteht, meinst ersteres das wiederholte Testen von Modellen oder Verfahren mit dem Ziel, schlussendlich signifikante und die eigene Theorie bestätigende Ergebnisse zu bekommen (Bettis, 2012). Problematisch macht data snooping also erst die Weigerung, nicht-signifikante Ergebnisse zu akzeptieren und aus diesen die nötigen Konsequenzen zu ziehen. Eine ähnliche Vorgehensweise stellt data peeking dar; sie ist immer dann vorhanden, wenn bei einem unerwünschten Ergebnis solange Daten erhoben werden, bis sich das erhoffte Ergebnis einstellt. Diese Vorgehensweise wird auch als optionales Stoppen bezeichnet (Francis, 2012a). Gerade bei der Messung von mehreren Variablen werden einige stets signifikant ausfallen; diejenigen, die dies nicht tun, werden im Falle von data peeking ausselektiert. Derselben Gefahr laufen nicht signifikant gewordene Ergebnisse von vermeintlichen Pilotstudien, wenn es sich bei diesen tatsächlich nur um geringfügige Veränderungen ein und desselben Experiments handelt, welche bis zum ersten signifikanten Ergebnis durchgeführt werden.

Bedenklich ist auch das Ausschließen von Versuchspersonen mit schlechten Ergebnissen aufgrund des zur Zeit der Testung vorhandenem Lärms o. Ä., wenn zugleich eine andere Versuchsperson, die ebenso Lärm ausgesetzte war, jedoch wünschenswerte Ergebnisse vorweisen kann, nicht ausgeschlossen wird (Francis, 2012a). Diese Entscheidungen werden oft nicht im Vorhinein festgelegt sondern post hoc gefällt, was sich nicht nur auf den Wunsch, signifikante Ergebnisse zu finden, zurückführen lässt, sondern auch auf die Ungewissheit der Forschenden, wie die von ihnen gefragten Entscheidungen bestmöglich zu fällen sind (Simmons et al., 2011). Eine weitere Form post hoc erfolgter Entscheidungen bezeichnet Kerr (1998) als HARKing (Hypothesizing After the Results are Known), wonach es vorkommt, dass Autor*innen einer Studie post hoc, also nach vollendetem Experiment und aufgrund der beobachteten Ergebnisse ihre Hypothesen erstellen, diese aber als a priori bestimmt ausgeben. Gegensätzlich dazu ist eine explorative Vorgehensweise, welche klar als solche erkenntlich gemacht wird. All diese Praktiken führen zweifelsohne zu einer Minderung der Teststärke und einer Zunahme falsch-positiver Studienergebnisse.

Warum werden diese Praktiken überhaupt angewandt? Eine Erklärung dafür ist, dass Forscher*innen unter ständigem Druck stehen Arbeiten zu publizieren, sei es wegen

Anerkennung durch Sichtbarkeit in der wissenschaftlichen Gemeinschaft, oder aus der Notwendigkeit die eigenen Karrierechancen nicht durch eine scheinbar dürftige Anzahl von Publikationen zu gefährden (McBee & Matthews, 2014).

Wie falsch-positive, aber auch falsch-negative Studienergebnisse ohne die Anwendung fragwürdiger Forschungspraktiken zustande kommen, wird im Folgenden genauer beschrieben. Wie es möglich ist, dass falsche Ergebnisse überhaupt in großem Ausmaß Eintritt in die Literatur erhalten können, wo doch Möglichkeiten im Rahmen der Wahrscheinlichkeitstheorie dies verhindern sollen, wird außerdem erklärt. In diesem Sinne soll zuallererst auf die populärste aller Methoden zur inferenzstatistischen Bewertung wissenschaftlicher Hypothesen eingegangen werden, dem Nullhypothesen-Signifikanztesten.

1.3 Nullhypothesen-Signifikanztesten

Das statistische Verfahren des Nullhypothesen-Signifikanztestens (kurz: NHST) zählt wohl zu den bekanntesten aller Verfahren, welche die Wissenschaft nutzt, um Hypothesen zu überprüfen und zu zeigen ob den untersuchten Variablen (statistisch) signifikanter Einfluss zukommt. In seiner einfachsten Form wird anhand einer (wenn möglich zufällig gezogenen sowie den Bedingungen zugeordneten) Stichprobe ermittelt, ob sich diese entsprechend einer zuvor formulierten Alternativhypothese oder der Nullhypothese verhält. Letztere beschreibt in der Regel einen Nulleffekt, also die Abwesenheit einer Wirkung der jeweiligen Indikation. Auf Basis dieser Untersuchung sollen nun anhand der verwendeten Stichprobe Rückschlüsse auf die Population gezogen werden (= Inferenzstatistik); man verlässt sich hierbei auf diejenige Grenze, die den Unterschied zwischen Effekt und Zufall kennzeichnen soll, das sogenannte Signifikanzniveau. Dieses Niveau ist willkürlich gewählt und liegt meist bei einem Alpha-Wert von .05. Ein anhand der Stichprobe berechneter *p*-Wert muss hierbei kleiner als das zuvor festgelegte Alpha sein ($p < .05$), um von einem statistisch signifikanten Ergebnis sprechen zu können. Ein α von 5% bedeutet, dass ein Ergebnis bei gültiger Nullhypothese mit einer Wahrscheinlichkeit von 5% zugunsten einer Alternativhypothese spricht. Man nennt diese (Risiko-)Entscheidung den Fehler 1. Art bzw. α-Fehler.

An dieser Stelle muss darauf hingewiesen werden, dass der *p*-Wert fälschlicherweise oft als Wahrscheinlichkeit für den Wahrheitsgehalt eines Ergebnisses interpretiert wird; ein *p*-Wert von .05 bedeutet nicht, dass das berechnete Ergebnis mit einer Wahrscheinlichkeit von 5% der Wahrheit entspricht (Greenwald, Gonzalez, Harris, & Guthrie, 1996). Ebenso irreführend ist die Interpretation, dass ein für die Alternativhypothese sprechendes Ergebnis nicht zustande hätte kommen können, wenn die Nullhypothese wahr ist (vgl. Fehler 1. Art; Cohen, 1994). Der *p*-Wert beschreibt lediglich die Wahrscheinlichkeit das jeweilige Ergebnis

zu erhalten, wenn die Nullhypothese wahr ist. Weiters ist nur möglich zu zeigen, ob ein beobachteter Unterschied in den untersuchten Gruppen vorhanden ist, nicht aber wie dieser Unterschied zustande gekommen ist (Cohen, 1994). So kann eine verworfene Nullhypothese in vielem begründet liegen, bspw. der missglückten Operationalisierung, einer fehlerhaften Durchführung, oder Besonderheiten der Stichprobe, aber auch am stets vorhandenen Messfehler. Abseits davon stellt die Verwerfung der Nullhypothese ein gänzlich ungeeignetes Mittel dar, um Rückschlüsse auf die Nützlichkeit der Alternativhypothese zu ziehen (Gelman, 2015).

Pashler und Harris (2012) warnen davor, das Signifikanzniveau mit der maximal möglichen Fehlerrate von Ergebnissen in der Literatur gleichzusetzen und demonstrieren diesen häufig begangenen Fehlschluss an einem Beispiel: Vermutet man den Anteil der tatsächlich existenten an von Forschenden gesuchten Effekten bei 10%, so werden Fehler 1. Art (bei einem Alpha von 5%) bei 4.5% aller durchgeführten Studien auftreten (90% x 5%). Bei einer angenommenen Teststärke von 80% (d. h. der 80-prozentigen Wahrscheinlichkeit, einen Effekt in der untersuchten Stichprobe zu finden, wenn dieser tatsächlich existiert) entspricht dies einer korrekten Zurückweisung der Nullhypothese in 8% aller Fälle (80% x 10%). Die Wahrscheinlichkeit fehlerhafter und publizierter Ergebnisse (unter der Annahme, dass sämtliche positiven Ergebnisse publiziert wurden) läge somit bei 36%. Dies entspricht dem Anteil falsch-positiver Ergebnisse dividiert durch die Summe falsch-positiver Ergebnisse und der korrekterweise zu verwerfenden Ergebnisse [4.5% / (4.5% + 8%)]. Der auf diese Weise ermittelte Prozentsatz steigt mit sinkender Teststärke (Pahler & Harris, 2012). Anhand eines weiteren Beispiels zeigen Tversky und Kahneman (1971), warum es uns generell schwerfällt die Wahrscheinlichkeit von Ergebnissen richtig einzuschätzen. Den befragten Personen (bei diesen handelte es sich um Teilnehmer*innen eines Treffens der Mathematical Psychology Group der American Psychological Association) wurde folgende Aufgabenstellung vorgelegt: Bei einem Experiment (n = 20) wurde ein die Theorie bestätigendes Ergebnis eingefahren (z = 2.23; p < .05, zweiseitig) und nun sollen 10 zusätzliche Subjekte getestet werden. Auf die Frage, wie hoch die Wahrscheinlichkeit eines weiteren signifikanten Ergebnisses sei (einseitig und getrennt für diese Gruppe), antwortete die Mehrheit der befragten Personen mit einer Schätzung von 85%. Tatsächlich konnten nur 9 von 84 Personen die korrekte Zahl (48%) angeben.

Krueger (2001) hebt in seiner Arbeit drei Kritikpunkte des NHST hervor. Nullhypothesen seien demnach immer falsch, da sich keine Wahrscheinlichkeit einer Punkthypothese bestimmen lasse und somit Falsifikation nicht möglich sei. Krueger rät dazu die subjektive Einstellung der Forschenden zu den verwendeten Hypothesen sichtbar zu machen, da diese die Interpretation der Ergebnisse maßgeblich beeinflussen würde.

Zweitens würden signifikante Ergebnisse nichts über die Wahrscheinlichkeit der Verwerfung der Nullhypothese verraten; der Effekt einer bereits durchgeführten Studie eigne sich am besten zur Schätzung einer korrekterweise angenommenen Alternativhypothese. Drittens würden *p*-Werte keine Rückschlüsse auf die Höhe der Wahrscheinlichkeit erfolgreich replizierter Ergebnisse zulassen; die Summe der beiden Wahrscheinlichkeiten Teststärke und Fehler 1. Art hingegen schon.

Unter diesen Aspekten muss man sich eingestehen, dass NHST eine durchaus ungeeignete Methode ist, um auf ihrer Basis dichotome Entscheidungen zu fällen, und die Kritik an dieser Vorgehensweise, so sehr sie die wissenschaftliche Methodik auch dominiert, gerechtfertigt und notwendig ist. Cohen (1994) rät komplett vom Gebrauch des NHST ab, da die Nullhypothese bei ausreichend umfangreicher Stichprobe ohnehin immer verworfen werden wird. Um es mit den Worten von Greenwald et al. (1996) auszudrücken: NHST gibt den Forschenden nicht die Antworten, welche sie sich von ihren Ergebnissen erhoffen.

An dieser Stelle mag man sich die Frage stellen, warum und wozu NHST noch praktiziert wird, wenn nur wenig an ihr für die wissenschaftliche Forschung hilfreich zu sein scheint? Sterling, Rosenbaum und Weinkam (1995) zeigen, dass 94.3% der Artikel aus psychologischen Fachzeitschriften Signifikanztests verwenden, wohingegen dies in medizinischen Fachzeitschriften „nur“ 69.2% tun; hierbei wurden 8 Zeitschriften aus verschiedenen Bereichen der Psychologie bzw. 3 Zeitschriften aus Bereichen der Medizin kontrolliert. Einerseits lässt sich die große Beliebtheit des NHST auf das Vorhandensein einer dichotomen Entscheidungsbasis und somit die Möglichkeit zur recht simplen Beantwortung einer oft komplexen Fragstellung mit *Ja* oder *Nein* zurückführen, andererseits auf die leichte Verständlichkeit von *p*-Werten im Gegenteil zu anderen statistischen Kennzahlen. Die bereits erwähnte, oft falsche Interpretation dieser Werte, etwa als Wahrscheinlichkeit für die Korrektheit der Nullhypothese, leistet sicherlich auch einen Beitrag (Greenwald et al., 1996).

Cumming (2014) rät dazu, sich möglichst vom Berichten der NHST-Ergebnisse abzuwenden. Neben einem 25 Punkte enthaltenden Guide mit Richtlinien zur Verbesserung psychologischer Forschung und Förderung deren Integrität, unterbreitet er einen achtstufigen Plan mit Fokus auf Effektstärken, ausführliche Beschreibung von Prozedur und Datenanalyse, Berechnung und Interpretation von Konfidenzintervallen (und deren visueller Darstellung) sowie meta-analytische Herangehensweise und die Vermeidung dichotomer Beantwortung von Forschungsfragen. Diesen Anwendungsvorschlägen wurde entgegengehalten, *p*-Werte seien der Spitze eines Eisberges gleichzusetzen, deren Verbannung das tieferliegende Problem nicht lösen würde (Leek & Peng, 2015; Savalei & Dunn, 2015). Über den Verlauf einer Studie hinweg würden immer wieder Entscheidungen

gefällt, deren Diskussion hierbei unter den Tisch falle; so würden verbannte *p*-Werte schlussendlich durch andere Kennzahlen ersetzt werden (Leek & Peng, 2015). Hiermit wären Forschende schnell wieder am Anfang ihrer Bemühungen, sich von der Überbewertung eines einzigen Werts loszulösen.

Für einen kurzen Sprung in die Geschichte empfiehlt sich Nuzzo (2014), welche anschaulich erklärt, warum *p*-Werten heutzutage ein Stellenwert zukommt, der so nie gedacht war. Um dennoch eine korrekte Anwendung des NHST zu ermöglichen, raten Greenwald et al. (1996) zu folgender Herangehensweise: Das Berichten eines exakten *p*-Werts anstelle der Angabe, dass dieser kleiner oder größer als ein bestimmtes Signifikanzniveau sei (z. B. $p = .034$ statt $p < .05$), sowie die Betrachtung des *p*-Werts als ergänzende Statistik und nicht als ausreichende Form der Bewertung von Hypothesen. Forscher*innen sollten darüber hinaus stets sämtliche von ihnen errechneten Ergebnisse der für die jeweilige Studie bedeutsamen Inhalte als auch die für eine Sekundäranalyse benötigten Daten berichten.

1.4 Teststärke

Einer der Einflüsse, welche falsch-positive bzw. falsch-negative Studienergebnisse bei Primär- oder Replikationsstudien bewirken, ist eine ungenügend große Teststärke (Power; 1 - β). Die Teststärke beschreibt die Höhe der Wahrscheinlichkeit, einen Effekt zu entdecken, wenn dieser tatsächlich vorhanden ist, d. h. die Wahrscheinlichkeit, sich zugunsten einer korrekten Alternativhypothese zu entscheiden. Ebenso lässt sich aus ihr der Fehler 2. Art (β-Fehler) berechnen. Er gibt jene Wahrscheinlichkeit an, mit der man sich bei einer wahren Alternativhypothese irrtümlich für die Nullhypothese entscheidet (falsch-negatives Ergebnis).

Dass eine niedrigere Teststärke zu wenig verlässlichen Ergebnissen führt, liegt vor allem in der geringen Wahrscheinlichkeit, einen Effekt überhaupt zu entdecken, als auch an der hohen Wahrscheinlichkeit, die Größe eines entdeckten Effekts zu überschätzen (Button et al., 2013). Zusätzlich variiert die Größe eines Effekts stark in Abhängigkeit des gewählten Analyseverfahrens. Erschwerend kommt hier eine Reihe von Bias hinzu. Kleine Studien mit signifikanten Ergebnissen, welche mit hoher Wahrscheinlichkeit falsch-positiv sind, werden eher publiziert als nicht-signifikante Ergebnisse, welche nur selten veröffentlicht und häufig komplett verschwiegen werden (Kühberger, Fritz, & Scherndl, 2014). Außerdem mangelt es kleinen Studien oft auch an anderer Stelle an Qualität; so scheinen sie oft opportunistischer Natur und ohne ausführliche Planung entstanden zu sein (Button et al., 2013).

Die Höhe der Teststärke steht in direktem Zusammenhang mit der Größe der untersuchten Stichprobe, dem gewählten Signifikanzniveau und der Effektstärke

(Schimmack, 2012). So wird beispielsweise eine große Stichprobe benötigt, um eine ausreichende Teststärke zu erlangen, wenn der gesuchte Effekt gering ist; bei einem großen Effekt reicht bereits eine kleine Stichprobe aus, um diesen bei ausreichender Teststärke nachweisen zu können. Rosnow & Rosenthal (1989) demonstrierten anhand eines Beispiels sehr anschaulich den Vergleich zweier Ergebnisse: Studie 1 mit relativ großer Stichprobe (n = 80) fällt signifikant aus, Studie 2 mit kleiner Stichprobe (n = 20) hingegen nicht. Bei alleiniger Betrachtung der p-Werte könnte man meinen, der Effekt wäre nicht repliziert worden. Die Effektstärken beider Studien sind jedoch ident. Betrachtet man auch noch die Teststärken, erkennt man, dass sich diese stark unterscheiden (Teststärke der Studie 1 beträgt .6, die der Studie 2 beträgt .18). Dieser Vergleich macht deutlich, wie wenig p-Werte über Studienergebnisse aussagen, und wie viel anders dieser Vergleich aussieht, wenn man den Fokus stattdessen auf Effekt und Teststärken legt.

Dazu kommt, dass nicht die Anzahl der durchgeführten Studien für die Effektstärke ausschlaggebend ist, sondern die Größe der Stichprobe; so erhält man in Summe dieselbe Effektstärke, egal ob 10 Studien mit je 100 Teilnehmer*innen durchgeführt werden, oder aber nur eine Studie mit 1,000 Teilnehmer*innen (Schimmack, 2012). Ein Vorteil der Kombination mehrerer Studien ist die Milderung von Bias-Auswirkungen, welche bei kleinen Studien generell größer ausfallen bzw. durch die Anwendung fragwürdiger Forschungspraktiken entstehen (Bakker, van Dijk, & Wicherts, 2012).

Hier könnte man schlussfolgern, es wäre am sinnvollsten eine möglichst große Stichprobe heranzuziehen, weil dadurch die Wahrscheinlichkeit des β-Fehlers gegen Null gehen müsste. Nun sind große Stichproben in der psychologischen Forschung einerseits eine Seltenheit (Kühberger et al., 2014), andererseits ist die Teststärke nur dann in großem Ausmaß von der Stichprobengröße abhängig, wenn diese alleiniger Zufallsfaktor ist (Westfall, Judd, & Kenny, 2015). In aller Regel interessieren sich Forschende jedoch für einen oder mehrere Stimuli, welche oft in unterschiedlichem Ausmaß variieren. Unter diesen Umständen stößt die Beeinflussbarkeit der Teststärke durch die Stichprobengröße bald an ihre Grenze und bringt diese in eine maximale Höhe, welche um einiges geringer als 1 ist. Für eine 32 Stimuli enthaltende Stichprobe (16 je Bedingung) mit einer Variabilität von V_S = 30% berechnen Westfall et al. (2015) eine maximal erreichbare Teststärke von .7, wohingegen eine Stichprobe mit 8 Stimuli (4 je Bedingung) und einer Variabilität von V_S = 10% eine maximale Teststärke von weniger als .5 erreicht. Auf Basis dieser Berechnungen raten Westfall et al. (2015) im Rahmen von Replikationsversuchen davon ab, bei gleichbleibender Anzahl der Stimuli ausschließlich die Größe der Stichprobe zu verändern, da sich so nur selten eine höhere Teststärke als in der Primärstudie erreichen lasse. Besser

sei es (auch bei einer direkten Replikation) ein neues Set an Stimuli zu verwenden sowie eine größere Anzahl an Stimuli, um auf diese Weise für eine höhere Teststärke zu sorgen.

Für eine anschauliche Darstellung der Wechselwirkung von Stichprobengröße, Teststärke und Effektgröße bzw. als Berechnungshilfe empfiehlt sich eine interaktive Grafik wie die von Magnusson (2015), abrufbar unter http://rpsychologist.com/d3/NHST/[1]; im Falle einer Replikation bzw. einer bereits vorhandenen Schätzung der Effektstärke lässt sich auf Basis dieses Modells die Anzahl an Versuchspersonen berechnen, welche für die gewünschte Teststärke benötigt werden. Cohen (1988) empfiehlt bspw. bei Medikamentenstudien eine Teststärke von mindestens 80% zu wählen, womit die Wahrscheinlichkeit einen vorhandenen Effekt nicht zu entdecken bei unter 20% liegt.

Lucas (2013) rät dazu, die erwartete Effektstärke bereits im Vorhinein (anhand früherer Forschung) zu ermitteln als auch zu begründen, und anhand der gewählten Stichprobe sowie dem erwarteten Effekt die Teststärke zu berechnen. Soll in einer Studie ein Effekt erforscht werden, der noch nie zuvor ermittelt wurde, muss dieser bspw. anhand einer Pilotstudie geschätzt werden (Stanley & Spence, 2014).

Da man generell das Problem hat, dass sich nie genau sagen lasse, wann genau ein Effekt überhaupt zu klein sei, um noch im Sinne der Theorie zu sein, als auch die wenigsten Experimente eine ausreichend große Stichprobe heranziehen können, welche benötigt werden würde, um sich bei einem wirklich kleinen Effekt bei hinreichender Teststärke für die Nullhypothese zu entscheiden, schlägt Simonsohn (2015) eine neue Methode vor. Diese soll es ermöglichen, die Nullhypothese in Zukunft nicht nur zu verwerfen, sondern einen Nulleffekt tatsächlich auch zu akzeptieren. Die Nichtexistenz eines Effekts zu beweisen ist schwer, leichter ist es zu beweisen, dass eine Methode nicht gut genug war, einen Effekt überhaupt erst entdecken zu können. Dies sei anhand einer Replikation möglich, die sich eines größeren „Teleskops“ bedient als die Primärstudie. Simonsohn (2015) verdeutlicht dies anhand der folgenden Analogie:

> Imagine an astronomer claiming to have found a new planet with a telescope. Another astronomer tries to replicate the discovery using a larger telescope and finds nothing. Although this does not prove that the planet does not exist, it does nevertheless contradict the original findings, because planets that are observable with the smaller telescope should also be observable with the larger one (S. 560).

Im Rahmen einer Replikation soll so gezeigt werden können, dass ein Effekt höchstwahrscheinlich nicht existiert. Berechnet man anhand der Primärstichprobe einen

[1] Dieses Modell basiert auf Einstichproben-Gauß-Tests.

Effekt, welcher bei einer Teststärke von 33% – dies entspricht der Wahrscheinlichkeit eines nicht-signifikanten Ergebnisses zu einem signifikanten im Verhältnis 2:1 und somit einer sehr geringen Teststärke – gerade noch auffindbar gewesen wäre ($d_{33\%}$), und zeigt sich bei erneuter Durchführung der Studie ein Effekt, welcher statistisch signifikant kleiner ist als $d_{33\%}$, so lässt sich dies nicht mit der Annahme vereinbaren, der Effekt wäre ausreichend groß gewesen, um anhand der Primärstichprobe überhaupt auffindbar gewesen zu sein (Simonsohn, 2015).

