universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

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

verfasst von / submitted by

Agnieszka Slowik, BSc

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

Master of Science (MSc)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 840
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Psychologie
Betreut von / Supervisor:	Assoz. Prof. DDDr. Martin Voracek

Danksagung

An erster Stelle möchte ich mich bei meinem Masterarbeitsbetreuer Martin Voracek bedanken. Nicht nur hat er es mir ermöglicht, mich mit der vorliegenden Thematik auseinanderzusetzen, seine gegenüber Studierenden wertschätzende Art, seine fachliche Kompetenz und seine immer zielführende Unterstützung ermöglichen Studierenden in einem Umfeld zu lernen, wie man es sich nur wünschen kann.

Bernhard Hussek möchte ich danken für sein unbedingtes Dasein, aber auch für das Schaffen der Ruhe und des Raums, der für