Neben den vielen falsch-positiven Ergebnissen werden von Fiedler, Kutzner und Krueger (2012) auch die Folgen falsch-negativer Ergebnisse beleuchtet und als viel problematischer für die Forschungsliteratur interpretiert. So lassen sich diese nicht ohne weiteres durch die Replikation von Studienergebnissen nachweisen, wie es bei falsch-positiven Ergebnissen der Fall ist. Wurde eine Alternativhypothese erst einmal abgelehnt, würde gemäß Fiedler et al. (2012) noch weniger Anreiz bestehen, diese zu replizieren, als bei falsch-positiven Ergebnissen.

Einen Vorschlag zur Förderung von Studien mit hoher Teststärke bieten Fraley und Vazire (2014). Anhand des sogenannten N-Pact Factor ordnen sie Fachzeitschriften nach Höhe der Teststärke der von ihnen publizierten Studien. Dies soll Forschenden ein Anreiz sein, die Teststärke ihrer Studien möglichst effizient einzusetzen, und Leser*innen die Möglichkeit geben, ihnen vorliegende Studienergebnisse besser einschätzen zu können.

Signifikant oder nicht, Effekte selbst sind immer noch keine ausreichende Bestätigung dafür, dass die verwendeten Messinstrumente tatsächlich die zu erforschen gewünschte Theorie untersucht haben und es sich um eine angemessene Operationalisierung des jeweiligen Konstrukts handelt. Effektstärken sagen uns auch nichts darüber, wie groß der Anteil falsch-positiver Studienergebnisse in der Literatur ist. Warum viele publizierte Studien gefährdet sind einen Fehler 1. Art zu beinhalten, wird im Folgenden anhand des Publikationsbias erklärt.

1.5 Publikationsbias

Einer der führenden Gründe warum (nicht nur) die psychologische Forschungsliteratur mit dem Problem falsch-positiver Studienergebnisse und somit mangelhaft replizierbarer Effekte zu kämpfen hat, findet sich im Publikationsbias. Bei einem Signifikanzniveau von .05 als Basis für die Berechnung der Anzahl falsch-positiver Studienergebnisse, kommen Pashler und Harris (2012) auf eine ihnen zufolge optimistisch geschätzte Rate von 36%. Nimmt man diese Zahl unter Berücksichtigung des bevorzugten Berichtens signifikanter Studienergebnisse in Fachzeitschriften her, kommt man schnell zu der Frage, wie es sein kann, dass mehrheitlich positive (signifikante) Ergebnisse berichtet

werden, sich dann aber herausstellt, dass es sich bei diesen zu über einem Drittel um nicht replizierbare Effekte handelt?

Wenn beinahe alle publizierten Studien ein signifikantes Ergebnis vorweisen, kann dies allein aufgrund der Wahrscheinlichkeit, welche den verwendeten Berechnungen unterliegt, nicht repräsentativ für alle tatsächlich durchgeführten Studien sein (Francis, 2012b). Hinweise, wie das Fehlen von nicht-signifikanten Studienergebnissen in der Forschungsliteratur, können auf eine selektive Berichterstattung durch den Publikationsbias deuten (Francis, 2012a, 2013). Dessen ursprüngliche Definition bezieht sich darauf, dass Signifikanz und Richtung von Studienergebnissen maßgeblich Einfluss darauf nehmen, ob diese Ergebnisse auch in Fachzeitschriften veröffentlicht werden oder nicht (Rothstein, Sutton, & Borenstein, 2005). So ist die Wahrscheinlichkeit einer Publikation bei einem signifikanten Ergebnis sehr viel höher, als wenn es sich um ein nicht-signifikantes Ergebnis handelt. Anhand von 221 TESS-Studien (Time-sharing Experiments for the Social Sciences) zeigte sich, dass Studien, welche Nullergebnisse aufweisen, nur zu 20.8% publiziert wurden; außerdem wurden 64.6% der Nullergebnisse gar nicht erst in Form eines Artikels gebracht (Franco, Malhotra, & Simonovits, 2014). Eine Ausnahme hiervon stellen Replikationen dar, deren Publikation sich unabhängig von der Art der Signifikanz um einiges schwieriger gestaltet als die von Primärergebnissen (Francis, 2012a). Werden diese Ergebnisse dennoch publiziert, so handelt es sich hierbei mehrheitlich um die Primärstudie bestätigenden Replikationen (Makel et al., 2012).

Eine höhere Wahrscheinlichkeit der Publikation signifikanter Effekte wird auch im File Drawer-Effekt beschrieben (Rosenthal, 1979). Dieser besagt, dass signifikante Ergebnisse bevorzugt publiziert werden, nicht-signifikante Ergebnisse jedoch selten und somit in der Schublade (drawer) der Forschenden verbleiben. Der Verdacht, nicht alle Studien würden ihre Ergebnisse berichten bzw. vollständig berichten (d. h. Nullergebnisse in der Schublade verschwinden lassen), lässt sich auch anhand des großen Interesses an Replikationen prominenter Studienergebnisse ablesen. So gibt es mit http://www.psychfiledrawer.org/top-20/ (Pashler, Spellman, Kang, & Holcombe, 2015) eine eigene Website die jene Studieneffekte listet, von welchen sich Forscher*innen am ehesten eine erfolgreiche Replikation wünschen.

Young, Ioannidis und Al-Ubaydli (2008) beschreiben die Tendenz, dass ungewöhnliche und extreme Ergebnisse eine größere Chance haben publiziert zu werden – demzufolge wäre das durchschnittliche Ergebnis zu einem bestimmten Effekt wahrscheinlich eine realistischere Einschätzung seiner tatsächlichen Größe. Young et al. (2008) prangern außerdem das Erschaffen künstlicher Knappheit durch Zeitschriften an: Eine gängige Begründung dafür, ein Manuskript nicht zu publizieren, sei es auf den mangelnden

Druckplatz hinzuweisen. In Anbetracht der Weiten des digitalen Raums ein nur wenig nachvollziehbares Argument. Diese Vorgehensweise begünstigt jedoch eine weitere fragwürdige Methode, welche darin besteht, möglichst viele signifikante Studienergebnisse vorweisen zu können, indem möglichst kurze Berichte veröffentlicht werden, welche folglich oft wenige Studien bzw. nur eine einzige Studie und geringe Stichproben enthalten. Im Extremfall führt dies zur sog. Salamitaktik, welche das Zerteilen einer Studie in mehrere kleine Studien beschreibt, um auf diese Weise mehr als einen Artikel publizieren zu können (Abraham, 2000). Leider zeigen vor allem kleine Studien einen Hang zu falsch-positiven Ergebnissen, da bei deren geringen Teststärke die Wahrscheinlichkeit eines (statistisch) signifikanten Effekts sehr hoch ist (Bertamini & Munafò, 2012). Kühberger et al. (2014) untersuchten den Zusammenhang zwischen Effektstärke und Stichprobengröße. Es wurde ein Stichprobe von 1,000 Artikel gezogen, welche verschiedenen psychologischen Forschungsrichtungen entstammten. In diesen wurden ausschließlich empirisch-quantitative, inferenzstatistische Forschungsmethoden angewandt: Häufigkeitstests, Testungen der Mittelwertdifferenzen und Varianzanalysen, Korrelationen, lineare Regressionen und Rangordnungstests. Es zeigte sich eine stark negative Korrelation dahingehend, dass Studien mit kleinen Stichproben größere Effektstärken berichteten als Studien mit großen Stichproben, r = -.54 mit 95% *KI* [- 0.6, - 0.5] bzw. r = -.45 mit 95% *KI* [- 0.53, - 0.36] beim Ausschluss extremer Stichprobengrößen. Weiters konnte gezeigt werden, dass Studien, deren p-Wert gerade noch signifikant ausfällt, zu jenen Studien, deren p-Werte gerade nicht mehr signifikant sind, im Verhältnis 3:1 stehen. Weder die Berechnung der Teststärken anhand der Artikel noch die Befragung der Autor*innen selbst ließ darauf schließen, dass das Erzielen einer ausreichend großen Teststärke bei der Planung der Studien berücksichtigt worden ist; 5% der Studien enthielten die berechnete Teststärke, 8% erwähnten sie. Dies lässt auf einen weitreichenden Einfluss durch den Publikationsbias schließen (Kühberger et al., 2014).

Eine Möglichkeit der Publikation zugunsten signifikanter Studien entgegenzuwirken wäre es, Studien im Vorab zu registrieren und somit zu verhindern, dass diese durch ein nicht-signifikantes Ergebnis dem File Drawer-Effekt und somit dem Publikationsbias erliegen (Open Science Collaboration, im Druck). Es gibt jedoch auch Forscher*innen, welche sich aus verschiedenen Gründen gegen das Berichten nicht-signifikanter Studien aussprechen. So warnt Francis (2012b) vor einer Flut von qualitativ wenig wertvollen Experimenten, sollte es Teil der wissenschaftlichen Praxis werden, diese vorab zu registrieren. Sein Vorschlag zur Bekämpfung des Publikationsbias beinhaltet die noch selektivere Auslese von Studien. So sollen beispielsweise Studien mit zu kleinen Stichproben nicht publiziert und Pilotstudien sowie explorative Forschung nicht als wissenschaftliche Entdeckungen deklariert werden. Würde es sich bei den durch Publikationsbias begünstigten Studien aber um solche mit

hoher Teststärke handeln, wäre dies nicht weiter problematisch. In dieselbe Richtung argumentieren de Winter und Happee (2013). Sie bestätigen die Sinnhaftigkeit der Veröffentlichung signifikanter wie nicht-signifikanter Ergebnisse, wenn hierdurch die Chance einer akkuraten Replikation von Studienergebnissen gesteigert wird – schließlich wäre mit der Zeit eine Regression zur Mitte erkennbar und somit eine genauere Einschätzung des untersuchten Effekts möglich. Anhand einer Simulationsstudie zeigten sie jedoch auch, dass selektives Publizieren von signifikanten Ergebnissen (unabhängig von der Richtung des Effekts) unter der Annahme einer sich selbst verbessernden Wissenschaft von Vorteil sein kann, da hierdurch weniger Studien für die Einschätzung eines Effekts durch eine Meta-Analyse benötigt werden würden. Somit raten de Winter und Happee (2013) erst dann zur selektiven Veröffentlichung von Studienergebnissen, wenn eine Hinwendung der Wissenschaft zur Selbstverbesserung erkennbar ist und vorhandene Effekte auch in Frage gestellt werden, also Replikationen stattfinden.

1.6 Publikationsbias bei Replikationen

Wie bereits erwähnt, wird durch Zufall allein schon eine bestimmte Anzahl von Studienergebnissen publiziert, die ein signifikantes Ergebnis vorweisen und die Nullhypothese irrtümlich verwerfen. Diese scheinbar erfolgreich replizierten Ergebnisse werden dann gemäß des False Hypothesis Bias eher publiziert als nicht-signifikante Ergebnisse (Sterling et al., 1995). Da Replikationen generell seltener publiziert werden, noch dazu, wenn sie nicht-signifikante (bzw. nicht in die Richtung des Primäreffekts deutende) Ergebnisse aufweisen, kommt es vor, dass früher oder später ein Ergebnis durch bloße Wahrscheinlichkeit signifikant wird, welches dann eher publiziert wird und wiederum eine falsche Hypothese bestätigt. Wenn von einer Studie also eine mehr als statistisch glaubwürdige Anzahl an erfolgreichen Replikationen berichtet wird, kommt man um den Verdacht nicht umhin, dass es sich hierbei entweder um nicht ordnungsgemäß durchgeführte Experimente handelt, oder deren Ergebnisse nicht vollständig berichtet wurden (Francis, 2012a).

Zur Veranschaulichung des Anteils an Replikationen in psychologischen Fachzeitschriften dient der von Makel et al. (2012) nach Impact Factor berechnete Wert einer Replikationsrate von 1.07% – knapp einer aus 100 veröffentlichten Artikel beinhaltet also eine Replikation. Von ursprünglich insgesamt 321 411 Artikel aus den Top 100 Fachzeitschriften der Psychologie (gemessen am Impact Factor) wurden diejenigen herausgefiltert, welche den Begriff „replicat*“ aufwiesen (5,051 Artikel bzw. 1. 57%), und anschließend jene 342 Artikel (68.4%) verwendet, bei denen es sich tatsächlich um Replikationen handelte. Bei diesen Replikationen gab es in 52.9% der Fälle mindestens eine gemeinsame Autorin bzw. einen gemeinsamen Autor. Dies scheint auch einen Einfluss

darauf zu haben, wie hoch die Rate erfolgreicher Replikationen ausfällt: Es zeigte sich eine Erfolgsquote von 91.7% bei Überschneidungen von Autor*innen im Vergleich zu 64.6% erfolgreichen Replikationen, wenn keine Überschneidung vorhanden ist. Ebenso berechneten Makel und Plucker (2014) die Replikationsrate für die Top 100 Fachzeitschriften der pädagogischen Forschung. Von ursprünglich 164 589 Artikel beinhalteten hier 461 den Begriff „replicat*“ und bei 221 dieser Artikel (47.9%) handelte es sich um Replikationen; dies entspricht einer Gesamtreplikationsrate von 0.13%.

Neben den Einschränkungen durch den Publikationsbias und den File Drawer-Effekt können sich noch weitere Bias behindernd auf die Durchführung von Replikationen auswirken: submission bias, funding bias, editor/reviewer bias, journal publication policy bias, promotion bias, journals-analyzed bias, novelty equals creativity bias (Makel & Plucker, 2014). Replikationsstudien haben beispielsweise eine geringere Chance publiziert zu werden, weil ihnen – vor allem direkten Replikationen – mangelhafte Originalität vorgeworfen wird (Pashler & Harris, 2012; Schmidt, 2009). Diese Meinung wird häufig auch von Editor*innen und Reviewer*innen vertreten: „Our attention is focused on avoiding replication! There are so many interesting subjects which have not been studied that it is a stupid thing to make the same work once again“ (Madden, Easley, & Dunn, 1995, S. 79). Dass die Durchführung von Replikationsstudien bei Forschenden nicht überaus populär ist, liegt neben den geringeren Publikationschancen auch an möglichen bzw. gerade dadurch gegebenen negativen Auswirkungen auf die eigene Karriere (McBee & Matthews, 2014).

1.7 Interpretation von Replikationen

Wie lässt sich überhaupt eine „erfolgreiche“ von einer „nicht-erfolgreichen“ Replikation unterscheiden? Die Ursache kann einerseits in der Primärstudie begründet liegen: Es kann sich beispielsweise um ein falsch-positives oder falsch-negatives Ergebnis handeln (Open Science Collaboration, 2012, 2014). Zufall durch statistische Wahrscheinlichkeit (α-Fehler, β-Fehler), unbekannte Moderatoren, fragwürdige Forschungspraktiken (willkürliches Beenden der Datenerhebung, selektives Berichten signifikanter Ergebnisse etc.), eine zu geringe Teststärke, oder Fehler in der Umsetzung können die Ursache sein. Erschwerend kommt noch die bevorzugte Publikation signifikanter, ungewöhnlicher und neuer Effekte (siehe Publikationsbias) hinzu, welche dazu führt, dass sich unter publizierten und somit potentiellen Primärstudien von vornherein eine höhere Rate falsch-positiver Ergebnisse findet. Falsch-positive bzw. falsch-negative Ergebnisse durch Zufall, mangelnde Teststärke, Fehler in der Umsetzung der Replikation (Methode, Stichprobe, Analyse, Interpretation), unbekannte Moderatoren, Mangel hinreichender Informationen zur Durchführung sowie fragwürdige Forschungspraktiken können wiederum Gründe für eine erfolgreiche oder nicht-erfolgreiche Replikation sein (Open Science

Collaboration, 2012, 2014). Eine scheinbar erfolgreiche, d. h. die Ergebnisse der Primärstudie bestätigende Replikation, bedeutet daher noch nicht, dass ein Effekt tatsächlich existiert (Open Science Collaboration, 2012). Wie unter anderem der Fall Diederik Stapel zeigt (Callaway, 2011), ist auch Betrug (sowohl bei der Primär- als auch der Replikationsstudie) nicht vollständig auszuschließen.

Wie können die Auswirkungen des Publikationsbias akkurat eingeschätzt, und wie kann ihnen entgegengewirkt werden? Im Folgenden wird eine Reihe von Methoden genannt, welche ein neues Licht auf Unmengen an scheinbar vorhandenen Effekten werfen sollen. Sie ermöglichen es, den Einfluss des Publikationsbias abzuschätzen und auf die Anwesenheit falscher Studienergebnisse hinzuweisen.

1.8 Lösungsansätze

Um die Auswirkungen des Publikationsbias einschätzen zu können, gibt es eine Reihe von Möglichkeiten. So lässt sich dessen Einfluss mit Hilfe des von Simonsohn, Nelson und Simmons (2014a, 2014b, im Druck) entwickelten Konzepts der *p*-curve offenlegen. Anhand des Vergleichs der *p*-Werte mehrerer Studien hilft sie zu erkennen, wie groß der Einfluss selektiver Berichterstattung auf einen bestimmten Effekt ist. Gemäß der Auftrittswahrscheinlichkeit falsch-positiver Ergebnisse unter den einzelnen Effekten ($p < .05$ = 5%, $p < .04$ = 4% usw.) lässt sich die Verteilung der signifikanten Effekte, unter der Annahme normalverteilter Daten, anhand einer Kurve darstellen (Simonsohn et al., 2014a). Sieht man sich nun die Verteilung wenig- bis hoch-signifikanter *p*-Werte eines untersuchten Effekts an, lässt sich diese Kurve mit einer Reihe von beschriebenen Kurvenvarianten (Simonsohn et al., 2014a) vergleichen, und sich somit abschätzen, ob selektive Berichterstattung oder *p*-hacking stattgefunden hat. Je größer ein vorhandener Effekt ist, desto rechtsschiefer wird die Verteilung sein, wenn kein *p*-hacking stattgefunden hat. Ist kein wahrer Effekt vorhanden, wird die Verteilung stetig ausfallen, wenn die Ergebnisse nicht durch *p*-hacking beeinflusst wurden, und eher linksschief, wenn *p*-hacking vorgekommen ist. Wenn durch *p*-hacking beeinflusste Studienergebnisse vorhanden sind und der Effekt wahr ist, wird die Verteilung mit zunehmender Stichproben- und Effektgröße rechtsschief werden. Die Möglichkeit zur Berechnung und Interpretation einer *p*-curve für einen gewünschten Effekt wird von Simonsohn, Nelson und Simmons (2015) unter http://p-curve.com/ angeboten.

Wie zuverlässig die Abschätzung von Publikationsbias und *p*-hacking anhand der *p*-curve ist, hängt zum einen von der verwendeten Anzahl an Studien bzw. *p*-Werten ab, deren Teststärke, sowie dem Ausmaß, in welchem *p*-hacking tatsächlich stattgefunden hat. So lässt sich mit einer Zahl von 20 *p*-Werten auch noch bei 50% Teststärke ein signifikantes

Ergebnis, welches für bzw. gegen *p*-hacking spricht, berechnen, wenn *p*-hacking tatsächlich existiert. Es gibt einige Limitierungen, die bei der Anwendung der *p*-curve zu berücksichtigen sind (Simonsohn et al., 2014b). Zu diesen zählt die Verwendung von Studien, welche einen unterschiedlich großen Effekt unter bestimmten Bedingungen vorhersagen; hier wird von der Inkludierung des *p*-Werts in die *p*-curve abgeraten, da auf diese Weise die Größe der Effektstärke falsch eingeschätzt werden würde. Die *p*-curve ist nicht bei Studien mit diskreten Teststatistiken anwendbar, zudem erhöht sich bei der Existenz einer mit der untersuchten Variable korrelierenden Kovariate die Wahrscheinlichkeit, dass die *p*-curve die Anwesenheit evidenter Daten verkennt. Vorteil und Einschränkung zugleich ist das Verwenden ausschließlich signifikanter Ergebnisse; obwohl dies keinen Bias verursacht beinhaltet die *p*-curve dadurch doch weniger erklärte Varianz, da sich keine Aussage über die Verteilung nicht-signifikanter Ergebnisse treffen lässt. Im Falle vermuteter Moderatorvariablen ist die Berechnung separater *p*-curves erforderlich. Erwähnt sei noch der unwahrscheinliche Fall, dass *p*-hacking auch zur Unterschätzung der Effektgröße führen kann. Simonsohn et al. (im Druck) führen ferner Lösungsmöglichkeiten für Probleme, wie das Berichten ausschließlich sehr kleiner signifikanter *p*-Werte (*p*-hacking, welches ausschließlich im signifikanten Bereich stattfindet), Missachtung der Voraussetzungen (z. B. müssen *p*-Werte statistisch unabhängig von anderen in Studien berichteten *p*-Werten sein) und beabsichtigte sowie unbeabsichtigte Fehler bei der Berichterstattung von Daten. Im Übrigen soll die *p*-curve nicht als Methode zur Verwerfung von Theorien hinter den untersuchten Effekten betrachtet werden; sie dient lediglich dem Sichtbarmachen selektiver Berichterstattung (Simonsohn et al., 2014a).

Die Berücksichtigung des Publikationsbias ist auch bei der Meta-Analyse von großer Bedeutung. Da Meta-Analysen den Anspruch erheben, eine genauere Einschätzung von Effekten zu sein als einzelne Studienergebnisse, muss hier besondere Rücksicht auf das Ergebnis potentiell verzerrende Phänomene – sprich File Drawer, Publikationsbias, graue Literatur – genommen werden (Borenstein, Hedges, Higgins, & Rothstein, 2009). Ein häufig verwendetes Hilfsmittel hierbei ist der Funnel Plot, eine grafische Darstellung in Form eines Streudiagramms. Die durch einzelne Punkte dargestellten Studien deuten hier bei asymmetrischer Anordnung (etwa offenbar fehlender Punkte/Studien im mittleren und unteren Bereich) auf die Anwesenheit des Publikationsbias (Light & Pillemer, 1984, zitiert nach Borenstein et al., 2009). Problematisch hierbei ist jedoch, dass die Form des Funnel Plot nicht nur durch den Publikationsbias beeinflusst wird, sondern auch durch Qualitätsunterschiede einzelner Studien, welche starke Schwankungen hinsichtlich der symmetrischen Anordnung von Studienpunkten bewirken können (Tang & Liu, 2000). Eine Möglichkeit, um den Summeneffekt der einzelnen Studien unter Berücksichtigung des Publikationsbias zu berechnen, ist die Trim and Fill-Methode. Diese beschreibt das Weglassen kleiner Studien zwecks Erlangung der Symmetrie des Funnel Plots und den

anschließenden Austausch dieser Studien mit anhand des Summeneffekts berechneter Ersatzstudien, welche die symmetrische Gegenstudie zu diesen echten, aber kleinen Studien darstellen (Mavridis & Salanti, 2014). Jedoch stellt der Publikationsbias auch hier nicht die einzig denkbare Ursache dar. Eine weitere Möglichkeit zur Kontrolle des Publikationsbias bieten Selektionsmodelle. Sie berechnen die Gesamteffektstärke, indem sie den einzelnen Studien basierend auf Basis der geschätzten Wahrscheinlichkeit, nach welcher diese publiziert wurden (vor allem in Abhängigkeit des *p*-Werts und der Stichprobengröße), Gewichte zuteilen (Mavridis & Salanti, 2014).

Eine andere Hilfestellung, um die Anwesenheit von zu vielen erfolgreichen Studienergebnissen zu ermitteln, genannt Incredibility Index, beschreibt Schimmack (2012). Anhand der berichteten Effektstärken eines Studiensets sollen post hoc die Teststärken berechnet werden und somit die Wahrscheinlichkeit, mit welcher ein signifikantes Ergebnis zu erwarten war. Auf Basis der Teststärke lässt sich somit ermitteln, wie viele Experimente der Studie gleich viele oder weniger signifikante Ergebnisse (als berichtet) erzielen hätten müssen. Auf diese Weise kann der Incredibility Index bei größeren Stichproben die Anwesenheit des Publikationsbias oder von fragwürdigen Forschungspraktiken anzeigen. Da die durchschnittliche Teststärke einer Gruppe von Studien auf Basis des Mittelwerts berechnet wird, läuft der Incredibility Index jedoch Gefahr die wahre Teststärke falsch einzuschätzen, da Stichprobenfehler bei geschätzten Teststärken in der Regel nicht normalverteilt sind (Yuan & Maxwell, 2005, zitiert nach Schimmack, 2014). Die Weiterentwicklung des Incredibility Index ist der Replicability-Index (Schimmack, 2014). Dieser „Doping"-Test für die Forschung soll die Replizierbarkeit bereits veröffentlichter Studien berechnen und ergibt sich aus der Differenz des Prozentsatzes signifikanter Ergebnisse und des Medians (als Schätzung der wahren Teststärke). Der Replicability-Index steigt einerseits mit zunehmender Teststärke an, da sich Studien mit größerer Teststärke eher replizieren lassen. Andererseits spricht eine zunehmende Diskrepanz zwischen der beobachteten Teststärke und dem Prozentsatz signifikanter Ergebnisse dafür, dass fragwürdige Forschungspraktiken angewandt wurden.

1.9 Incentives

Der Mangel an Incentives für eine transparente, Replikationen gutheißende Wissenschaft mag in der Forschungspolitik einzelner Institutionen oder an gängigen Richtlinien für beispielsweise die Vergabe von Forschungsgeldern begründet liegen (Nosek et al., 2015). Einen wichtigen Aspekt stellen hierbei Richtlinien dar, an welchen sich Zeitschriften bei der Publikation von Studien orientieren. Im Rahmen eines am Center for Open Science abgehaltenen Treffens, hat das Transparency and Openness Promotion (TOP) Committee ein Set solcher Richtlinien erstellt, welches die Transparenz von in

Fachzeitschriften veröffentlichten Arbeiten erhöhen soll (Alter et al., 2015). Darüber hinaus können diese bei der Vergabe von Forschungsgeldern u. Ä. berücksichtigt und gegebenenfalls angepasst werden. Es werden drei Ebenen der Transparenz unterschieden – so verlangt die erste Richtlinie eine ausführliche Zitation sämtlicher verwendeter Daten, Programmcodes sowie anderer Methoden inklusive deren Identifikation anhand bspw. des DOI (Digital Object Identifier). Geordnet nach Transparenz *soll* diese Zitation nun entweder stattfinden (Level 1 an Transparenz), *muss* sie stattfinden (Level 2), oder es wird die Publikation der Studie *verweigert*, wenn diese Standards nicht anhand der Zitationen nachgewiesen werden können (Level 3). Die Richtlinien zwei bis vier behandeln das Ausmaß, in welchem Daten, Methoden und Materialien zugänglich gemacht werden. Richtlinie fünf beinhaltet das Ausmaß, in welchem Standards zu Design und Analyse befolgt werden sollen bzw. wurden. Richtlinie sechs und sieben beschäftigen sich mit der Präregistrierung der Studien und des Analyseplans. Erstere soll das Auffinden von Studien auch dann ermöglichen bzw. erleichtern, wenn diese nicht publiziert worden sind, zweite die Nachvollziehbarkeit der Studienart (explorativ oder konfirmativ) erleichtern. Die letzte und achte Richtlinie beschäftigt sich mit der Replikation von Studienergebnissen, d. h. wie weit diese gefördert, veröffentlicht und beurteilt wird. Eine auf der Homepage des Centers for Open Science zugänglich gemachte Liste enthält bis dato 526 Fachzeitschriften und 51 Organisationen, welche ihre Unterstützung hinsichtlich der genannten Richtlinien kundgetan haben (Stand 21.10.2015; https://cos.io/top/). Eine weitere Seite des Center for Open Science enthält eine Liste jener Zeitschriften, welche die Präregistrierung von Berichten bereits ermöglichen (Chambers et al., 2015). Auf diese Weise soll das Zurückhalten von negativen Studienergebnissen unterbunden werden. Auch soll die Interpretation von Ergebnissen nachvollziehbar gemacht werden, da Hypothesen nicht mehr post hoc an die Resultate angepasst werden können.

Um dem Argument, Transparenz sei zwar notwendig, aber kostenintensiv und zeitaufwendig, da sie mit extra Arbeit und wenig bis keinen Forschungsgeldern verbunden sei, entgegenzukommen, empfiehlt Buck (2015) die Implementierung qualitativ hochwertiger Open Source-Werkzeuge, wie sie bereits durch das Open Science Framework (https://osf.io/), iPython (http://ipython.org/) oder das Galaxy Project (https://galaxyproject.org/) angeboten werden.

Einen weiteren Incentive zur Quantifizierung von Qualität schlagen Hartshorne und Schachner (2012) vor. Ein eigenes Open-Access Journal soll Forschenden einen Anreiz für die Durchführung replikativer Studien bieten. Dieses würde eine Datenbank für Replikationsstudien beinhalten; die Qualität dieser Datenbank würde durch Crowdfunding und Peer-Review gesichert, und die Berechnung eines Replikationswerts anhand erfolgreich

und nicht-erfolgreich replizierter Studien ermöglichen. Dies entspräche einer Art Suchmaschine für Replikationen (replication tracker). Ähnlich der Nennung weiterer Artikel, wenn diese den gesuchten Artikel zitieren, wie es z. B. durch Google Scholar ermöglicht wird, würde der replication tracker jene Studien listen, welche die gewünschte Studie replizieren bzw. nicht replizieren konnten. Außerdem könnte anhand zweier Ratings angegeben werden, wie hoch der Anteil an erfolgreichen Replikationen ist, und wie sehr Ergebnisse daraufhin deuten, dass es sich um einen wahren Effekt handelt. Es würden hierbei ausschließlich strikte Replikationen verwendet werden, da primär herausgefunden werden soll, welche Ergebnisse und nicht welche Theorien korrekt sind.

1.10 Replikationsprojekte

Zahlen wie jene der 1.07%igen Replikationsrate in psychologischen Fachzeitschriften (Makel et al., 2012) lassen vermuten, dass Replikationen keine erstrebenswerte Art wissenschaftlicher Forschung darstellen (Madden et al., 1995). Eine Liste zu aktuell laufenden Replikationsprojekten der Zeitschrift Perspectives on Psychological Science findet sich auf der Homepage der Association for Psychological Science, abrufbar unter http://www.psychologicalscience.org/index.php/replication/ongoing-projects. Darüber hinaus bedarf es Studien, welche über die Replikation einzelner Effekte hinausgehen und eine Schätzung der Replizierbarkeit über das ganze Fachgebiet der Psychologie erlauben (vgl. Begley & Ellis, 2012). Das erste Projekt, welches beide dieser Punkte adressiert, ist das Many Labs-Projekt (Klein et al., 2014). Dieses verdankt seinen Namen den unterschiedlichen Forschungslaboratorien, in welchen eine Reihe von Experimenten von unterschiedlichen Wissenschaftler*innen in unterschiedlichen Settings und mit unterschiedlichen Stichproben durchgeführt worden sind. Bei den insgesamt 12 Studien (13 Effekte) handelte es sich um eine gezielte Auswahl unter Berücksichtigung von Design, Dauer, online-Tauglichkeit und Diversität (ein Teil der für das Projekt gewählten Effekte galt bereits als gut replizierbar). Ziel war die Erforschung des theoretischen Konstrukts der untersuchten Experimente über die Charakteristika von Primärstudie und –stichprobe, einschließlich kultureller Bedingungen hinaus. Die einzelnen Experimente der 13 Effekte wurden zu einem Ganzen fusioniert und am Computer von insgesamt 6,344 Proband*innen ausgeführt. 27 Datenansammlungen fanden in Labors statt, 9 online – hierbei 11 von 25 außerhalb der USA. Schlussendlich galten je nach Interpretation 10 bis 11 Effekte als repliziert; zwei Effekte zum Thema Priming wurden nicht (erfolgreich) repliziert.

Ein ähnliches Projekt, welches sich jedoch in Art und Umfang vom Many Labs-Projekt unterscheidet, stellt das im Folgenden beschriebene Reproducibility Project Psychology der Open Science Collaboration (2015) dar.

2 Reproducibility Project Psychology (OSC, 2012, 2014, 2015)

Das Reproducibility Project Psychology (RP:P) stellt das bis dato größte Replikations-Projekt der Psychologie dar und wurde 2011 von Brain Nosek ins Leben gerufen. Dieser nahm es sich zur Aufgabe, gemeinsam mit 270 teilnehmenden Autorinnen und Autoren aus verschiedensten Institutionen und Ländern dieser Welt, eine Schätzung der Replizierbarkeitsrate einer Reihe von Studien der psychologischen, genauer der sozial- und kognitionspsychologischen Forschung durchzuführen (Open Science Collaboration, 2015). Die Open Science Collaboration agierte mit Hilfe des Open Science Framework (https://osf.io/) des Center for Open Science, einer öffentlich zugänglichen Plattform anhand derer Planung, Durchführung sowie die Archivierung von Materialien, Daten und Ergebnissen von Beginn bis Ende des Projektes koordiniert wurden (Open Science Collaboration, 2012, 2014).

Im Fokus des RP:P stand die Replikation der veröffentlichten Studienergebnisse von insgesamt drei der renommiertesten aller psychologischen Fachzeitschriften. Anhand standardisierter Protokolle wurde das Vorgehen einzelner Autor*innen dokumentiert; auf diese Weise sollte eine möglichst hohe Transparenz und Qualität der einzelnen Replikationsstudien gewährleistet werden. Sämtliche im Rahmen der Replikation bedeutsamen Informationen wurden im Vorhinein ermittelt und festgehalten, darunter die Berechnung der für eine erfolgreiche Replikation der jeweiligen Studie benötigte Stichprobe anhand der gewählten Teststärke, die verwendeten Materialien, Skripte und schlussendlich auch Daten, Analyse und Ergebnisbericht. Die Ergebnisse der einzelnen Replikationsstudien wurden mit denen der Primärstudien verglichen, statistisch analysiert und kombiniert sowie interpretiert (Open Science Collaboration, 2012, 2014).

Ziel dieses Projekts war es einerseits aufzuzeigen, wie hoch die Replizierbarkeit psychologischer Forschungsarbeiten tatsächlich war, andererseits Aufschluss über mögliche Faktoren geben zu können, anhand derer die erfolgreiche oder nicht-erfolgreiche Replikation von Effekten erklärt werden konnte (Open Science Collaboration, 2015).

Es folgt eine detaillierte Beschreibung der Vorgehensweise des RP:P, sowie eine Demonstration am Bespiel der Replikation von Experiment 7 der Studie „Increasing and decreasing motor and cognitive output: A model of general action and inaction goals“ von Albarracín et al. (2008), welche dem Journal of Personality and Social Psychology entnommen wurde.

2.1 Methoden

Die für das Projekt gewählten Studien entstammen aus im Jahr 2008 veröffentlichten Forschungsarbeiten dreier prominenter Fachzeitschriften der psychologischen Forschung: Journal of Experimental Psychology: Learning, Memory, and Cognition, Journal of Personality and Social Psychology und Psychological Science. Die Forscher*innen der Open Science Collaboration konnten hiervon diejenige bzw. diejenigen Studien wählen, welche sie replizieren würden. Einschränkungen ergaben sich dadurch, dass nicht jede der veröffentlichten Studien zur Replikation freistand; dies war zum Beispiel der Fall, wenn die Einzigartigkeit der Stichprobe nicht rekonstruiert werden konnte, der Zugang zu spezieller benötigter Forschungsausrüstung nicht gegeben war oder historisch bedingte Umstände der untersuchten Phänomene deren Replikation nicht ermöglichen würden bzw. sinnvoll erscheinen ließen. Weiters wurde eine sukzessive Freigabe der Studien veranlasst, um auf diese Weise einem eventuell auftretenden Bias hinsichtlich der Wahl der Studien entgegenzuwirken (Open Science Collaboration, 2012, 2014).

Aus den hierdurch 158 wählbaren von 488 Artikel entstanden schließlich 113 Replikationen (111 Artikel); von diesen konnten aufgrund der Abgabefrist 100 Replikationen im finalen Bericht des RP:P berücksichtigt werden (Open Science Collaboration, 2015). In Summe entspricht dies einem Artikelanteil der jeweiligen Zeitschriften von 61% (Psychological Science), 56% (Journal of Personality and Social Psychology) und 72% (Journal of Experimental Psychology: Learning, Memory, and Cognition). In 84% der Fälle wurde die jeweils zuletzt berichtete Studie des Artikels repliziert; hierbei wiederum wurde sich nach Absprache mit den Autor*innen der Primärstudie für einen zentralen, zu replizierenden Effekt entschieden. Jede der Replikationsstudien musste eine Teststärke von mindestens 80% erreichen; die Größe der dazu benötigten Stichprobe wurde im Vorhinein errechnet und berichtet.

Zur Ermittlung und zum Vergleich der Replizierbarkeit der Ergebnisse wurden Signifikanz, *p*-Werte, Effektstärken, subjektive Einschätzungen der Replikationsautor*innen sowie Meta-Analysen herangezogen (Open Science Collaboration, 2015). Die Ergebnisse der Primär- und Replikationsstudien wurden in signifikante ($p \leq .05$) und nicht-signifikante ($p > .05$) untergeteilt. Nicht-signifikante Studienergebnisse, welche im Original als signifikant interpretierten worden waren, wurden in der Replikation ebenso gehandhabt. Die Verteilung der *p*-Werte der Primärstudien wurde mit jener der Replikationsstudien verglichen (Wilcoxon-Vorzeichen-Rang-Test bzw. *t*-Test für unabhängige Stichproben).

Effektstärken wurden, wenn möglich, in Korrelationskoeffizienten umgewandelt und anhand von vier Tests verglichen. Mit Hilfe eines Zweistichproben-*t*-Tests sowie Wilcoxon-

Vorzeichen-Rang-Tests wurde die zentrale Tendenz der Primär- und Replikationseffektstärken berechnet. Das Verhältnis der Studien, deren Effekt in der Primärstudie stärker ausfiel als in der Replikationsstudie, zu den Studien, wo das nicht der Fall war, wurde ermittelt, ebenso das Verhältnis jener Studienpaare, deren Primäreffekt im Konfidenzintervall des Replikationseffekts lag (Open Science Collaboration, 2015).

Weiters wurden Primär- und Replikationsergebnisse einer Meta-Analyse unterzogen, und anhand der subjektiven Einschätzung der Replikationsautor*innen, welche die Frage „Haben Ihre Ergebnisse den Primäreffekt repliziert“ mit ja oder nein beantworten mussten, verglichen (Open Science Collaboration, 2015).

2.2 Ergebnisse

Von 100 Primärstudien konnten 97 ein signifikantes Ergebnis berichten. Es wurden die durchschnittlichen Teststärken der Primärstudien berechnet, wonach 89 der Replikationsstudien ebenfalls ein positives Ergebnis vorweisen sollten. Tatsächlich taten das lediglich 35 Studien, d. h. 36.1% der Effekte wurden erfolgreich repliziert, 95% *KI* [26.6, 34.2].

Um dem Nachteil der streng dichotomen Unterteilung in erfolgreiche und nicht-erfolgreiche Ergebnisse anhand eines *p*-Werts von .05 zu beheben, wurden die Standardfehler der Korrelation bei 73 Studien berechnet; demnach lag innerhalb von 30 der Konfidenzintervallen der Replikationsstudien auch der Primäreffekt (41.1%). Für weitere 22 Studien wurden die Konfidenzintervalle mit Hilfe von $F(df_1 > 1, df_2)$ und c² berechnet; von diesen enthielten 68.2% den Primäreffekt. Kombiniert entspricht dies einer Gesamtrate von 47.4% erfolgreichen Replikationen. Nachteilig an dieser Art der Replikationskontrolle ist jedoch, dass ein replizierter Effekt, so er auch in dieselbe Richtung wie der Primäreffekt zeigt, nicht als repliziert gilt, wenn er signifikant kleiner ist dieser ist. Außerdem gilt eine Replikation auch dann als erfolgreich, wenn das Ergebnis zwar nahe Null liegt, jedoch nicht mit ausreichender Genauigkeit vom Primärergebnis unterschieden werden kann (Open Science Collaboration, 2015).

Insgesamt 82 der 99 Primärstudien, zu welchen Effektstärken berechnet werden konnten, zeigten eine größere Effektstärke als in der Replikationsstudie (Open Science Collaboration, 2015). Auch im Schnitt waren die Effektstärken der Primärstudien (M = 0.403, SD = 0.188) signifikant größer als jene der Replikationsstudien (M = 0.197, SD = 0.267), Wilcoxon’s W = 7137, $p < .001$.

Im Rahmen der anschließend berechneten Meta-Analyse gab es von den 75 für diese Analyseform geeigneten Effektstärken insgesamt 51 Effekte (68%), deren 95%-Konfidenzintervall den Nulleffekt nicht enthielt (Open Science Collaboration, 2015).

Die subjektive Einschätzung der Replizierbarkeit durch die Autor*innen der Replikationsstudien ähnelte der Unterteilung anhand der *p*-Werte und lag bei 39 erfolgreichen von insgesamt 100 Replikationen (Open Science Collaboration, 2015).

Zwischen den verschiedenen Forschungsbereichen dieser Studien zeigten sich ebenfalls Unterschiede. So konnten dem Bereich der kognitiven Psychologie zuzuordnende Studien zu 50% repliziert werden (21 von 42 Studien), Studien der Sozialpsychologie zu 25% (14 von 55 Studien). Vergleicht man Interaktionseffekte mit einfachen Effekten, zeigt sich, dass erstere zu 22% (8 von 37), letzte zu 47% (23 von 49) erfolgreich repliziert werden konnten. Studien, deren Primäreffekt mit einem *p*-Wert von < .02 signifikant ausfiel, konnten zu 41% (26 von 63) erneut einen *p*-Wert < .05 erreichen; Studien, deren Primäreffekt einen *p*-Wert < .001 aufwies, wurden zu 63% (20 von 32) erfolgreich repliziert (Open Science Collaboration, 2015).

2.3 Diskussion

Die Ergebnisse der fünf oben angewandten Messungen der Replizierbarkeit, mögen sie jede für sich genommen auch unzureichend sein, zeigen als Ganzes doch recht deutlich, dass die Anzahl erfolgreich replizierter Studien (35) stark von der aufgrund statistischer Berechnungen erwarteten Anzahl (89) abweicht (Open Science Collaboration, 2015). Es zeigte sich eine Tendenz, wonach Studien, deren Primärergebnis stark signifikant (geringer *p*-Wert) ausgefallen war, eher erfolgreich repliziert wurden.

Gründe, wieso viele der Ergebnisse nicht erfolgreich repliziert werden konnten, wurden bereits ausreichend erörtert (vgl. *Einleitung*). Es soll hier noch einmal darauf hingewiesen werden, dass ein nicht-replizierter Effekt mehrere Ursachen haben kann, und eine nicht-signifikante Replikation keinen Beweis eines nicht-vorhandenen oder fehlerhaft erlangten Effekts darstellt. Entscheidende Unterschiede in der Methodologie, geringe Teststärken oder Auswirkungen des Publikationsbias mögen zur selektiven Auswahl von Studien durch Fachzeitschriften beigetragen haben (Open Science Collaboration, 2015).

Anzumerken seien ebenso Limitierungen durch die Vorgehensweise des Reproducibility Project Psychology selbst (Open Science Collaboration, 2015). Es wurde jeweils nur ein statistischer Wert eines einzigen Experiments einer Studie zur Replikation herangezogen. Inwieweit selektive Einflüsse – trotz des Bemühens, diese durch strategisches Vorgehen bei der Auswahl der Artikel durch die Replikationsautor*innen gering

zu halten – sich auf das Ergebnis ausgewirkt haben, lässt sich hier nicht beantworten. Die durch dieses Projekt erlangten Erkenntnisse geben zuallererst Auskunft über die Replizierbarkeit von Forschungsergebnissen als Ganzes, jedoch nur ungenügend Information über die einzelnen Effekte, zu deren Erforschung mehr als eine Replikation nötig gewesen wäre (vgl. Klein et al., 2014). Ergebnisse wurden durch dieses Projekt weder als wahr noch als falsch klassifiziert – dies war weder möglich noch Ziel. Vielmehr sei dieses Projekt als Aufruf zu verstehen, wonach sich die Wissenschaft wieder ihrer Wurzeln besinnen solle, nämlich der Wiederholbarkeit von Forschungsergebnissen genüge zu tragen, um ein solides Fundament aus abgesicherten Resultaten zu schaffen. Die Open Science Collaboration (im Druck) gibt über dieses Projekt hinaus Anleitung dafür, die eigene Forschungsarbeit so zu gestalten, dass diese für eine mögliche Replikation optimal genutzt werden kann.

3 Albarracín et al. (2008)

In einem Versuch, die Auswirkungen von Priming auf Aktivität zu erfassen, bedienten sich Albarracín et al. (2008) der Konzepte der *generellen Aktion* bzw. der *generellen Inaktion*. Unter genereller Aktion werden unterschiedliche Arten motorischer als auch kognitiver Leistung verstanden, wie etwa das Aneignen von Wissen, aber auch das Kritzeln auf einem Blatt Papier. Hierbei kann es sich sowohl um wichtige und gezielt ausgeführte Aktivitäten handeln, als auch um solche, die auf den ersten Blick weniger sinnvoll erscheinen oder aber keine besondere Anstrengung erkennen lassen. Aktion und Inaktion werden außerdem als kontinuierliche Ausprägung ein und derselben Variable (Aktivität) betrachtet. Ausgangsbasis der durchgeführten Experimente ist die Vermutung, dass sich der Level an Aktivierung unabhängig von der Art der Aktivität durch ein Ziel regulieren lässt. Das Priming von Aktivität sollte gemäß Albarracín et al (200) in einem Aktionsziel resultieren, das Priming von Inaktivität in einem Inaktionsziel. Beispielsweise soll das Priming von Aktivität dazu führen, dass die Versuchsperson unter mehreren Aufgaben eine aktive Aufgabe wählt, und in weiterer Folge mehr kognitive oder motorische Leistung zeigt, als dies bei einer inaktiven Aufgabe der Fall gewesen wäre.

Auf Basis dieses Konzepts wollten Albarracín et al. (2008) nun zeigen, dass Menschen sich zuerst für ein Verhalten (motorische oder kognitive Leistung) entscheiden, und erst später Art und Qualität der Ergebnisse dieses Verhaltens prüfen, d. h. nicht notwendigerweise sinnvolle Leistungen anstreben. Außerdem wurde ein top-down Mechanismus angenommen, welcher sowohl triviale als auch nicht-triviale Leistung durch dieselben Prozesse steuert. Inaktion wird bei alldem als erwünschtes Endstadium und nicht als gescheiterte Aktion verstanden.

Um Effekte genereller Aktion und Inaktion zu demonstrieren, sollte gezeigt werden, dass sich das Priming von aktiven und inaktiven Wörter entsprechend auf die Wahl weiterer Aktivitäten auswirkt (Albarracín et al., 2008). Versuchspersonen, welche mit dem Konzept der generellen Aktion geprimt worden waren, sollten in Folge dessen eine aktive Aufgabe wählen, waren sie hingegen mit Inaktion geprimt worden, sollten sie inaktiven Aufgaben (z. B. eine Pause machen) den Vorzug geben. Außerdem bestand die Möglichkeit, dass die anhand der gewählten Aufgabe erbrachte motorische bzw. kognitive Leistung stärker ausfallen würde, wenn zuvor das Konzept genereller Aktion geprimt worden war.

In der Erwartung, dass in Folge geprimter Aktion mehr motorische Aktivität vorhanden sein würde als bei geprimter Inaktion, sollte das erste Experiment von Albarracín et al. (2008) zeigen, dass Priming anhand von Wörtern des Konzeptes Aktion (z. B. „gehen“) mehr Versuchspersonen dazu bewegen würde, ein Papierflugzeug zu basteln oder auf einem Blatt

zu kritzeln, als das Priming inaktiver Wörter (z. B. „rasten“). In Experiment 2 sollte das Priming aktiver Wörter zu mehr motorischer Leistung (dem Essen von Trauben) führen, als das Priming inaktiver Wörter. Das dritte Experiment sollte anhand einer Videoaufnahme, in welcher Schüler*innen beim Ausführen diverser täglicher Aktivitäten zu sehen waren, zeigen, dass die mit aktiven Wörter geprimten Versuchspersonen mehr bedeutungsvolles Verhalten identifizieren konnten (gemessen an der Verwendungshäufigkeit der Leertaste), als Personen, die mit inaktiven Wörter geprimt worden waren. Diese drei Experimente sollten außerdem zeigen, dass die durch das Priming genereller Aktion hervorgebrachten Prozesse motorische Leistung stärker fördern würden, als die durch das Priming von Inaktion hervorgebrachten Prozesse. In Experiment 4 sollte supraliminales Priming aktiver Wörter (im Vergleich zu inaktiven Wörter) zu einem höheren Anteil korrekt abgerufener Inhalte eines zuvor gelesenen Textes führen; in Experiment 5 zu einem höheren Anteil richtig gelöster verbaler und mathematischer Aufgaben. Experiment 4 und 5 sollten somit die Effekte von Priming auf kognitive Leistungen zeigen. Die letzten beiden Experimente (6 und 7) sollten zeigen, dass generelle Ziele die Effekte von aktivem und inaktivem Priming mediieren würden. Die Ergebnisse aller Experimente konnten diese Vermutungen bestätigen. Die Effekte von Experiment 7 – Effekte der Befriedigung geprimter Ziele auf das Level an Aktivierung – sollen nun im Rahmen des Reproducibility Project Psychology (Open Science Collaboration, 2015) repliziert werden. Aufbau und Ablauf des Experiments in der Primär- sowie der Replikationsstudie werden im Folgenden beschrieben.

3.1 Methoden

Anhand von Experiment 7 soll demonstriert werden, dass Versuchspersonen über eine höhere Aktivität aufweisen, wenn das Konzept Aktion geprimt worden war, und eine niedrigere Aktivität, wenn das Konzept Inaktion geprimt worden war. Ein Aktionsziel ist immer dann vorhanden, wenn Versuchspersonen anhand aktiver Wörter geprimt werden und im Anschluss daran nicht aktiv handeln können, d. h. eine inaktive Aufgabe bewältigen müssen. Ein Inaktionsziel ist dann vorhanden, wenn Versuchspersonen mit inaktiven Wörtern geprimt worden sind und danach nicht inaktiv handeln können, d. h. eine aktive Aufgabe durchführen müssen. Sollten Primingeffekte zielorientiert sein, würde ein unbefriedigtes Aktionsziel demzufolge zu mehr Aktivität führen, als ein befriedigtes Aktionsziel. Umgekehrt wäre es bei einem unbefriedigten Inaktionsziel – dieses würde zu geringerer Aktivität als ein befriedigtes Inaktionsziel führen.

3.1.1 Stichprobe

Die Stichprobe setzte sich zusammen aus 98 Versuchspersonen: Studierende der University of Florida, welche als Gegenleistung für ihre Studienteilnahme Punkte für einen

Psychologie-Einführungskurs erhalten haben. Es wurde ein 3 (Priming: Aktion, Inaktion, neutral) x 2 (Aufgabe: aktiv, inaktiv) faktorielles Design verwendet; die Versuchspersonen wurden je einer dieser sechs Bedingungen zugewiesen. Der Aufbau des Experiments gestaltete sich wie folgt.

3.1.2 Wortvervollständigungstest

Erste Aufgabe der Versuchspersonen war es, eine Wortvervollständigungsaufgabe durchzuarbeiten. Die Versuchspersonen wurden je nach Bedingung mit aktiven, inaktiven oder neutralen Kontrollwörtern geprimt. Danach sollte jede Person insgesamt zwanzig Wortanfänge vervollständigen; bei je acht von diesen handelte es sich um die Anfänge aktiver oder inaktiver Wörter (aktiv: „motivation“, „doing“, „behavior“, „engage“, „action“, „make“, „go“, „active“ bzw. inaktiv: „still“, „pause“, „interrupt“, „calm“, „freeze“, „unable“, „stop“, „paralyze“), bei den restlichen 12 um Kontrollwörter. Die Personen der Kontrollbedingung erhielten ausschließlich Kontrollwörter. Es wurden jeweils der erste oder die ersten beiden Buchstaben des Wortes präsentiert. Die Vervollständigung dieses Wortes (oder eines anderen, passenden Wortes) konnte folgendermaßen geschehen: Aus den Anfangsbuchstaben „ma“ konnte beispielsweise das Wort „make“ gebildet werden (aktive Bedingung), aus „st“ „stop“ (inaktive Bedingung) und aus „la“ für „lamp“ (Kontrollbedingung). Diese Wortanfänge wurden den Versuchspersonen einzeln und in zufälliger Reihenfolge vorgegeben. Wörter mit hohen Assoziationen zu Aktivität bzw. Inaktivität wurden anhand des Computerized Edinburgh Associative Thesaurus (Kiss, Armstrong, Milroy, & Piper, 1973) bestimmt; die Wörter der Kontrollgruppe hatten keine besondere Assoziation zu Aktivität oder Inaktivität. Die Wortvervollständigungsaufgabe wurde den Versuchspersonen am Computer vorgegeben. Diese Form des Priming entsprach der ersten der unabhängigen Variablen.

3.1.3 Aufgabe

Im Anschluss an die Wortvervollständigungsaufgabe wurde den Versuchspersonen eine aktive oder inaktive Aufgabenstellung präsentiert. Inhalt der aktiven Aufgabenstellung war es, die Augen zu schließen und den Kopf „leer“ zu bekommen; im Rahmen der inaktiven Aufgabestellung mussten die Versuchspersonen auf einem bereitgelegten Blatt Papier zeichnen, oder aus diesem ein Papierflugzeug basteln. Nach Ablauf von zwei Minuten wurde diese Aufgabe beendet. Diese Aufgabenstellung entsprach der zweiten der unabhängigen Variablen.

3.1.4 Text über Vegetarismus

Nach der Durchführung der aktiven bzw. inaktiven Aufgabe wurden die Versuchspersonen gebeten einen Text zu lesen, und anschließend ihre Gedanken zu diesem Text zu notieren. Der Text beinhaltete Informationen zum Thema Vegetarismus; es handelte sich hierbei um einen Auszug aus White und Frank (1994). Im Anschluss wurden die Versuchspersonen angehalten, ihre Gedanken zum eben gelesenen Text in einem dafür vorgesehenen Feld zu notieren, wobei jeder Gedanke einer Zeile entsprechen sollte. Die Anzahl der Gedanken diente hierbei als abhängige Variable.

3.1.5 Stimmung

Zum Abschluss der Studie wurde die Stimmung der Versuchspersonen erhoben; dazu wurden diese gefragt, wie glücklich bzw. wütend sie sich gerade fühlten („Right now, do you feel happy?" bzw. "Right now, do you feel angry?"), und sie konnten ihr Antwort anhand einer 9-stufigen Skala (1 = not at all, 9 = extremely) angeben.

3.2 Ergebnisse

Eine ANOVA verwendend, kamen Albarracín et al. (2008) zu folgenden Ergebnissen: Es wurde eine signifikante zweifache Interaktion (p = .02) der Variablen Priming und Aufgabe nachgewiesen, $F(2, 92) = 4.36$, sowie ein signifikanter Haupteffekt (p = .007) für Aufgabe, $F(1, 92) = 7.57$. Die Variable Priming zeigte keinen signifikanten Effekt, $F(2, 92) = 1.88$.

Die Berechnung der Kontraste ergab eine signifikant höhere Gedankenanzahl bei Aktion-Priming gefolgt von einer inaktiven Aufgabe, d. h. einem unbefriedigten Aktionsziel (M = 6.94, SD = 3.46) als Inaktion-Priming gefolgt von einer aktiven Aufgabe, d. h. einem unbefriedigten Inaktionsziel (M = 3.36, SD = 0.81), mit p = .001. Diese beiden Bedingungen wiederrum zeigten signifikante Unterschiede (je p = .02) zur jeweiligen Kontrollbedingung (M = 4.17, SD = 2.04 bei inaktiver Aufgabe; M = 5.00, SD = 2.24 bei aktiver Aufgabe). Diese Ergebnisse wurden von Albarracín et al. (2008) als Replikation der Ergebnisse der vorherigen Experimente gesehen.

Für einen Effekt der Befriedigung aktiver bzw. inaktiver Ziele sprachen weitere Kontraste. Es wurden bei Inaktion-Priming gefolgt von einer aktiven Aufgabe, d. h. einem unbefriedigten Inaktionsziel (M = 3.36, SD = 0.81) signifikant weniger Gedanken zum Text notiert (p = .006) als bei einer inaktiven Aufgabe, d. h. einem befriedigten Inaktionsziel (M = 6.50, SD = 2.80). Dazu passend zeigte sich eine signifikant höhere Gedankenanzahl (p = .02) bei Aktion-Priming gefolgt von inaktiven Aufgaben, d. h. einem unbefriedigten Aktionsziel (M = 6.94, SD = 3.46) im Kontrast zu Aktion-Priming gefolgt von aktiven Aufgaben, d. h. einem befriedigten Aktionsziel (M = 4.58, SD = 2.76).

Vergleicht man weiter, zeigt sich eine signifikant höhere Gedankenanzahl bei Aktion-Priming in Kombination mit aktiver Aufgabe, d. h. einem befriedigten Aktionsziel (M = 4.58, SD = 2.76), als bei Inaktion-Priming gefolgt von einer aktiven Aufgabe, d. h. einem unbefriedigten Inaktionsziel (M = 3.36, SD = 0.81, p = .05), und als bei den Kontrollbedingungen (p = .02). Ein befriedigtes Inaktionsziel zeigt hingegen eine ähnliche Gedankenanzahl (M = 6.50, SD = 2.80) wie ein unbefriedigtes Aktionsziel (M = 6.94, SD = 3.46), p = .72. Diese beiden Bedingungen zeigten wiederum eine höhere Gedankenanzahl als der Durchschnitt der beiden Kontrollbedingungen (M = 4.58, SD = 2.76), p = .001.

3.3 Diskussion

Die Ergebnisse der Experimente von Albarracín et al. (2008) unterstützen die Vermutung, dass sich Versuchspersonen um die Befriedigung aktiver bzw. inaktiver Ziele bemühen, wenn zuvor Aktion bzw. Inaktion geprimt wurde. Dieser Theorie folgend verhielt sich auch die Anzahl der von den Versuchspersonen notierten Gedanken, wenn die angebotene Aufgabe nicht dem zuvor geprimten Konzept entsprach; hier schien der Effekt durch Priming verzögert aufzutreten. Somit konnte demonstriert werden, dass unbefriedigte Aktionsziele zu mehr aktivem Verhalten (mehr notierten Gedanken) führen als befriedigte Aktionsziele. Ebenso führen unbefriedigte Inaktionsziele zu weniger aktivem Verhalten (weniger notierten Gedanken) als befriedigte Inaktionsziele.

4 Replikation von Albarracín et al. (2008)

Im Rahmen des Reproducibility Project Psychology wurde das siebte Experiment von Albarracín et al. (2008) repliziert. Als zu untersuchender Effekt – hierbei musste gemäß der Richtlinien des RP:P ein einziger, zentraler Effekt bestimmt werden – wurde der Interaktionseffekt von Priming und Aufgabe gewählt. Zusätzlich sollte der Haupteffekt Aufgabe bestimmt und eine Kontrastanalyse unter Berücksichtigung der Stimmung vorgenommen werden.

4.1 Methoden

4.1.1 Analyse der Teststärke

Sämtliche Replikationsstudien des Reproducibility Project Psychology mussten eine Teststärke von mindestens 80% aufweisen. Um die Anzahl an benötigten Versuchspersonen zu berechnen, die die jeweilige Teststärke von 80%, 90% bzw. 95% garantieren sollte, wurde die Software *G*Power 3* (Faul, Erdfelder, Lang, & Buchner, 2007) verwendet. Als Testfamilie wurde „*F*-tests“ und als statistischer Test „ANOVA: Fixed effects, special, main effects and interactions“ gewählt; die Art der Teststärkenanalyse war „a priori“.

Zur Kalkulation der Effektstärke der Primärstudie wurde Daniel Lakens' Tabelle zur Berechnung von Effektstärken verwendet (Lakens, 2013). Dies ergab eine Effektstärke von η_p^2 = .087 für den von Albarracín et al. (2008) berichteten Interaktionseffekt von Priming und Aufgabe, $F(2, 92) = 4.36$, p = .02. Mit Hilfe von G*Power 3 wurde diese Effektstärke konvertiert und als f = .309 zur weiteren Berechnung verwendet. Die Fehlerwahrscheinlichkeit wurde mit .05 angegeben, der Zähler der Freiheitsgrade *df* mit 2 und die Anzahl der Gruppen mit 6. Hieraus ergab sich eine Anzahl von 165 Versuchspersonen für eine 95%ige Teststärke, 136 Versuchspersonen für eine 90%ige und 105 für eine 80%ige Teststärke.

Eine weitere Teststärkenanalyse wurde vorgenommen, um sichergehen zu können, mit 165 Versuchspersonen ebenfalls einen signifikanten Haupteffekt für die Variable Aufgabe erreichen zu können. Dazu wurden dieselben Angaben wie bei der Berechnung des Interaktionseffekts verwendet. Basierend auf dem berichteten Effekt für Aufgabe, $F(1, 92) = 7.56$, p = .007 (Albarracín et al., 2008) lässt sich mit Hilfe von Lakens' Tabelle (Lakens, 2013) eine Effektstärke von η_p^2 = .076 berechnen. Mit G*Power 3 wurde diese zu f = .2868 konvertiert. Die Anzahl der Freiheitsgrade betrug 1. Für eine Teststärke von 95% war somit eine Anzahl von 160 Versuchspersonen nötig. Daher sollte die für den Interaktionseffekt berechnete Anzahl von 165 Versuchspersonen hinreichend sein.

4.1.2 Versuchspersonen

Insgesamt wurden schlussendlich 109 Versuchspersonen per LABS[2], E-Mailverteiler und Flyer rekrutiert. Die Daten von 4 Personen wurden aufgrund fehlerhafter Speicherung von der weiteren Analyse ausgeschlossen; somit bestand die untersuchte Stichprobe aus 105 Personen (58.1% Frauen), davon 61 Psychologie-Studierende. Das durchschnittliche Alter entsprach einem Mittelwert von 25.93 (SD = 7.703); außerdem gaben 22.9% der Versuchspersonen an, sich vegetarisch oder vegan zu ernähren. Jede Versuchsperson erhielt als Entschädigung für die Studienteilnahme 5 Euro in bar ausbezahlt.

4.1.3 Versuchsplan

Der Studie lag ein 3 (Priming: Aktion, Inaktion, neutral) x 2 (Aufgabe: aktiv, inaktiv) faktorielles Design zugrunde. In jeder Bedingung entsprach die abhängige Variable der Anzahl an notierten Gedanken zum gelesenen Text.

[2] System zur Anmeldung an psychologischen Studien zum Erlangen von Kurspunkten für Psychologie-Student*innen der Universität Wien.

4.1.4 Material und Durchführung

Die Versuchspersonen wurden je einer von insgesamt 6 Bedingungen zugeteilt. Schlussendlich ergab sich folgende Einteilung: Die Bedingungen Aktion-Priming/aktive Aufgabe, Inaktion-Priming/inaktive Aufgabe sowie Kontroll-Priming/aktive Aufgabe enthielten je 18 Versuchspersonen, die Bedingung Aktion-Priming/inaktive Aufgabe enthielt 19 Versuchspersonen, die Bedingung Inaktion-Priming/aktive Aufgabe enthielt 17 und die Bedingung Kontroll-Priming/inaktive Aufgabe 15 Versuchspersonen. Die sechs Versionen des Experiments wurden mit der Experimentalsoftware *E-Prime Version 2.0* (Psychology Software Tools, Inc., 2012) erstellt; die Daten anschließend mit *IBM SPSS Statistics 20* (IBM Corp., 2011) analysiert. Die gesamte Testung erfolgte am Computer und wurde unter Anleitung und in Anwesenheit der Testleiterin durchgeführt. Jede der Versuchspersonen musste zu Beginn des Experiments einen Kopfhörer aufsetzen, über welchen sie anhand eines Glockentons über das Ende der sich im Experiment befindlichen Aufgabe informiert wurde. Der vollständige Ablauf der Testung, sowie die Übersetzung der Originalversion ins Deutsche, befinden sich im Anhang (Appendix A. 1).

4.1.4.1 Wortvervollständigungstest

Zuerst wurden die Versuchspersonen mithilfe zu vervollständigender Wörter entsprechend der jeweiligen Bedingung (aktiv, inaktiv oder Kontrollbedingung) geprimt. Insgesamt mussten sie zwanzig einzeln vorgegebene Wortanfänge (z.B. „ge“ für „gehen“ in der aktiven Bedingung, „un“ für „unfähig“ in der inaktiven Bedingung, „Li“ für „Licht“ in der Kontrollbedingung) vervollständigen. Es handelte es sich hierbei je nach Bedingung um 8 Aktionswörter bzw. 8 Inaktionswörter mit 12 Kontrollwörtern, oder 20 Kontrollwörter. Die Auflistung der von Albarracín et al. (2008) verwendeten Wörter sowie deren Übersetzung ins Deutsche können dem Anhang entnommen werden (Appendix A. 2).

4.1.4.2 Aufgabe

Im Anschluss an die Wortvervollständigungsaufgabe erhielten die Versuchspersonen die Instruktion, eine zweiminütige Pause einzulegen, welche sie – je nach Instruktion – entweder aktiv oder inaktiv gestalten sollten. Die aktive Aufgabenstellung bestand darin, auf einem bereitgelegten Blatt Papier zu zeichnen oder ein Papierflugzeug aus diesem zu basteln; in der inaktiven Aufgabenstellung wurden die Versuchspersonen gebeten ihre Augen zu schließen und den Kopf „frei“ zu bekommen. Ein kurzer Glockenton über die zu Beginn des Experiments angebrachten Kopfhörer informierte die Versuchsperson über das Ende der Pause.

4.1.4.3 Text über Vegetarismus

Nach der zweiminütigen Pause wurden die Versuchspersonen gebeten einen Auszug aus dem Artikel „Health effects and prevalence of vegetarianism“ (White & Frank, 1994) zu lesen und darauf hingewiesen, dass sie im Anschluss an den gelesenen Text ihre Gedanken zu diesem notieren werden würden. Nach dem Lesen des Textes wurden die Versuchspersonen erneut darauf hingewiesen, ihre Gedanken zum Text zu notieren und dabei jeden einzelnen Gedanken extra zu kennzeichnen (Appendix A. 3).

4.1.4.4 Ergänzungen

Zum Abschluss wurden die Versuchspersonen gebeten, ihre aktuelle Stimmung auf einer 9-stufigen Skala einzuschätzen, sowie Fragen nach Alter und Geschlecht zu beantworten. Außerdem wurden die Versuchspersonen gefragt, ob sie sich vegetarisch oder vegan ernährten („Ernährst du dich vegetarisch oder vegan?“) und ob sie Psychologie-Student*innen waren (Appendix A. 1).

4.2 Ergebnisse

Um mögliche Unterschiede in der Anzahl notierter Gedanken in Abhängigkeit der jeweiligen Priming-Wörter (Aktion, Inaktion, neutral) in Kombination mit der Aufgabe (aktiv, inaktiv) festzustellen, wurde eine ANOVA verwendet (Abbildung 1). Es zeigte sich keine signifikante Interaktion der beiden Faktoren Priming und Aufgabe, $F(2, 103) = 2.601$, $p = .079$ ($\eta_p^2 = .048$), auch keine signifikanten Haupteffekte Priming, $F(2, 103) = 0.043$, $p = .957$ ($\eta_p^2 = .001$) oder Aufgabe, $F(1, 103) = 0.020$, $p = .887$ ($\eta_p^2 < .001$).

Weiters wurde eine zweifaktorielle ANOVA der Faktoren Aufgabe und Priming-kombiniert durchgeführt, wobei der Faktor Priming-kombiniert sich hier in zwei Gruppen unterteilt (Priming versus Kontrollgruppe). Es wurde keine signifikante Interaktion der Faktoren Priming-kombiniert und Aufgabe gefunden, $F(1, 105) = 2.890$, $p = .092$ ($\eta_p^2 = .027$), ebenso keine Haupteffekte Priming-kombiniert, $F(1, 105) = 0.038$, $p = .847$ ($\eta_p^2 = .000$), und Aufgabe, $F(1, 105) = 0.518$, $p = .473$ ($\eta_p^2 = .005$).

Eine Kontrastanalyse wurde durchgeführt, um mögliche Effekte der Stimmung der Versuchspersonen zu berücksichtigen. Verglichen wurde die Gedankenanzahl der Bedingungen Aktion-Priming/inaktive Aufgabe mit Aktion-Priming/aktive Aufgabe, Inaktion-Priming/aktive Aufgabe und Inaktion-Priming/inaktive Aufgabe für diejenige Personen, welche eine positive oder negative Stimmung angaben. Es wurde kein signifikanter Effekt gefunden, $F(1, 94) = 0.017$, $p = .897$ ($\eta_p^2 = .000181$). Für die Gruppe der Personen mit berichteter negativer Stimmung konnte bei einer Anzahl von 4 Versuchspersonen keine Kontrastanalyse vorgenommen werden.

Für die ergänzende Analyse des Faktors Vegetarismus/Veganismus wurde wiederum eine ANOVA verwendet. Diese konnte keine signifikanten Unterschiede zwischen Vegetarier*innen/Veganer*innen und Nicht-Vegetarier*innen/Nicht-Veganer*innen feststellen, $F(1, 107) = 1.419$, $p = .236$ ($\eta_p^2 = .013$).

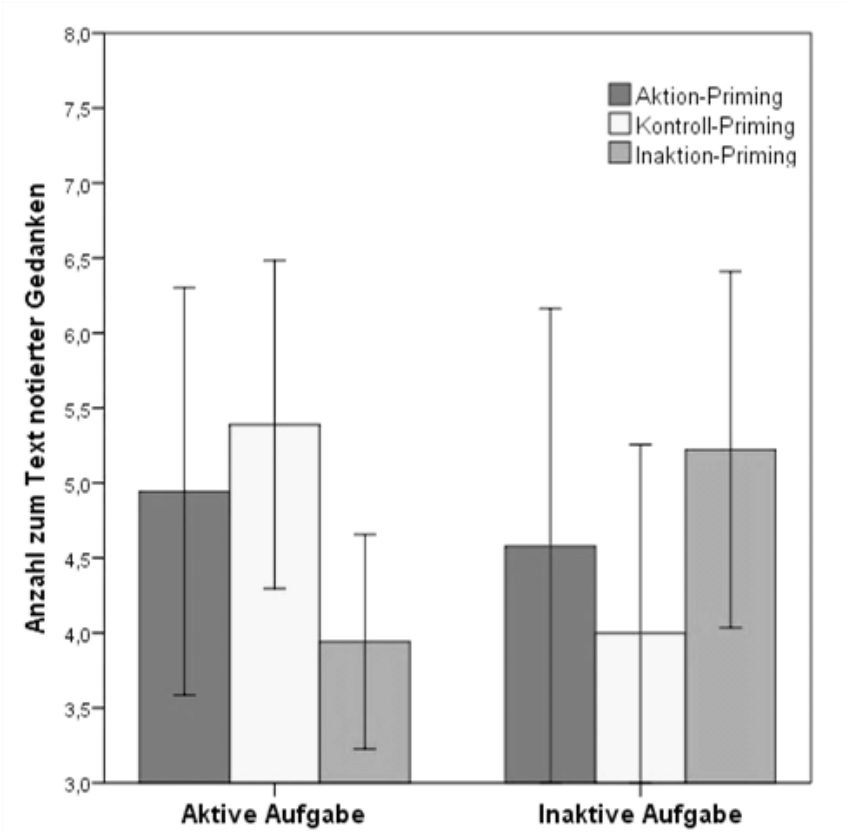

Abbildung 1. Auswirkung von Priming und Aufgabe auf Aktivität (Anzahl notierter Gedanken). Fehlerbalken stellen das 95%-Konfidenzintervall dar.

4.3 Diskussion

Die von Albarracín et al. (2008) postulierten Ergebnisse konnten im Rahmen dieser Replikation nicht nachgewiesen werden. Weder der Interaktionseffekt zwischen Priming und Aufgabe, noch die Haupteffekte Priming und Aufgabe zeigten signifikante Ergebnisse. Die durchgeführte Kontrastanalyse ergab unter Berücksichtigung der positiven Stimmung ebenfalls keine signifikanten Unterschiede zwischen den untersuchten Bedingungen.

4.4 Meta-Analyse

Die Ergebnisse der Replikationsstudie (2015) sowie der Primärstudie (Albarracín et al., 2008) wurden einer Meta-Analyse unterzogen; diese Berechnungen waren nicht Teil des Reproducibility Project Psychology.

Bei der Meta-Analyse wurde den Fokus auf die Frage nach der Befriedigung von generellen Zielen gelegt: So zeigte sich bei Albarracín et al. (2008), dass ein unbefriedigtes im Vergleich zu einem befriedigten Aktionsziel in mehr Aktivität resultiert (Aktion-Priming/inaktive Aufgabe vs. Aktion-Priming/aktive Aufgabe). Ähnlich folgte auf ein unbefriedigtes Inaktionsziel weniger Aktivität als auf ein befriedigtes Inaktionsziel (Inaktion-Priming/aktive Aufgabe vs. Inaktion-Priming/inaktive Aufgabe). Sämtliche Berechnungen erfolgten mit R Studio (R Studio Team, 2015); das Skript hierzu befinden sich im Anhang (Appendix A. 4).

Hypothese 1: Ein unbefriedigtes Aktionsziel (Aktion-Priming + inaktive Aufgabe) führt zu einer höheren Aktivität (Anzahl notierter Gedanken) als ein befriedigtes Aktionsziel (Aktion-Priming + aktive Aufgabe)

Hypothese 2: Ein unbefriedigtes Inaktionsziel (Inaktion-Priming + aktive Aufgabe) führt zu einer niedrigeren Aktivität (Anzahl notierter Gedanken) als ein befriedigtes Inaktionsziel (Inaktion-Priming + inaktive Aufgabe).

Anhand eines Random Effects Model, genauer des DerSimonian Laird-Tests, wurden die Ergebnisse dieser Fragestellungen aus Albarracín et al. (2008) mit den entsprechenden Replikationsergebnissen verglichen. Da die Stichprobenverteilung auf die einzelnen Gruppen der Primärstudie nicht bekannt ist, wurde diese per Münzwurf entschieden, wobei auf eine möglichst ausgewogene Stichprobengröße geachtet wurde. Folgende Zuteilung wurde ermittelt: Aktion-, Inaktion- und Kontroll-Priming in Kombination mit inaktiver Aufgabe sowie für Aktion-Priming mit inaktiver Aufgabe je n = 16, Aktion- und Kontroll-Priming in Kombination mit aktiver Aufgabe je n = 17.

4.4.1 Ergebnisse

Die Meta-Analyse zu Hypothese 1 ergab einen nicht-signifikanten Summeneffekt (standardisierte Mittelwertdifferenz) von d = 0.299 (SD = 0.427), 95%-Konfidenzintervall [-0.538, 1.136], p = .484. Für Hypothese 2 wurde ein signifikanter Summeneffekt (standardisierte Mittelwertdifferenz) von d = -1.037 (SD = 0.424), p = .015 berechnet. Das 95%-Konfidenzintervall [-1.869, -0.205] beinhaltet somit keinen Nulleffekt (Abbildung 2).

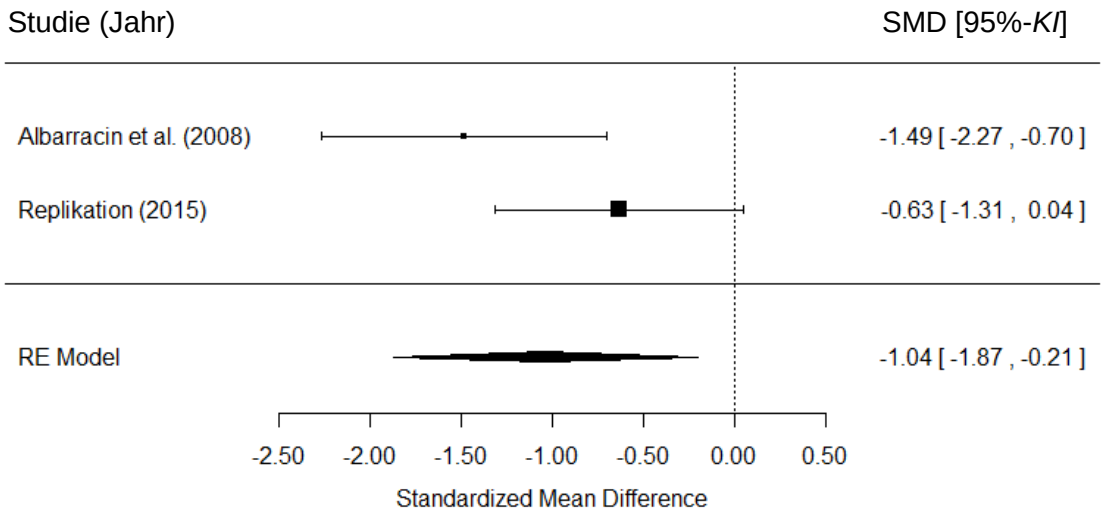

Abbildung 2. Forest-Plot zu Hypothese 2, p = .015

4.4.2 Diskussion

Das nicht signifikante Ergebnis von Hypothese 1 liegt nahe dem Nulleffekt und deutet in die entgegengesetzte Richtung der Hypothese. Hypothese 2 zeigt hingegen ein signifikantes Ergebnis; die beiden Effekte deuten also in dieselbe Richtung. Das Konfidenzintervall des Replikationsergebnisses beinhaltet jedoch auch den Nulleffekt. In Anbetracht der nicht signifikanten Interaktions- und Haupteffekte, sowie der ebenfalls nicht signifikanten Kontrastanalyse, sollte dieses Ergebnis jedoch mit Vorsicht interpretiert werden. Weitere Replikationsstudien sind nötig, um ein eindeutiges Ergebnis bestimmen zu können.

4.5 Small Telescope

Der Small Telescope-Ansatz von Simonsohn (2015) –diese Berechnung ist nicht Teil des Reproducibility Project Psychology – sieht vor, anhand der Ergebnisse der Primärstudie einen sehr kleinen Effekt zu berechnen, welcher der Studie eine Teststärke von 33% bemessen würde: eine entsprechend kleine Effektstärke d von zum Beispiel 0.2 würde als $d_{33\%}$ = 0.2 bezeichnet werden. Auf dieser Grundlage soll nun errechnet werden, ob die Teststärke der Primärstudie groß genug gesetzt wurde, um einen möglichen Effekt überhaupt entdeckt haben zu können. Erhält eine Replikationsstudie einen Effekt, welcher signifikant kleiner ist als der Effekt $d_{33\%}$ der Primärstudie, spricht dies gegen die Annahme, der berichtete Effekt wäre anhand der Stichprobe der Primärstudie auffindbar gewesen.

Mit Hilfe der von Simonsohn (2014) zur Verfügung gestellten Anleitung und Skripten (zu finden unter https://osf.io/adweh/) wurde für folgende Hypothesen die Effektstärke $d_{33\%}$ berechnet (adaptiertes Skript für R in Appendix A. 5):

Hypothese 1: Ein unbefriedigtes Aktionsziel (Aktion-Priming + inaktive Aufgabe) führt zu einer höheren Aktivität (Anzahl notierter Gedanken) als ein befriedigtes Aktionsziel (Aktion-Priming + aktive Aufgabe)

Hypothese 2: Ein unbefriedigtes Inaktionsziel (Inaktion-Priming + aktive Aufgabe) führt zu einer niedrigeren Aktivität (Anzahl notierter Gedanken) als ein befriedigtes Inaktionsziel (Inaktion-Priming + inaktive Aufgabe).

4.5.1 Ergebnisse

Die berechnete Effektstärke $d_{33\%}$ für Hypothese 1 beträgt 0.549; wäre dies die wahre Effektstärke, bestünde eine 9.5%ige Chance, dass die Testung der Mittelwertunterschiede (bei n = 19 und n = 18) eine gleich kleine oder kleinere Effektstärke d ergeben hätte, als von der Replikation (2015) berichtet wurde. Die Nullhypothese, der Effekt $d_{33\%}$ war anhand der Primärstudie (Albarracín et al., 2008) zu entdecken und wird daher beibehalten, p = .095.

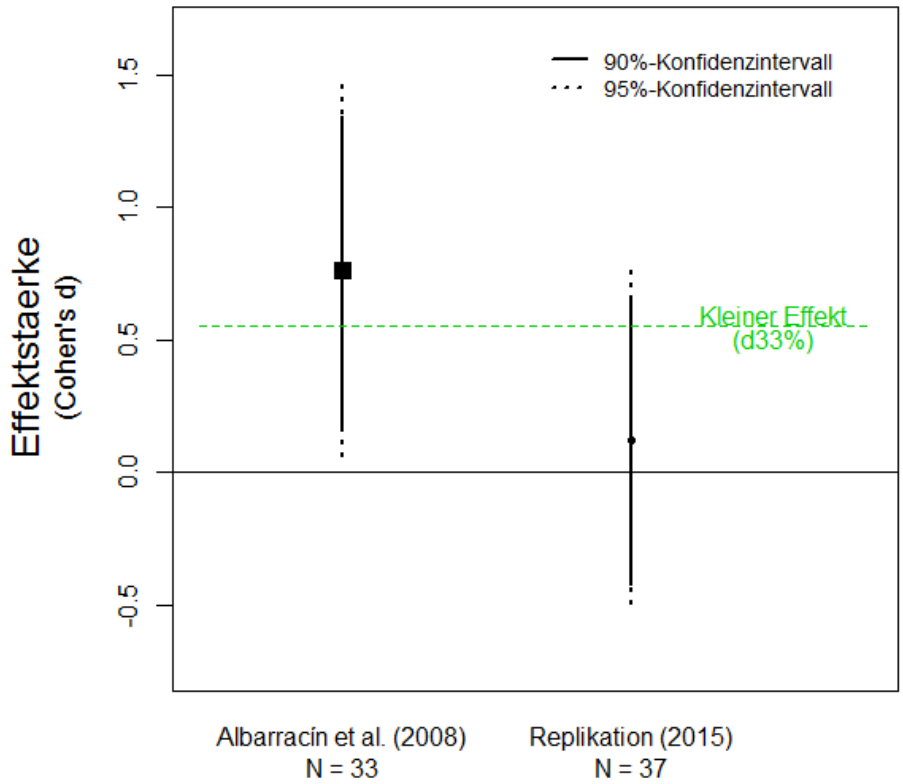

Abbildung 3. Ergebnisse der Hypothese 1 anhand der Stichprobe der Primärstudie von Albarracín et al. (2008) und der Replikationsstudie (2015) in Form der Effektstärken mit den jeweiligen 90%- und 95%-Konfidenzintervallen. Die strichlierte Linie entspricht jener Effektstärke, welcher der Primärstudie 33% Teststärke zusprechen würde ($d_{33\%}$ = 0.549).

Bei einer wahren Effektstärke von $d_{33\%}$ = 0.558 der Hypothese 2 gäbe es eine 59.8%ige Chance, dass diese oder eine noch kleinere Effektstärke anhand der Stichprobe der Replikation (n = 17 und n = 18) entdeckt hätte werden können, p = .598. Auch in diesem Fall wird die Nullhypothese beibehalten.

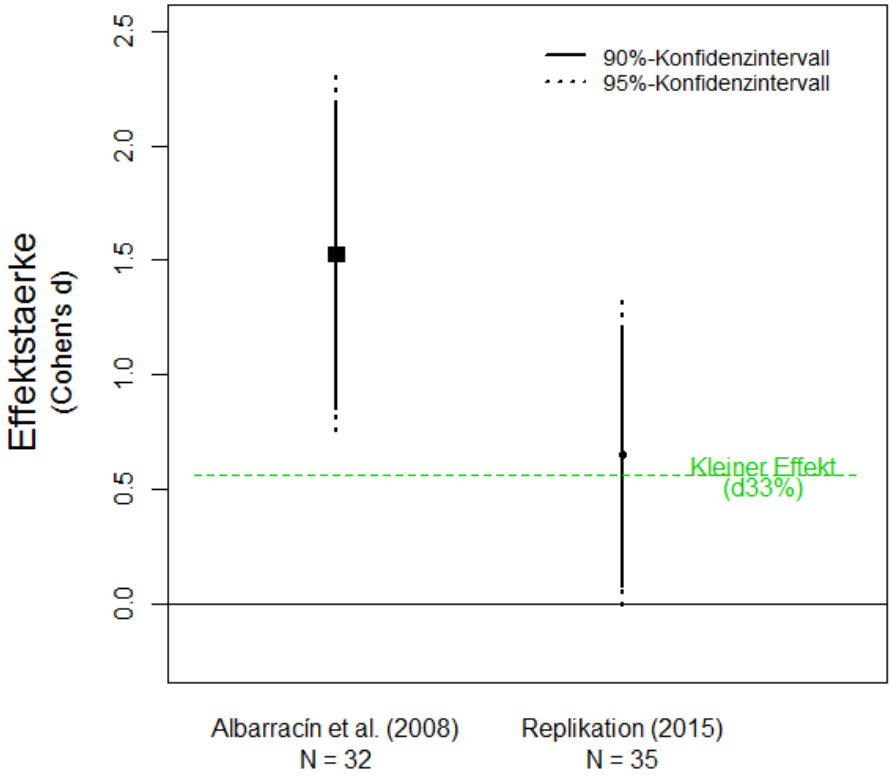

Abbildung 4. Ergebnisse der Hypothese 2 anhand der Stichprobe der Primärstudie von Albarracín et al. (2008) und der Replikationsstudie (2015) in Form der Effektstärken mit den jeweiligen 90%- und 95%-Konfidenzintervallen. Die strichlierte Linie entspricht jener Effektstärke, welcher der Primärstudie 33% Teststärke zusprechen würde ($d_{33\%}$ = 0.558).

5 Diskussion

Sind die Ergebnisse wissenschaftlicher Forschung wiederholbar? Um diese Frage zu beantworten bedarf es der Auseinandersetzung mit grundlegenden Problemen der wissenschaftlichen Erkenntnisgewinnung, allen voran der umstrittenen Nullhypothesen-Signifikanztestung, der fehlerhaften Handhabung von Teststärken, der Anwendung fragwürdiger Forschungspraktiken und der schwerwiegenden Auswirkungen des Publikationsbias. Anhand des bis zum heutigen Tag einzigartigen Reproducibility Project Psychology (Open Science Collaboration, 2015), sowie dessen Demonstration anhand der Replikation von Albarracín et al. (2008), wurde die Problematik nicht reproduzierbarer Forschung verdeutlicht. Darüber hinaus werden Lösungsvorschläge präsentiert, die helfen sollen diesen Problemen zu entgehen oder ihnen zumindest entgegenzuwirken.

5.1 Zusammenfassung

Folgende Probleme wurden im Rahmen dieser Arbeit genauer thematisiert.

5.1.1 Fragwürdige Forschungspraktiken

Die großflächige Verbreitung fragwürdiger Forschungspraktiken (Questionable Research Practices) lässt sich mittlerweile nicht nur vermuten, sondern auch bestätigen. Ob Fortsetzung (57%) oder Beendung (22%) der Datenerhebung zugunsten signifikanter Ergebnisse, opportunistischem Entfernen einzelner Datenpunkte (40%) oder dem vollständigen Weglassen ganzer Variablen in den Berichten (66%) – es gibt wenig, das nicht längst Einzug in den Alltag vieler Psycholog*innen gefunden hat (John et al., 2012). Diese und andere Praktiken lassen falsch-positive wie falsch-negative Studienergebnisse von der statistischen Ausnahme durch Zufall zur unwissenschaftlichen Praxis werden.

5.1.2 Nullhypothesen-Signifikanztesten

Die traditionelle Technik des Nullhypothesen-Signifikanztests (Sterling et al., 1995) hat Forscher*innen an die Goldfrage nach der statistischen Signifikanz gewöhnt – „Ist der *p*-Wert geringer als .05?“ – und die Frage nach der praktischen Signifikanz – „Wie groß ist die Effektstärke?“ – in den Hintergrund rücken lassen (Schimmack, 2012). Der reizvolle Vorteil der ersten Frage liegt auf der Hand: Erfolg oder Scheitern einer Studie lässt sich mit einem knappen *Ja* oder *Nein* anhand der (willkürlich) gezogenen Grenze eines *p*-Werts von meist .05 rasch beantworten (Rosnow & Rosenthal, 1989). Jenseits der fehlerhaften jedoch immer noch existierenden Interpretation des *p*-Werts als Wahrscheinlichkeit für den Wahrheitsgehalt eines Ergebnisses – ein p-Wert von bspw. .02 bedeutet nicht, dass der untersuchte Effekt mit 2-prozentiger Wahrscheinlichkeit existiert (Greenwald et al., 1996;

Cohen, 1994) – oder des Schlusses auf die Nützlichkeit der Alternativhypothese (Gelman, 2015), lassen sich aus diesem Wert alleine keine Schlüsse praktischer Relevanz ziehen. So demonstrieren Rosnow und Rosenthal (1989) sehr anschaulich, wie *p*-Werte ohne die Berücksichtigung von Stichprobengröße und Effektstärke zu einer der Wahrheit entgegengesetzten Interpretation führen können.

5.1.3 Teststärke

Um *p*-Werte akkurat nutzen, d. h. anhand von ihnen tatsächlich die Wahrscheinlichkeit, mit welcher ein Ergebnis bei gültiger Nullhypothese für die Alternativhypothese spricht, ablesen zu können, muss vor allem die Teststärke berücksichtigt werden. Diese beschreibt die Höhe der Wahrscheinlichkeit einen Effekt zu entdecken, wenn dieser vorhanden ist. Sie ergibt sich durch das Zusammenspiel des Signifikanzlevels, der Stichprobengröße sowie der Größe des untersuchten Effekts (Schimmack, 2012). Neben der Wahrscheinlichkeit, dass ein Effekt tatsächlich existiert, haben auch fragwürdige Forschungspraktiken (John et al., 2012), darunter die sog. Freiheitsgrade der Forschenden (z. B. Ende der Datenaufnahme, Anzahl verwendeter Variablen) und diverse Bias (Button et al., 2013) einen Einfluss auf diese Wahrscheinlichkeit. Geringe Teststärken führen selten zur Entdeckung eines Effekts, auch wenn dieser vorhanden ist; stattdessen führen sie zur Überschätzung der Effektgröße, wenn ein Effekt gefunden wird (Button et al., 2013). Es empfiehlt sich, die für einen gesuchten Effekt benötigte Teststärke stets im Voraus einer Studie zu ermitteln, sei es anhand von bereits durchgeführten Studien oder, im Falle einer explorativen Untersuchung, anhand von Pilotstudien (Stanley & Spence, 2014).

5.1.4 Publikationsbias

Die Anwendung fragwürdiger Forschungsmethoden sowie zu kleine Stichproben und dadurch geringe Teststärken führen zu einer Schwemme falsch-positiver Ergebnisse, welche leicht Einzug in die Literatur erhalten. Pashler und Harris (2012) vermuten eine ermittelte Rate von 36% falsch-positiver Ergebnisse in der Forschungsliteratur als optimistische Schätzung (basierend auf einer Teststärke von 80%). Die bevorzugte Publikation signifikanter, interessanter, neuartiger wie kontroverser Ergebnisse (Young et al., 2008) wird anhand des Publikationsbias (publication bias) beschrieben. Neben den hauptsächlich berichteten signifikanten Ergebnissen in der Literatur (Franco et al., 2014; Open Science Collaboration, 2015), muss es jedoch den Gesetzen der Wahrscheinlichkeit folgend auch nicht-signifikante Studienergebnisse geben (Francis, 2012a). Viele dieser nicht-signifikanten Ergebnisse werden Opfer des File Drawer-Effekts (Rosenthal, 1979), demzufolge nicht-signifikante Studien oft in der Schublade (drawer) der Forschenden bleiben, da nur eine geringe Chance auf Publikation besteht. Dieselbe Gefahr laufen nicht-signifikante

Replikationen eines Effekts. Gerade weil es schwieriger ist, replikative Studien zu veröffentlichen (Francis, 2012a), lassen statistisch unwahrscheinlich viele Bestätigungen eines Effekts ebenso auf den Publikationsbias schließen (Makel et al., 2012).

5.1.5 Reproducibility Project Psychology (OSC, 2015)

Im Sinne einer weitläufigen Schätzung des Anteils replizierbarer Effekte in der psychologischen Literatur wurden insgesamt 100 Replikationen von Studien aus drei Zeitschriften sozial- und kognitionspsychologischer Forschung (Journal of Experimental Psychology: Learning, Memory, and Cognition, Journal of Personality and Social Psychology, Psychological Science) aus dem Jahr 2008 durchgeführt. Das Reproducibility Project Psychology (RP:P) wurde über eine Plattform des Center for Open Science koordiniert und beheimatete 270 Forscher*innen aus verschiedenen Institutionen und Ländern der Welt. Um einer möglichst transparenten und standardisierten Vorgehensweise gerecht zu werden, wurden sämtliche Information, Methoden und Daten der einzelnen Studien online und frei zugänglich dokumentiert (siehe https://osf.io/ezcuj/; Aarts et al., 2012). Allen Studien gemeinsam war die Voraussetzung einer Mindestgröße der Teststärke von 80% – die dafür nötige Stichprobengröße wurde im Vorhinein berechnet. Je nach Art der Berechnung der Replizierbarkeit der Ergebnisse ergab sich eine Gesamtanzahl erfolgreicher Replikationen von 36.1% (signifikante Ergebnisse) bis 47.4% (der Primäreffekt befand sich im 95%-Konfidenzintervall der Replikationsstudie). Eine Meta-Analyse konnte bei 75 Studienergebnissen durchgeführt werden; bei 51 von diesen (68%) enthielt das 95%-Konfidenzintervall keinen Nulleffekt. Einer subjektiven Bewertung der Autor*innen der Replikationsstudien zufolge wurden 39 der 100 Ergebnisse erfolgreich repliziert. Mehrheitlich handelte es sich bei den replizierten Effekten um solche aus der kognitionspsychologischen Forschung (50%); Effekte aus dem Bereich der Sozialpsychologie wurden zu 25% repliziert. Einfache Effekte wurden zu 22% und Interaktionseffekte zu 47% erfolgreich repliziert. War der *p*-Wert der Primärstudien unter .02, erreichten diese zu 41% erneut ein signifikantes Ergebnis; Studien mit einem *p*-Wert unter .001 wurden zu 63% erfolgreich repliziert.

Wie sind diese Ergebnisse nun zu interpretieren? Wichtig ist es darauf hinzuweisen, dass diese Ergebnisse nicht notwendigerweise bedeuten, dass kein Effekt existiert. So könnte der Effekt der Primär- wie der Replikationsstudie auf Zufall basieren. Es könnte sich auch um falsch-negative oder falsch-positive Ergebnisse, verursacht durch unzureichende Teststärken aufgrund zu kleiner Stichproben, die Anwendung fragwürdiger Forschungspraktiken, Fehler in der Umsetzung, Einfluss durch unbekannte Moderatoren, oder eine selektive Studienwahl (Publikationsbias) durch Fachzeitschriften, handeln.

Ziel des RP:P war es, eine möglichst umfangreiche Schätzung der Replizierbarkeit zu bewerkstelligen. Infolgedessen wurde aus jeder Studie ein Effekt als zentraler Effekt gewählt – diesen galt es zu replizieren. Studien, welche mit umfangreichen und im Rahmen des Projekts nicht zu bewerkstelligenden Ressourcenaufwand verbunden waren, wurden von der Replikation ausgeschlossen. Daneben sei auch die Möglichkeit einer selektiven Auswahl der Studien durch Autor*innen des Replikationsprojekts genannt, wobei die Freigabe der Studien nach und nach erfolgte, um einen solchen Bias gering zu halten. Wie weit die Ergebnisse des RP:P auf andere Bereiche der psychologischen Forschung übertragbar sind (z. B. Klinische Psychologie) bleibt zudem offen.

5.1.6 Replikation von Albarracín et al. (2008)

Albarracín et al. (2008) erforschten die Auswirkung von Priming auf generelle Aktions- und Inaktionsziele. In einer Reihe von Experimenten wurde die Auswirkung des Primings aktiver, inaktiver und neutraler Wörter auf das von den Versuchspersonen erbrachte Level an Aktivität hinsichtlich der auf das Priming folgenden Aufgaben gemessen. Mit Experiment 7 dieser Studie wurde gezeigt, dass dieser Effekt je nach Art einer zwischengeschalteten Aufgabe (aktiv oder inaktiv) verstärkt oder abgeschwächt werden konnte. Die Interaktion zwischen Priming und Aufgabe, $F(2, 92) = 4.36$ ($p = .02$), sowie der Haupteffekt Aufgabe, $F(1, 92) = 7.57$ ($p = .007$) fielen in der Primärstudie signifikant aus. Weiters konnte anhand einer Kontrastanalyse gezeigt werden, dass aktiv geprimte Personen in Kombination mit einer inaktiven Aufgabe mehr Aktivität (unbefriedigtes Aktionsziel; $M = 6.94$, $SD = 3.46$) im Sinne der Anzahl niedergeschriebener Gedanken zu einem Text zeigten als bei aktivem Priming mit aktiver Aufgabe (befriedigtes Aktionsziel; $M = 4.58$, $SD = 2.76$), $p = .02$. Außerdem wurden weniger Gedanken nach inaktivem Priming in Kombination mit einer aktiven Aufgabe notiert (unbefriedigtes Inaktionsziel, $M = 3.36$, $SD = 0.81$) als bei inaktivem Priming mit inaktiver Aufgabe (befriedigtes Inaktionsziel; $M = 6.50$, $SD = 2.80$), $p = .006$. Diese Ergebnisse wurden dahingehend interpretiert, dass das Erreichen eines zuvor geprimten Ziels den Effekt von Priming abschwächen bzw. diesem entgegenwirken kann.

Die Replikation von Experiment 7 wurde mit einer Stichprobe von 105 Personen durchgeführt. Ein 3 (Priming: aktiv, inaktiv, neutral) x 2 (Aufgabe: aktiv, inaktiv) faktorielles Design wurde ausgeführt. Das Priming der Versuchspersonen erfolgte zu Beginn anhand eines Wortvervollständigungstests, danach mussten sie je nach Art der Aufgabe entweder ein Papierflugzeug basteln bzw. einen Zettel bemalen oder die Augen schließen und rasten. Anschließend wurden sie gebeten einen Text zum Thema Vegetarismus zu lesen und Gedanken zu diesem Text zu notieren. Die Anzahl der notierten Gedanken wurde hierbei als Indikator für Aktivierung verwendet. Zusätzlich wurden die Versuchspersonen am Ende der Studie nach ihrer aktuellen Stimmung und ihrem Ernährungsstil (vegetarisch/vegan) befragt;

hiermit sollte eine eventuell unterschiedliche Antworttendenz berücksichtigt werden, welche durch die aktuelle Stimmung bzw. den Inhalt des Textes verursacht worden sein könnte. Es wurde weder eine Interaktion, $F(1, 105) = 2.890$, $p = .092$, noch der Haupteffekt Aufgabe ($F(1, 105) = 0.518$, $p = .473$) oder Priming ($F(2, 103) = 0.043$, $p = .957$) signifikant. Eine Kontrastanalyse zum Vergleich der einzelnen Gruppen fiel bei Versuchspersonen mit berichteter positiver Stimmung nicht signifikant aus, $F(1, 94) = 0.017$, $p = .897$. Es wurden keine Gruppen unter der Berücksichtigung negativer Stimmung verglichen, da nur 4 Personen eine negative Stimmung berichtet hatten. Eine Meta-Analyse zum Vergleich von unbefriedigten und mit befriedigten Inaktionszielen ergab einen signifikanten Summeneffekt von $d = -1.037$ (SD = 0.424), $p = .015$. Folglich weist der Replikationseffekt in dieselbe Richtung wie jener der Primärstudie. Weiters überschneiden sich die Konfidenzintervalle; sie beinhalten jedoch auch den Nulleffekt, 95%-*KI* [-1.869, -0.205], und sollten aufgrund der vorherigen Ergebnisse mit Vorsicht interpretiert werden.

Die Berechnung der bei einer Teststärke von 33% entdeckbaren Effektstärke $d_{33\%}$ nach Simonsohn (2015) zeigte außerdem, dass die Primärstudie über eine ausreichend große Stichprobe verfügte, um die jeweiligen Effekte zum Vergleich befriedigter und unbefriedigter Aktionsziele bzw. unbefriedigter und befriedigter Inaktionsziele tatsächlich entdeckt haben zu können.

Als Limitierung der Studie sei ein möglicher Unterschied, verursacht durch die Übersetzung der für das Priming verwendeten Wörter, nicht vollständig ausgeschlossen; die Überprüfung des Effekts der Wörter (Priming von Aktivität und Inaktivität) durch Vorstudien war im Rahmen des Projekts nicht vorgesehen.

5.1.7 Lösungsansätze

Es wurde eine Reihe von Methoden vorgeschlagen, welche die Handhabung mit Problemen der Teststärke, fragwürdiger Forschungspraktiken und des Publikationsbias erleichtern sollen. Zur Evaluation von Replikationsergebnissen schlägt Simonsohn (2015) vor, die Nullhypothese in Zukunft zu akzeptieren, wenn eine Replikation mit großer Teststärke einen Effekt nicht findet, dieser aber zuvor anhand einer Studie mit geringer Teststärke entdeckt wurde. Anstelle des Beweisens der Nicht-Existenz eines Effekts soll somit gezeigt werden, dass eine Stichprobe nicht geeignet war, einen Effekt dieser Größe überhaupt erst entdecken zu können.

Eine Möglichkeit, den Publikationsbias oder *p*-hacking sichtbar zu machen, stellt die *p*-curve von Simonsohn et al. (2014a, 2014b, 2015, im Druck) dar. Anhand der Verteilung berechneter und signifikanter *p*-Werte unterschiedlicher Studien zu einem Effekt soll sichtbar gemacht werden, ob selektives Analysieren oder Berichten stattgefunden hat. Je größer ein

tatsächlich vorhandener Effekt ist, desto rechtsschiefer wird die Verteilung unter der Abwesenheit von *p*-hacking ausfallen; beim Vorkommen von *p*-hacking wird die Verteilung bei geringer Effektstärke linksschief sein und erst mit zunehmender Größe rechtsschief werden.

Mit Hilfe des Incredibility Index kann anhand der Teststärke post hoc ermittelt werden, wie viele nicht-signifikante Ergebnisse gemäß dieser Teststärke berichtet hätten werden müssen (Schimmack, 2012). Die Weiterentwicklung dieses Index (Replicability Index) soll angeben, wie groß die Chance ist, eine Studie erfolgreich zu replizieren (Schimmack, 2014).

Die Open Science Collaboration (im Druck) gibt außerdem Anleitung dafür, wie die eigene Forschungsarbeit so gestaltet werden kann, dass diese für potentielle Replikationen optimal genutzt werden kann. Weiters wurde vom Transparency and Openness Promotion (TOP) Committee ein Bündel an Richtlinien erstellt, welches der Transparenz von Studien dienlich sein soll (Alter et al., 2015). Diese Richtlinien können von Fachzeitschriften oder beispielsweise bei der Vergabe von Forschungsgeldern berücksichtigt werden.

5.2 Ausblick

Es bietet sich ein wachsendes Feld von Möglichkeiten und hilfreichen Werkzeugen zur Durchführung, Evaluation und Verbesserung des Bereichs replikativer Forschung, sowohl der psychologischen Wissenschaft als auch anderer Disziplinen. Große Unternehmen wie das Reproducibility Project Psychology machen deutlich, welches Gewicht der Überprüfung von Studienergebnissen beikommt und beikommen sollte. Ein ähnliches Großprojekt wird aktuell im Bereich der Medizin realisiert; das Reproducibility Project Cancer Biology (https://osf.io/e81xl/; Errington et al., 2015) wiederholt Studien zur Krebsforschung aus den Jahren 2010 bis 2012. Im Sinne einer Fortsetzung des Many Labs-Projekt werden im Rahmen von Many Labs 2 weitere 28 Effekte auf ihre Replizierbarkeit über verschiedene Stichproben und Settings hinaus untersucht (https://osf.io/8cd4r/; Klein et al., 2015). Ausgehend vom Center for Open Science wird zudem gerade ein Projekt konkretisiert, welches die Zugänglichkeit von Daten und Materialien von Studien sowie durchgeführte Präregistrierungen anhand von Abzeichen (badges) kenntlich machen soll (https://osf.io/tvyxz/; Blohowiak, et al., 2015).

Einen anderen Ansatz verfolgen Dreber et al. (2015); sie demonstrieren die Vorhersage erfolgreicher Replikationen anhand von Prognosemärkten (prediction markets). 92 Forscher*innen des Reproducibility Project Psychology wurden je 100 US-Dollar zur Verfügung gestellt, um darauf zu wetten, ob sie eine erfolgreiche Replikation von einzelnen Studien (insgesamt 41, mit Ausnahme der eigenen Replikationsstudie) für wahrscheinlich halten. Auf diese Weise wurden die Ergebnisse der Replikationen – 16 von 25 Studien

wurden erfolgreich repliziert – zu 71% korrekt vorhergesagt. Anhand der subjektiven Einschätzung derselben Personen vor der Eröffnung des Marktes waren es nur 58%. Vorhersagemärkte scheinen somit ein geeignetes Mittel zu sein, die Replizierbarkeit von Ergebnissen publizierter Studien einzuschätzen.

In einem eher unkonventionellen Versuch, Forscher*innen zur Präregistrierung ihrer Studien zu bewegen, plant das Center for Open Science die ersten 1,000 Forscher*innen, welche ihre Studien im Vorab registrieren lassen, mit einem Preis von je 1,000 US-Dollar zu belohnen (https://osf.io/x5w7h/; Mellor et al., 2015). Jedoch sollte Präregistrierung nicht als Lösung für verlässliche Ergebnisse betrachtet werden (Nosek & Lakens, 2014). Falsch-positive wie falsch-negative Ergebnisse werden weiterhin existieren – sei es aufgrund von Zufall, unbekannten Moderatoren oder Unterschieden im Setting. Registrierte Berichte sollen außerdem nicht als Abraten von explorativen Analysen verstanden werden, sondern dabei helfen, diese klar von konfirmatorischen Analysen unterscheiden zu können.

Für die Zukunft lässt sich hoffen, dass Replizierbarkeit bereits im Lehrplan auszubildender Institutionen einen würdigen Platz erhält. Bis es soweit ist, kann man sich auch anderweitig mit diesem Thema auseinandersetzen. So wird bspw. auf der Lernplattform Coursera einen Kurs zum Thema Reproducible Research angeboten (https://www.coursera.org/course/repdata). Dieser bietet Teilnehmer*innen das nötige Know-how, um ihre Daten und Programmcodes so zu berichten, dass diese von anderen Forschenden einfach und schnell genutzt werden können, um Studienergebnisse anhand desselben Analysevorgangs zu replizieren.

Literaturverzeichnis

Aarts, A. A., Anderson, C. J., Anderson, J., van Assen, M. A. L. M., Attridge, P. R., Attwood, A. S., … Zuni, K. (2015, October 1). Reproducibility Project: Psychology. Abgerufen von https://osf.io/ezcuj/

Abraham, P. (2000). Dublicate and salami publications. *Journal of Postgraduate Medicine, 46*, 67–69.

Albarracín, D., Handley, I. M., Noguchi, K., McCulloch, K. C., Li, H., Leeper, J., … Hart, W. P. (2008). Increasing and decreasing motor and cognitive output: A model of general action and inaction goals. *Journal of Personality and Social Psychology, 95*, 510–523. doi: 10.1037/a0012833

Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., … Yarkoni, T. (2015). Transparency and openness promotion (TOP) guidelines. Abgerufen von https://osf.io/9f6gx/

Association for Psychological Science (2015). *Ongoing replication projects*. Abgerufen von http://www.psychologicalscience.org/index.php/replication/ongoing-projects

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543–554. doi: 10.1177/1745691612459060

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483,* 531–533. doi: 10.1038/483531a

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407–425. doi: 10.1037/a0021524

Bertamini, M., & Munafò, M. R. (2012). Bite-size science and its undesired side effects. *Perspectives on Psychological Science, 7*, 67–71. doi: 10.1177/1745691611429353

Bettis, R. A. (2012). The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal, 33*, 108–113. doi: 10.1002/smj.975

Blohowiak, B. B., Cohoon, J., de-Wit, L., Eich, E., Farach, F. J., Hasselman, F., … Giner-Sorolla, R. (2015). Badges to acknowledge open practices. Abgerufen von https://osf.io/tvyxz/

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex: Wiley.

Brandt, M. J., Ijzerman, H., Dijksterhuis, A., Farach, F. J., Geller, Giner-Sorolla, R., … van't Veer (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224. doi: 10.1016/j.jesp.2013.10.005

Buck, S. (2015). Solving reproducibility. *Science, 348*: 1403. doi: 10.1126/science.aac8041

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafõ, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376. doi: 10.1038/nrn3475

Callaway, E. (2011). Report finds massive fraud at Dutch universities. *Nature, 479*, 15. doi: 10.1038/479015a

Chambers, C., Banks, G. C., Bishop, D. V. M., Bowman, S. D., Button, K. S., Crockett, M., … Rogelberg, S. (2015). Registered reports. Abgerufen von https://osf.io/8mpji/

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Hillsdale, NJ: Erlbaum.

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

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7–29. doi: 10.1177/0956797613504966

De Winter, J., & Happee, R. (2013). Why selective publication of statistically significant results can be effective. *PLoS ONE, 8*: e66463. doi: 10.1371/journal.pone.0066463

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE, 7*: e29081. doi: 10.1371/journal.pone.0029081

Doyen, S., Klein, O., Simons, D. J., & Cleeremans, A. (2014). On the other side of the mirror: priming in cognitive and social psychology. *Social Cognition, 32*, 12–32. doi: 10.1521/soco.2014.32.supp.12

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksoon, S., Wilson, B., Chen, Y., … Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research, *PNAS*, doi: 10.1073/pnas.1516179112

Errington, T. M., Tan, F. E., Lomax, J., Perfito, N., Iorns, E., Gunn, W., … Williams, S. R. (2015). Reproducibility Project: Cancer Biology. Abgerufen von https://osf.io/e81xl/

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7*, 661–669. doi: 10.1177/1745691612462587

Fraley, R. C., & Vazire, S. (2014). The *N*-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE, 9*: e109019. doi: 10.1371/journal.pone.0109019

Francis, G. (2012A). Publication bias and the failure of replication in experimental psychology. *Psychological Bulletin and Review, 19*, 975–991. doi: 10.3758/s13423-012-0322-y

Francis, G. (2012B). The psychology of replication and replication in psychology. *Perspectives on Psychological Science, 7*, 585–594. doi: 10.1177/1745691612459520

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology, 57*, 153–169. doi: 10.1016/j.jmp.2013.02.003

Franco, A, Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*, 1502–1505. doi: 10.1126/science.1255484

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian Perspective. *Journal of Management, 41*, 632–643.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? Psychophysiology, 33, 175–183.

Hartshorne, J. K. & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience, 6*. doi: 10.3389/fncom.2012.00008

IBM Corp. Herausgegeben 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7*, 645–654. doi: 10.1177/1745691612464056

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–532. doi: 10.1177/0956797611430953

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

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.). *The computer and literary studies* (pp. 153–165). Edinburgh: University Press.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr. Bahnik, S., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A “Many Labs” replication project. *Social Psychology, 45*, 142–152. doi: 10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … Friedman, M. (2015). Many Labs 2: Investigating variation in replicability across sample and setting. Abgerufen von https://osf.io/8cd4r/

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56*, 16–26. doi: 10.1037//0003-066X.56.1.16

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE, 9*: e105825. doi: 10.1371/journal.pone.0105825

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*. doi: 10.3389/fpsyg.2013.00863

Leek, J. T., & Peng, R. D. (2015). *P* values are just the tip of the iceberg. *Nature, 520*. doi: 10.1038/520612a

Lishner, D. A. (2015). A concise set of core recommendations to improve the dependability of psychological research. *Review of General Psychology, 19*, 52–68. doi: 10.1037/gpr0000028

Lucas, R. E. (2013). Improving the replicability and reproducibility of research published in the Journal of Research in Personality. *Journal of Research in Personality, 47*, 453–454. doi:10.1016/j.jrp.2013.05.002

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*, 151–159.

Madden, C. S., Easley, R. W., & Dunn, M. G. (1995). How journal editors view replication research, *Journal of Advertising, 24*, 77–87.

Magnusson, C. (29. September 2015). Understanding statistical power and significance testing: An interactive visualization [Web Log Eintrag]. Abgerufen von http://rpsychologist.com/d3/NHST/

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*, 537–542. doi: 10.1177/1745691612460688

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43*, 304–316. doi: 10.3102/0013189X14545513

Mavridis, D., & Salanti, G. (2014). How to assess publication bias: funnel plot, trim-and-fill method and selection models. *Evidence-Based Mental Health, 17*. doi: 10.1136/eb-2013-101699.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147–163.

McBee, M. T., & Matthews, M. S. (2014). Welcoming quality in non-significance and replication work, but moving beyond the p-value: Announcing new editorial policies for quantitative research in JOAA. *Journal of Advanced Academics, 25*, 73–87. doi: 10.1177/1932202X14532177

McCarthy, R. J. (2014). Close replication attempts of the heat priming-hostile perception effect. *Journal of Experimental Social Psychology, 54*, 165–169. doi: 10.1016/j.jesp.2014.04.014

Mellor, D., Esposito, J., Hardwicke, T. E., Nosek, B. A., Cohoon, J., Soderberg, C. K., & Kidwell, M. (2015). Preregistration challenge: Plan, test, discover. Abgerufen von https://osf.io/x5w7h/

Murray Lindsay, R., & Ehrenberg, A. S. C. (1993). The design of replicated studies. *The American Statistician, 47*, 217–228.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. Social Psychology, 45, 137–141. doi: 10.1027/1864-9335/a000192

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, 1422–1425. doi: 10.1126/science.aab2374

Nuzzo, R. (2014). Scientific method: statistical errors. Nature 506, 150–152, doi: 10.1038/506150a

Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657–660. DOI: 10.1177/1745691612462588

Open Science Collaboration (2014). The Reproducibility Project: A model of large-scale collaboration for empirical research on reproducibility. In V. Stodden, F. Leisch, & R. Peng (Eds.), *Implementing reproducible computational research* (A volume in the R series) (pp. 299–323). New York, NY: Taylor & Francis.

Open Science Collaboration [Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., Babel, M., Bahník, Š., Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., Bosco, F. A., Bowman, S. D., Brandt, M. J., Braswell, E., Brohmer, H., Brown, B. T., Brown, K., Brüning, J., Calhoun-Sauls, A., Callahan, S. P., Chagnon, E., Chandler, J., Chartier, C. R., Cheung, F., Christopherson, C. D., Cillessen, L., Clay, R., Cleary, H., Cloud, M. D., Cohn, M., Cohoon, J., Columbus, S., Cordes, A., Costantini, G., Cramblet Alvarez, L. D., Cremata, E., Crusius, J., DeCoster, J., DeGaetano, M. A., Della Penna, N., den Bezemer, B., Deserno, M. K., Devitt, O., Dewitte, L., Dobolyi, D. G., Dodson, G. T., Donnellan, M., Donohue, R., Dore, R. A., Dorrough, A., Dreber, A., Dugas, M., Dunn, E. W., Easey, K., Eboigbe, S., Eggleston, C., Embley, J., Epskamp, S., Errington, T. M., Estel, V., Farach, F. J., Feather, J., Fedor, A., Fernández-Castilla, B., Fiedler, S., Field, J. G., Fitneva, S. A., Flagan, T., Forest, A. L., Forsell, E., Foster, J. D., Frank, M. C., Frazier, R. S., Fuchs, H., Gable, P., Galak, J., Galliani, E. M., Gampa, A., Garcia, S., Gazarian, D., Gilbert, E., Giner-Sorolla, R., Glöckner, A., Goellner, L., Goh, J. X., Goldberg, R., Goodbourn, P. T., Gordon-McKeon, S., Gorges, B., Gorges, J., Goss, J., Graham, J., Grange, J. A., Gray, J., Hartgerink, C., Hartshorne, J., Hasselman, F., Hayes, T., Heikensten, E., Henninger, F., Hodsoll, J., Holubar, T., Hoogendoorn, G., Humphries, D. J., Hung, C. O., Immelman, N., Irsik, V. C., Jahn, G., Jäkel, F., Jekel, M., Johannesson, M., Johnson, L. G., Johnson, D. J., Johnson, K. M., Johnston, W. J., Jonas, K., Joy-Gaba, J. A., Kappes, H. B., Kelso, K., Kidwell, M. C., Kim, S. K., Kirkhart, M., Kleinberg, B., Knežević, G., Kolorz, F. M., Kossakowski, J. J., Krause, R. W., Krijnen,

J., Kuhlmann, T., Kunkels, Y. K., Kyc, M. M., Lai, C. K., Laique, A., Lakens, D., Lane, K. A., Lassetter, B., Lazarević, L. B., LeBel, E. P., Lee, K. J., Lee, M., Lemm, K., Levitan, C. A., Lewis, M., Lin, L., Lin, S., Lippold, M., Loureiro, D., Luteijn, I., Mackinnon, S., Mainard, H. N., Marigold, D. C., Martin, D. P., Martinez, T., Masicampo, E. J., Matacotta, J., Mathur, M., May, M., Mechin, N., Mehta, P., Meixner, J., Melinger, A., Miller, J. K., Miller, M., Moore, K., Möschl, M., Motyl, M., Müller, S. M., Munafo, M., Neijenhuijs, K. I., Nervi, T., Nicolas, G., Nilsonne, G., Nosek, B. A., Nuijten, M. B., Olsson, C., Osborne, C., Ostkamp, L., Pavel, M., Penton-Voak, I. S., Perna, O., Pernet, C., Perugini, M., Pipitone, R. N., Pitts, M., Plessow, F., Prenoveau, J. M., Rahal, R. M., Ratliff, K. A., Reinhard, D., Renkewitz, F., Ricker, A. A., Rigney, A., Rivers, A. M., Roebke, M., Rutchick, A. M., Ryan, R. S., Sahin, O., Saide, A., Sandstrom, G. M., Santos, D., Saxe, R., Schlegelmilch, R., Schmidt, K., Scholz, S., Seibel, L., Selterman, D. F., Shaki, S., Simpson, W. B., Sinclair, H. C., Skorinko, J. L., Slowik, A., Snyder, J. S., Soderberg, C., Sonnleitner, C., Spencer, N., Spies, J. R., Steegen, S., Stieger, S., Strohminger, N., Sullivan, G. B., Talhelm, T., Tapia, M., te Dorsthorst, A., Thomae, M., Thomas, S. L., Tio, P., Traets, F., Tsang, S., Tuerlinckx, F., Turchan, P., Valášek, M., van 't Veer, A. E., Van Aert, R., van Assen, M., van Bork, R., van de Ven, M., van den Bergh, D., van der Hulst, M., van Dooren, R., van Doorn, J., van Renswoude, D. R., van Rijn, H., Vanpaemel, W., Vásquez Echeverría, A., Vazquez, M., Velez, N., Vermue, M., Verschoor, M., Vianello, M., Voracek, M., Vuu, G., Wagenmakers, E. J., Weerdmeester, J., Welsh, A., Westgate, E. C., Wissink, J., Wood, M., Woods, A., Wright, E., Wu, S., Zeelenberg, M., & Zuni, K.] (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943-952. doi: 10.1126/science.aac4716

Open Science Collaboration (im Druck). Maximizing the reproducibility of your research. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions*. New York, NY: Wiley.

Palus. (26. August 2015). Diederik Stapel ups count to 55 retractions [Web Log Eintrag]. Abgerufen von http://retractionwatch.com/2015/08/26/stapel-is-up-to-55-retractions-another-article-determined-to-be-fraudulent/#more-31371

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–536. doi: 10.1177/1745691612463401

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*, 528–530. doi: 10.1177/1745691612465253

Pashler, H., Spellman, B., Kang, S., & Holcombe, A. (1. Oktober 2015). Abgerufen von http://www.psychfiledrawer.org/top-20/

Psychology Software Tools, Inc. [E-Prime 2.0]. (2012). Abgerufen von http://www.pstnet.com

Rosenbaum, P. R. (2001). Replicating effects and biases. *The American Statistician, 55*, 223–227.

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

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276–1284.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 1-7). Chichester: Wiley.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/

Savalei, V., & Dunn, E. (2015). Is the call to abandon *p*-values the red herring of the replicability crisis? *Frontiers in Psychology, 6*. http://dx.doi.org/10.3389/fpsyg.2015.00245

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551–566. doi: 10.1037/a0029487

Schimmack, U. (2014). Quantifying statistical research integrity: The replicabilty-index. Abgerufen von www.r-index.org/uploads/3/5/6/7/3567479/introduction_to_the_r-index__14-12-01.pdf

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90–100. doi: 10.1037/a0015108

Sherman, R. A., & Wood, D. (2014). Estimating the expected replicability of a pattern of correlations and other measures of association. *Multivariate Behavioral Research, 49*, 17–40. doi: 10.1080/00273171.2013.822785

Simmons, D. J., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. doi: 10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (im Druck). Better *p*-curves. *Journal of Experimental Psychology: General*.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*, 559–569. doi: 10.1177/0956797614567341

Simonsohn, U. (2014). R Code. Abgerufen von https://osf.io/adweh

Simonsohn, U., Nelson, L., & Simmons, J. (2015) *P*-curve [Version 2015-03-02]. Abgerufen am 29. September 2015, von http://p-curve.com/

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*, 534–457. doi: 10.1037/a0033242

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*, 666–681. doi: 10.1177/1745691614553988

Spruyt, A., Hermans, D., Pandelaere, M., De Houwer, J., & Eelen, P. (2004). On the replicability of the affective priming effect in the pronunciation task. *Experimental Psychology, 51*, 109–115. doi: 10.1027/1618-3169.51.2.109

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science, 9*, 305–318. doi: 10.1177/1745691614528518

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*, 108–122. doi: 10.2307/2684823

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*, 59–71. doi: 10.1177/1745691613514450

Tang, J.-L., & Liu, J. L. Y. (2000). Misleading funnel plot for detection of bias in meta-analysis. *Journal of Clinical Epidemiology, 53*, 477–484. doi: 10.1016/S0895-4356(99)00204-8

Tversky, A., & Kahneman, D. (1971). Belief In the law of small numbers. *Psychological Bulletin, 76*, 105–110.

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science, 10*, 390–399. doi: 10.1177/1745691614564879

White, R., and Frank, E. (1994). Health effects and prevalence of vegetarianism. *Western Journal of Medicine, 160*, 465–471.

Vaux, D. L., Fidler, F., & Cumming, G. (2012). Replicates and repeats: What is the difference and is it significant? *EMBO Reports, 13*, 291–296. doi: 10.1038/embor.2012.36

Yong, E. (2012). Bad copy: In the wake of high-profile controversies, psychologists are facing up to problems with replication. *Nature, 485*, 298–300. doi: 10.1038/485298a

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices pay distort science. *PLoS Medicine, 5*, e201. doi: 10.1371/journal.pmed.0050201

Appendix

A. 1 Instruktion der Primärstudie mit Übersetzung ins Deutsche

	Primärstudie	**Replikation**
Instruktion für den Wortverständigungs-test	This first task is a quick measure of verbal ability. You will be shown the first letter or letters of 20 words. After the first letter or letters there will be a blank. Fill in the blank to make a word. For example, if you see: Ap____ You could fill in the blank with "ple" to make "Apple." Type in your responses using your keyboard. Try to make a different word every time.	Die erste Aufgabe besteht aus einer kurzen Messung der sprachlichen Fähigkeiten. Dir werden die ersten Buchstaben von 20 Wörter gezeigt. Nach diesen ersten Buchstaben befindet sich eine Leerstelle. Fülle die Leerstellen aus, um ein Wort zu bilden. Siehst du zum Beispiel: Ap____ dann kannst du die Leerstelle mit „fel" ausfüllen um das Wort „Apfel" zu bilden. Notiere deine Antwort mit Hilfe der Tastatur. Versuche jedes Mal andere Wörter zu bilden.
Priming-Stimuli	Appendix A. 1.	Appendix A. 1.
Instruktion: Inaktive Aufgabe	You are now going to be given a break from the rest of this study's procedure. Take the next two minutes to clear your mind. Please rest, with your eyes closed. You will be notified when two minutes have elapsed.	Du kannst jetzt eine Pause vor dem restlichen Teil der Studie machen. Nutze die nächsten zwei Minuten um deinen Kopf frei zu bekommen. Bitte ruhe dich aus und halte deine Augen dabei geschlossen. Du wirst verständigt, sobald die zwei Minuten vorbei sind.
Instruktion: Aktive Aufgabe	You are now going to be given a break from the rest of this study's procedure. Take the next two minutes to clear your mind.	Du kannst jetzt eine Pause vor dem restlichen Teil der Studie machen. Nutze die nächsten zwei Minuten um deinen Kopf

	Please use the piece of paper provided to draw on or to fold into a paper airplane. You will be notified when two minutes have elapsed.	frei zu bekommen. Bitte verwende das vorbereitete Blatt Papier um darauf zu zeichnen oder ein Papierflugzeug zu basteln. Du wirst verständigt, sobald die zwei Minuten vorbei sind.
Instruktion für Text	Please read the following excerpt from "Health effects and prevalence of vegetarianism" (White and Frank, 1994). After you are finished, you will be asked to list your thoughts on the text.	Bitte lies den folgenden Auszug aus „Gesundheitliche Auswirkungen und Verbreitung von Vegetarismus" („Health effects and prevalence of vegetarianism", White & Frank, 1994). Wenn du fertig bist, wirst du gebeten deine Gedanken zum Text zu notieren.
Text	Appendix A. 3.	Appendix A. 3.
Instruktion nach dem Text (Primärstudie)	Please type any thoughts that you had while reading the previous text into the space below. Please list one thought per line.	Bitte notiere die Gedanken, die du während des Lesens des Textes hattest im unteren Bereich. Bitte schreibe je Zeile nur einen Gedanken.
Instruktion nach dem Text (Replikation)	Please type any thoughts that you had while reading the previous text into the space below. Write three stars (***) at the beginning of each sentence.	Bitte notiere die Gedanken, die du während des Lesens des Textes hattest im unteren Bereich. Schreibe am Anfang eines jeden Gedanken drei Sterne (***).
Stimmung	Right now, do you feel happy? 1 2 3 4 5 6 7 8 9 not at all extremely	Jetzt gerade, fühlst du dich glücklich? 1 2 3 4 5 6 7 8 9 überhaupt nicht extrem
	Right now, do you feel angry? 1 2 3 4 5 6 7 8 9 not at all extremely	Jetzt gerade, fühlst du dich wütend? 1 2 3 4 5 6 7 8 9 überhaupt nicht extrem

Zusätzliche Frage (nicht Teil der Primärstudie)	Are you a vegetarian / vegan? Yes No	Ernährst du dich vegetarisch oder vegan? Ja Nein

A. 2 Priming-Stimuli mit Übersetzung ins Deutsche

(vollständige Wörter mit verwendeten Wortanfängen)

	Primärstudie	**Replikation**
Priming: Aktion	motivation (mo)	Motivation (Mo)
	doing (do)	tuen (tu)
	go (g)	gehen (ge)
	behavior (be)	Verhalten (Ve)
	engage (en)	engagieren (en)
	make (ma)	machen (ma)
	action (ac)	Aktion (Ak)
	active (ac)	aktiv (ak)
Priming: Inaktion	still (st)	noch (no)
	pause (pa)	stoppen (st)
	interrupt (in)	unterbrechen (un)
	calm (ca)	Ruhe (Ru)
	freeze (fr)	frieren (fr)
	unable (un)	unfähig (un)
	stop (st)	stoppen (st)
	paralyze (pa)	lähmen (lä)
Priming: neutral	shoes (sh)	Schuhe (Sc)
	home (ho)	daheim (da)
	lamp (la)	Lampe (La)
	green (gr)	Grün (Gr)
	people (pe)	Menschen (Me)
	mall (ma)	Kaufhaus (Ka)
	food (fo)	Essen (Es)
	time (ti)	Zeit (Ze)
	moon (mo)	Mond (Mo)
	foot (fo)	Fuß (F)
	light (li)	Licht (Li)

	big (b)	groß (gr)
	candles (ca)	Kerzen (Ke)
	spoke (sp)	Speiche (Sp)
	shop (sh)	Geschäft (Ge)
	salty (sa)	salzig (sa)
	watch (wa)	Uhr (Uh)
	time (ti)	Zeit (Ze)
	fictional (fi)	fiktiv (fi)
	book (bo)	Buch (Bu)

A. 3 Text zum Thema Vegetarismus mit Übersetzung ins Deutsche

Text in Originalsprache

"Vegetarians consume less total food energy (calories) than do omnivores because their diets are composed of less total fat and protein (but more total complex carbohydrate). Vegetarians' diets are also comparatively low in cholesterol and saturated fat, have a higher ratio of polyunsaturated to saturated fats and have more fiber.

"The traditional concern about vegetarians' diets has been the adequacy of protein intake, especially of the eight essential amino acids. Data indicate, however, that vegetarians typically receive adequate protein. The American Dietetic Association finds that "vegetarian diets usually meet or exceed requirements for protein" but recommends the consumption of a variety of high-protein plant foods such as grains, nuts, seeds, and legumes each day to obtain enough of each essential amino acid. Eating enough food to meet energy requirements provides adequate protein for pregnant and lactating women as well. In Diet for a Small Planet, Frances Moore Lappé provides an example of the way in which protein requirements can be met with a vegan diet. With a breakfast that included 1 cup of cooked oatmeal (5.4 grams of protein) and 0.5 ounces of sunflower seeds (3.5 grams), a lunch that included 2 tablespoons of peanut butter (7.8 grams) and 2 slices of whole wheat bread (4.8 grams), a dinner that included 1 cup of cooked beans (15.6 grams), 1 cup of cooked brown rice (3.8 grams), and 1.3 cups of broccoli (6.2 grams), with assorted other minor protein food sources (including mushrooms, carrots, apples, raisins, fruit juice, honey, and oil totaling an additional 7.9 protein grams), a total of 57.7 grams of protein and 1,993 calories would be consumed; the recommended dietary allowance for a woman weighing 58 kg (128 lb) consuming a 2,000 calorie diet would be 44 grams of protein. As more calories are consumed, more grams of protein are proportionately consumed. Lacto-ovo vegetarians have additional protein sources: a cup of milk has 9 grams of protein, an ounce of cheese

has 5 to 8 grams, and a medium-sized egg has 6 grams of protein.

"The vitamin and mineral states of vegetarians have been studied by analyzing both food intake and plasma micronutrient levels. Vegetarians have adequate or better intake for most vitamins, including A, C, E, thiamine, and riboflavin. Pyridoxine (vitamin B_6) intake has been found to be low, but one study found this true for omnivores as well, and the ratio of pyridoxine to protein intake—a physiologically appropriate correction, given the function of vitamin B_6 in amino acid metabolism—was found adequate among vegetarians. Tissue folate levels appear normal, and one study found that vegetarians have a higher plasma α-tocopherol (vitamin E)-to-cholesteral ratio than omnivores, which may contribute to their lower risk of atherosclerosis (as discussed later).

"Although vegetarians have adequate intake and levels of many vitamins, they tend to have lower vitamin B_{12} and D intake and plasma levels than do omnivores. This is not surprising, as both these vitamins occur primarily in animal-delivered foods (particularly fortified dairy products in the case of vitamin D). "

White, R., and Frank, E. (1994). Health effects and prevalence of vegetarianism. Western Journal of Medicine, 160:465-471.

Übersetzung ins Deutsche

"Vegetarier konsumieren weniger Gesamtnahrungsenergie (Kalorien) als Omnivore [Anm. = Allesesser], da ihre Nahrung aus weniger Gesamtfett und Eiweiß besteht (aber mit höherem Gesamtanteil an komplexen Kohlenhydraten). Die Nahrung von Vegetariern enthält außerdem vergleichsweise wenig Cholesterin und gesättigte Fette, hat einen höheren Anteil an mehrfach ungesättigten im Vergleich zu gesättigten Fetten und mehr Ballaststoffe.

"Die herkömmliche Sorge hinsichtlich vegetarischer Ernährung ist eine angemessene Eiweißaufnahme, insbesondere jene der acht essentiellen Aminosäuren. Daten deuten jedoch darauf hin, dass Vegetarier normalerweise ausreichend Eiweiß zu sich nehmen. Die *Amerikanische Gesellschaft für Ernährung* findet, dass „vegetarische Ernährung im Allgemeinen den Bedarf an Eiweiß deckt oder überschreitet", empfiehlt jedoch den täglichen und abwechslungsreichen Verzehr von hoch eiweißhaltiger Pflanzennahrung, wie etwa Getreide, Nüssen, Samen und Hülsenfrüchte, um ausreichend essentielle Aminosäuren zu erhalten. Hinreichende Nahrungsaufnahme um Energieanforderungen zu erfüllen, versorgt auch schwangere und stillende Frauen mit ausreichend Eiweiß. In ihrem Buch *Die Öko-Diät* liefert *Frances Moore Lappé* ein Beispiel dazu, wie mit veganer Ernährung der Bedarf an Eiweiß gedeckt werden kann. Mit einem Frühstück bestehend aus einer Tasse gekochter

Haferflocken (5,4 g Eiweiß) und 0,5 Unzen Sonnenblumensamen (3,5 g), einem Mittagessen bestehend aus 2 Esslöffeln Erdnussbutter (7,8 g) und 2 Scheiben Vollkornbrot (4,8 g), einem Abendessen bestehend aus einer Tasse gekochter Bohnen (15,6 g), 1 Tasse gekochten dunklen Reis (3,8 g) und 1,3 Tassen Brokkoli (6,2 g), dazu noch verschiedene andere gering eiweißhaltige Quellen (darunter Pilze, Karotten, Äpfel, Rosinen, Fruchtsaft, Honig und Öl von insgesamt 7,9 g Eiweiß), werden eine Gesamtmenge von 57,7 g Eiweiß und 1.993 Kalorien konsumiert; die empfohlene Nahrungsmenge für eine Frau von 58 kg (128 Pfund) bei einer Aufnahme von 2.000 Kalorien Nahrung beträgt 44 Gramm Eiweiß. Je mehr Kalorien konsumiert werden, desto mehr Gramm Eiweiß werden proportional dazu verbraucht. Ovo-Lakto-Vegetarier verfügen über zusätzliche Eiweißquellen: eine Tasse Milch beinhaltet 9 Gramm Eiweiß, eine Unze Käse 5 bis 8 Gramm und ein mittelgroßes Ei 6 Gramm Eiweiß.

"Der Vitamin- und Mineralienstand von Vegetariern wurde untersucht, indem sowohl Nahrungsaufnahme als auch Plasmaspiegel hinsichtlich der Mikronährstoffe analysiert wurden. Vegetarier zeigen eine ausreichende oder bessere Aufnahme der meisten Vitamine, darunter A, C, E, Thiamin und Riboflavin. Die Pyridoxin(Vitamin B_6)-Aufnahme hat sich als gering erwiesen, jedoch zeigte dies eine Studie auch bei Omnivoren, und der Anteil an Pyridoxin an der Eiweißaufnahme – eine physiologisch angemessene Korrektur, angesichts der Funktion von Vitamin B_6 im Aminosäurenstoffwechsel – war bei Vegetariern ausreichend. Folsäure-Level im Gewebe erscheinen normal und eine Studie fand, dass Vegetarier ein höheres α-Tocopherol(Vitamin E)-zu-Cholesterin Verhältnis im Plasma aufweisen als Omnivore, welches zu deren geringerem Arteriosklerose-Risiko (siehe unten) beitragen könnte.

"Auch wenn Vegetarier über ausreichende Aufnahme und Level von vielen Vitaminen verfügen, tendieren sie zu einer niedrigeren Aufnahme von Vitamin B_{12} und D und niedrigerem Plasmalevel als Omnivore. Dies überrascht nicht, da diese beiden Vitamine primär in tierischen Nahrungsmittel vorhanden sind (insbesondere angereicherten Milchprodukten im Falle von Vitamin D)."

White, R., and Frank, E. (1994). Health effects and prevalence of vegetarianism. Western Journal of Medicine, 160:465-471.

A. 4 Skript Meta-Analyse (R Studio)

```
> ############################## Meta-Analyse ##############################
> ##########################################################################
> ############################## Hypothese 1 ##############################
>
```

```
> ## Ein unbefriedigtes Aktionsziel (Aktion-Priming + inaktive Aufgabe)
> ## führt zu einer höheren Aktivität (Anzahl notierter Gedanken) als
> ## ein befriedigtes Aktionsziel (Aktion-Priming + aktive Aufgabe)
>
> #----------------------------- load data -----------------------------#
> Hypothese1 <- read.csv("C:\\Users\\...\\Hypothese1.csv",
+                        sep = ";")
> detach(Hypothese1)
> attach(Hypothese1)
>
> #---------------------------- load packages -------------------------#
> library("metafor")
> library("car")
>
> #-------------------- calculate SMD & frame outcomes ----------------#
> Hyp1 <- escalc(m1i=m1, m2i=m2, sd1i=sd1, sd2i=sd2, n1i=n1, n2i=n2,
+                measure = "SMD")
>
> metaHyp1 <- data.frame(Hypothese1, Hyp1)
>
> #------------------------- DerSimonian_Laird ----------------------#
> meta_dl <- rma(yi, vi, data = metaHyp1,
+                slab = paste(Studie, sep = ";"),
+                method = "DL")
> meta_dl
>
> ############################ Ende Hypothese 1# ########################
> ######################################################################
> ############################## Hypothese 2 ############################
>
> ## Ein unbefriedigtes Inaktionsziel (Inaktion-Priming + aktive AUfgabe)
> ## führt zu einer niedrigeren Aktivität (Anzahl notierter Gedanken) als
> ## ein befriedigtes Inaktionsziel (Inaktion-Priming + inaktive Aufgabe)
>
> #----------------------------- load data -----------------------------#
> Hypothese2 <- read.csv("C:\\Users\\...\\Hypothese2.csv",
```

```
+                         sep = ";")
> detach(Hypothese2)
> attach(Hypothese2)
>
> #---------------------------- load packages -------------------------#
> library("metafor")
> library("car")
>
> #--------------------- calculate SMD & frame outcomes ---------------#
> Hyp2 <- escalc(m1i=m1, m2i=m2, sd1i=sd1, sd2i=sd2, n1i=n1, n2i=n2,
+                measure = "SMD")
>
> metaHyp2 <- data.frame(Hypothese2, Hyp2)
>
> #--------------------------- DerSimonian_Laird ----------------------#
> meta_dl2 <- rma(yi, vi, data = metaHyp2,
+                slab = paste(Studie, sep = ";"),
+                method = "DL")
> meta_dl2
>
> ############################## Ende Hypothese 2 ########################
> ########################################################################
> ################################### Ende ###############################
```

A. 5 Skript Small Telescope (R Studio),

adaptiert nach Simonsohn (2014, https://osf.io/adweh/)

```
> ##################### satisfied action goals ####################
>
> # Hypothese 1: Ein unbefriedigtes Aktionsziel (Aktion-Priming +
> # inaktive Aufgabe) führt zu einer höheren Aktivität (Anzahl
> # notierter Gedanken) als ein befriedigtes Aktionsziel
> # (Aktion-Priming + aktive Aufgabe)
>
> ##################### Albaracín et al. (2008) ##################
> #---------------------- Berechnung der KI --------------------#
```

```
>
> #Create tnonct function, like SAS
> tnonct = function(delta,pr,x,df) pt(x,df = df, ncp = delta)-pr
>
> #Get the ncp high
> ncp_low90=uniroot(tnonct,c(-10, 10),pr=.95, x=2.173, df=31)$root
> ncp_low95=uniroot(tnonct,c(-10, 10),pr=.975,x=2.173, df=31)$root
>
> #Get the ncp low
> ncp_high90=uniroot(tnonct,c(-10,10),pr=.05,x=2.173,df=31)$root
> ncp_high95=uniroot(tnonct,c(-10,10),pr=.025,x=2.173,df=31)$root
>
> #Go from ncp to d
> dhigh95=ncp_high95/sqrt(16.5/2)
> dhigh90=ncp_high90/sqrt(16.5/2)
> dlow95=ncp_low95/sqrt(16.5/2)
> dlow90=ncp_low90/sqrt(16.5/2)
>
> #print result
> c(dhigh95,dhigh90,dlow90,dlow95)
>
> #----------------------- Berechnung von d33% -------------------#
>
> library(pwr)
> pwr.t.test(n=16.5,power=1/3)$d
>
> ######################## Replikation (2015) ####################
> #----------------------- Berechnung der KI --------------------#
>
> #Create tnonct function, like SAS
> tnonct = function(delta,pr,x,df) pt(x,df = df, ncp = delta)-pr
>
> #Get the ncp high
> ncp_low90=uniroot(tnonct,c(-10, 10),pr=.95, x=0.361, df=35)$root
> ncp_low95=uniroot(tnonct,c(-10, 10),pr=.975,x=0.361, df=35)$root
>
```

```
> #Get the ncp low
> ncp_high90=uniroot(tnonct,c(-10,10),pr=.05,x=0.361,df=35)$root
> ncp_high95=uniroot(tnonct,c(-10,10),pr=.025,x=0.361,df=35)$root
>
> #Go from ncp to d
> dhigh95=ncp_high95/sqrt(18.5/2)
> dhigh90=ncp_high90/sqrt(18.5/2)
> dlow95=ncp_low95/sqrt(18.5/2)
> dlow90=ncp_low90/sqrt(18.5/2)
>
> #print result
> c(dhigh95,dhigh90,dlow90,dlow95)
>
> #-------------------------------------------------------------#
> # evaluate the probability of observing a t-test with an
> # t <=.361 for that ncp
> # (ncp = sqrt(n/2)*d33%  -> ncp = sqrt(18.5/2)*.549)
>
> pt(.361,df=35,ncp=1.67)
>
> ###################################################################
> #R Code behind Figures 1 & 2 in "Small Telescopes"
> #Written by Uri Simonsohn (uws@wharton.upenn.edu)
> #Last update: 2014 12 16
> ###################################################################
>
> Figure 1s.png
>
> library(pwr)     #Library with power functions
> setwd("C:\\Users\\...\\")  #Set directory where
>                                        #figures will be saved
>
> #This function plots 90&95% confidence intervals for 2 studies,
> # original and replication
> repCI=function(d,h90,h95,l90,l95,d33,StudyLabels)
+         ###########################################################
```

```
+ #Syntax:
+ #     d: point estimates for effect size of original and two
+ #        replications
+ #     h90/h95:  high bounds for 90/95% confidence intervals
+ #     l90/l95:  low bounds for 90/95% confidence intervals
+ #     StudyLabels: three string values to put under each
+ #                  confidence interval
+ #   d33: scalar with d33 effect size for sample size of original
+ #        study
+ ######################################################################
+ {
+         x=c(1.5,4.5)
+         #Margins for graph to allow y-label with two lines of text
+         par(mar=c(4.1,6.1,.5,4.1))
+         #Draw the point estimates
+         plot(x,d, xlim=c(0,7),ylim=c(min(l95)-.2,max(h95)+.2),
+              col='black',pch=c(15,16),cex=c(1.5,.75),xlab="",
+              xaxt="n",ylab="")
+         #Label y-axis
+         mtext("Effektstaerke",side=2,line=4.2,cex=1.5)
+         mtext("(Cohen's d)",side=2,line=3,cex=1.1)
+         #Label x-axis
+         axis(side=1,at=x,labels=StudyLabels,line=1,tck=0,lwd=0)
+         #Draw the CI lines
+         for (i in 1:2)
+         {
+                 #90% CI
+                 lines(c(x[i],x[i]),c(h90[i],l90[i]),lwd=2)
+                 #95-90%
+                 lines(c(x[i],x[i]),c(h95[i],l95[i]),lty=3,
+                       col="black",lwd=2)
+         }
+
+         #Add line at 0
+         abline(h=0,col=153)
+         #Add line at d33%
```

```
+         lines(c(0,7),c(d33,d33),col=51,lty=2)
+         #Text near the 33% line
+         text(x=6,y=d33+.05,"Kleiner Effekt",col=51)
+         text(x=6,y=d33-.05,"(d33%)",col=51)
+         #Add CI legend
+         legend(x=3.5,max(h95)+.2,legend=c("90%-Konfidenzintervall"
+                                           ,"95%-Konfidenzintervall"),
+                cex=.85,lty=c(1,3),col=c("black","black"),lwd=2,
+                bty="n")
+ }
>
> ###################################################################
> #Function that computes confidence interval for given t and n
> #First a function that allows finding the corresponding NCP
> pgap = function(ncp_est, t, n, p) pt(t,df=2*n-2,ncp=ncp_est)-p
> #We now apply it for each of the percentiles needed
> ci=function(t,n)
+ {
+         #95% CI
+         h95=(uniroot(pgap, c(-34, 34), p=.025,t = t, n = n)$root)/
+                 sqrt(n/2)
+         l95=(uniroot(pgap, c(-34, 34), p=.975, t = t, n = n)$root)/
+                 sqrt(n/2)
+         #90% CI
+         h90=(uniroot(pgap, c(-34, 34), p=.05, t = t, n = n)$root)/
+                 sqrt(n/2)
+         l90=(uniroot(pgap, c(-34, 34), p=.95, t = t, n = n)$root)/
+                 sqrt(n/2)
+         #estimate
+         d=2*t/sqrt(2*n)
+         #return results
+         return(c(h95,h90,d,l90,l95))
+ }
>
> ###################################################################
>
```

```
> #FIGURE 1 - Enter data from the studies
> #Compute ds and confidence intervals based on reported test
> # statistic
> r1=ci(t=2.173,n=16.5)    #d and confidence intervals for Original
> r2=ci(t=.361, n=18.5)  #d and confidence intervals for Replication
>
> #Convert the vectors of results into the vectors being fed on the
> # plotting function
> h95=c(r1[1],r2[1])
> h90=c(r1[2],r2[2])
> d=c(r1[3],r2[3])
> l90=c(r1[4],r2[4])
> l95=c(r1[5],r2[5])
> #compute d33
> d33=pwr.t.test(n=16.5,power=1/3)$d
>
> #Labels for the studies
> StudyLabels=c("Albarracín et al. (2008)\nN = 33",
+               "Replikation (2015)\nN = 37")
>
> #plot it
> repCI(d=d,h90=h90,h95=h95,l90=l90,l95=l95,d33=d33,
+       StudyLabels=StudyLabels)
> #plot it
>
> png("Figure 1 - Final.png",width=2800,height=1600,res=300)
> repCI(d=d,h90=h90,h95=h95,l90=l90,l95=l95,d33=d33,
+       StudyLabels=StudyLabels)
> dev.off()
>
> ##################################################################
> ############### unsatisfied inaction goals  ####################
>
> # Hypothese 2: Ein unbefriedigtes Inaktionsziel (Inaktion-Priming
> # + aktive Aufgabe) führt zu einer niedrigeren Aktivität (Anzahl
> # notierter Gedanken) als ein befriedigtes Inaktionsziel
```

```
> # (Inaktion-Priming + inaktive Aufgabe).
>
> ##################### Albaracín et al. (2008) ####################
> #---------------------- Berechnung der KI ----------------------#
>
> #Create tnonct function, like SAS
> tnonct = function(delta,pr,x,df) pt(x,df = df, ncp = delta)-pr
>
> #Get the ncp high
> ncp_low90=uniroot(tnonct,c(-10, 10),pr=.95, x=4.309, df=30)$root
> ncp_low95=uniroot(tnonct,c(-10, 10),pr=.975,x=4.309, df=30)$root
>
> #Get the ncp low
> ncp_high90=uniroot(tnonct,c(-10,10),pr=.05,x=4.309,df=30)$root
> ncp_high95=uniroot(tnonct,c(-10,10),pr=.025,x=4.309,df=30)$root
>
> #Go from ncp to d
> dhigh95=ncp_high95/sqrt(16/2)
> dhigh90=ncp_high90/sqrt(16/2)
> dlow95=ncp_low95/sqrt(16/2)
> dlow90=ncp_low90/sqrt(16/2)
>
> #print result
> c(dhigh95,dhigh90,dlow90,dlow95)
>
> #----------------------- Berechnung von d33% -------------------#
>
> library(pwr)
> pwr.t.test(n=16,power=1/3)$d
>
> ######################## Replikation (2015) ####################
> #---------------------- Berechnung der KI ----------------------#
>
> #Create tnonct function, like SAS
> tnonct = function(delta,pr,x,df) pt(x,df = df, ncp = delta)-pr
>
```

```
> #Get the ncp high
> ncp_low90=uniroot(tnonct,c(-10, 10),pr=.95, x=1.921, df=33)$root
> ncp_low95=uniroot(tnonct,c(-10, 10),pr=.975,x=1.921, df=33)$root
>
> #Get the ncp low
> ncp_high90=uniroot(tnonct,c(-10,10),pr=.05,x=1.921,df=33)$root
> ncp_high95=uniroot(tnonct,c(-10,10),pr=.025,x=1.921,df=33)$root
>
> #Go from ncp to d
> dhigh95=ncp_high95/sqrt(17.5/2)
> dhigh90=ncp_high90/sqrt(17.5/2)
> dlow95=ncp_low95/sqrt(17.5/2)
> dlow90=ncp_low90/sqrt(17.5/2)
>
> #print result
> c(dhigh95,dhigh90,dlow90,dlow95)
>
> #-------------------------------------------------------------#
> # evaluate the probability of observing a t-test with an
> # t <=1.921 for that ncp
> # (ncp = sqrt(n/2)*d33%  -> ncp = sqrt(17.5/2)*.558)
>
> pt(1.921,df=33,ncp=1.651)
>
> ######################################################################
> ######################################################################
> #R Code behind Figures 1 & 2 in "Small Telescopes"
> #Written by Uri Simonsohn (uws@wharton.upenn.edu)
> #Last update: 2014 12 16
> ######################################################################
>
> Figure 2s.png
>
> library(pwr)     #Library with power functions
> setwd("C:\\Users\\...\\")  #Set directory where
>                                        #figures will be saved
```

```
>
> #This function plots 90&95% confidence intervals for 2 studies,
> #original and replication
> repCI=function(d,h90,h95,l90,l95,d33,StudyLabels)
+         ############################################################
+ #Syntax:
+ #     d: point estimates for effect size of original and two
+ #        replications
+ #     h90/h95:  high bounds for 90/95% confidence intervals
+ #     l90/l95:  low bounds for 90/95% confidence intervals
+ #     StudyLabels: three string values to put under each
+ #                  confidence interval
+ #   d33: scalar with d33 effect size for sample size of original
+ #        study
+ ####################################################################
+ {
+         x=c(1.5,4.5)
+         #Margins for graph to allow y-label with two lines of text
+         par(mar=c(4.1,6.1,.5,4.1))
+         #Draw the point estimates
+         plot(x,d, xlim=c(0,7),ylim=c(min(l95)-.2,max(h95)+.2),
+              col='black',
+              pch=c(15,16),cex=c(1.5,.75),xlab="",xaxt="n",ylab="")
+         #Label y-axis
+         mtext("Effektstaerke",side=2,line=4.2,cex=1.5)
+         mtext("(Cohen's d)",side=2,line=3,cex=1.1)
+         #Label x-axis
+         axis(side=1,at=x,labels=StudyLabels,line=1,tck=0,lwd=0)
+         #Draw the CI lines
+         for (i in 1:2)
+         {
+                 #90% CI
+                 lines(c(x[i],x[i]),c(h90[i],l90[i]),lwd=2)
+                 #95-90%
+                 lines(c(x[i],x[i]),c(h95[i],l95[i]),lty=3,
+                       col="black",lwd=2)
```

```
+         }
+
+         #Add line at 0
+         abline(h=0,col=153)
+         #Add line at d33%
+         lines(c(0,7),c(d33,d33),col=51,lty=2)
+         #Text near the 33% line
+         text(x=6,y=d33+.05,"Kleiner Effekt",col=51)
+         text(x=6,y=d33-.05,"(d33%)",col=51)
+         #Add CI legend
+         legend(x=3.5,max(h95)+.2,legend=c("90%-Konfidenzintervall",
+                                           "95%-Konfidenzintervall"),
+                cex=.85,lty=c(1,3),col=c("black","black"),lwd=2,
+                bty="n")
+ }
>
> ######################################################################
> #Function that computes confidence interval for given t and n
> #First a function that allows finding the corresponding NCP
> pgap = function(ncp_est, t, n, p) pt(t,df=2*n-2,ncp=ncp_est)-p
> #We now apply it for each fo the percentiles needed
> ci=function(t,n)
+ {
+         #95% CI
+         h95=(uniroot(pgap, c(-34, 34), p=.025,t = t, n = n)$root)/
+                 sqrt(n/2)
+         l95=(uniroot(pgap, c(-34, 34), p=.975, t = t, n = n)$root)/
+                 sqrt(n/2)
+         #90% CI
+         h90=(uniroot(pgap, c(-34, 34), p=.05, t = t, n = n)$root)/
+                 sqrt(n/2)
+         l90=(uniroot(pgap, c(-34, 34), p=.95, t = t, n = n)$root)/
+                 sqrt(n/2)
+         #estimate
+         d=2*t/sqrt(2*n)
+         #return results
```

```
+         return(c(h95,h90,d,l90,l95))
+ }
>
> ######################################################################
>
> #FIGURE 2 - Enter data from the studies
> #Compute ds and confidence intervals based on reported test
> #statistic
> r1=ci(t=4.309,n=16)    #d and confidence intervals for Original
> r2=ci(t=1.921, n=17.5)  #d and confidence intervals for Replication
>
> #Convert the vectors of results into the vectors being fed on the
> #plotting function
> h95=c(r1[1],r2[1])
> h90=c(r1[2],r2[2])
> d=c(r1[3],r2[3])
> l90=c(r1[4],r2[4])
> l95=c(r1[5],r2[5])
> #compute d33
> d33=pwr.t.test(n=16,power=1/3)$d
>
> #Labels for the studies
> StudyLabels=c("Albarracín et al. (2008)\nN = 32",
+               "Replikation (2015)\nN = 35")
>
> #plot it
> repCI(d=d,h90=h90,h95=h95,l90=l90,l95=l95,d33=d33,
+       StudyLabels=StudyLabels)
> #plot it
>
> png("Figure 2 - Final.png",width=2800,height=1600,res=300)
> repCI(d=d,h90=h90,h95=h95,l90=l90,l95=l95,d33=d33,
+       StudyLabels=StudyLabels)
> dev.off()
```

Zusammenfassung

Die Replikationskrise (crisis of confidence) der psychologischen und sozialwissenschaftlichen Forschung hat sich mit Beginn dieses Jahrzehnts einen Weg in die Mitte der wissenschaftlichen Aufmerksamkeit gebahnt und eine zunehmende Verunsicherung hinsichtlich der Standhaftigkeit von Forschungsergebnissen bewirkt. Der Anwendung fragwürdiger Forschungspraktiken ist es zu verdanken, dass die Glaubhaftigkeit vieler als hinreichend nachgewiesen geltender Studienergebnisse neuerdings in Zweifel gezogen wird. Neben diesen Praktiken erschweren zusätzlich die immer häufiger sichtbar gemachten Einschränkungen sowie der fehlerhafte Gebrauch klassischer Methoden der Wissenschaft, wie der Nullhypothesen-Signifikanztestung oder der opportunistische Gebrauch von Teststärken, eine informative Interpretation von Resultaten. Hinzu kommt erschwerend eine der Realität nicht gerecht werdende Abbildung von Studien in wissenschaftlichen Zeitschriften – Erkenntnisse, welche nicht den Schutzmantel der Signifikanz, der Originalität oder des Außergewöhnlichen tragen, fallen vielfach dem Publikationsbias zum Opfer. So trifft vor allem die als wenig erkenntnisreich gebrandmarkten replikativen Studien mehrheitlich das Schicksal, der wissenschaftlichen Öffentlichkeit vorenthalten zu werden und ihr Ende im file drawer der Forschenden zu finden. Um dieser Problematik Rechnung zu tragen, nahm es sich das Reproducibility Project Psychology (Open Science Collaboration, 2015) zur Aufgabe, 100 Replikationen von Studien aus drei prominenten psychologischen Fachzeitschriften zu erstellen und somit eine Schätzung der Rate replizierbarer psychologischer Forschungsergebnisse zu verwirklichen. Der Ablauf dieses Projekts sowie Wege zur Interpretation seiner Ergebnisse werden an der Replikation von Albarracín et al. (2008) demonstriert; die hierbei untersuchten Effekte zum Priming von Aktions- und Inaktionszielen konnten nicht wiederholt werden. Zur weiteren Veranschaulichung dieser Ergebnisse wurden die Berechnung einer Meta-Analyse sowie die Evaluation der Ergebnisse anhand des Small Telescope-Ansatzes nach Simonsohn (2015) vorgenommen. Letztgenannte Methoden sind zwei aus einer Reihe von Hilfsmittel, welche die Replikation von Studienergebnissen in Zukunft erleichtern und die akkurate Einschätzung dieser Ergebnisse fördern sollen.

Abstract

Since the beginning of this decade the focus of attention has been a psychological and social science crisis of confidence, which addresses a growing disbelief in the results of scientific studies. The use of questionable research practices, along with the misuse and misinterpretation of standard methods, led to the questioning discoveries long believed to be true. Furthermore, there is a lack of publication of not significant, extraordinary or innovative findings (publication bias). In order to estimate the rate of reproducible research of psychological science, the reproducibility project psychology (Open Science Collaboration, 2015) conducted a total of 100 replications, addressing studies from three major journals. Process and purpose of the project is demonstrated by the replication of Albarracín et al. (2008), a study concerning the priming of action and inaction goals and whose results could not have been replicated. Moreover a meta-analysis and an evaluation of the replication study, using the small telescope-approach by Simonsohn (2015), were conducted. The aforementioned methods, alongside other useful tools, will hopefully encourage scientists to start valuing replication of scientific discoveries.

Curriculum Vitae

Persönliche Daten	
Geburtsdatum	Juni 1990, Linz an der Donau
Staatsbürgerschaft	Österreichische Staatsbürgerin
Studium	
Seit 2009	**Diplomstudium der Psychologie** – Universität Wien Studienschwerpunkte ▪ Klinische und Gesundheitspsychologie (Kinderpsychologie und -psychiatrie) ▪ Sozialpsychologie (Autorität, Politik und Philosophie) 1. Diplomprüfung abgelegt am 22. Februar 2012
2008 – 2009	**Diplomstudium der Humanmedizin** – Medizinische Universität Innsbruck
Praktika	
08/09 2014	**EXIT-sozial Verein für psychosoziale Dienste, BAGUA – Freizeit, Kommunikation & Beschäftigung**, Linz
2013 – 2014	**ASZ – Das Arbeitsmedizinische und Sicherheitstechnische Zentrum in Linz GmbH**, Linz (6-Wochen-Pflichtpraktikum)

Publikationen & wissenschaftliche Arbeiten

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. Doi: 10.1126/science.aac4716

Kidwell, M., Lazarevic, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., … Nosek, B. A. (28. Oktober 2015). The Effect of Badges on Availability of Data and Materials. Abgerufen von https://osf.io/rfgdw/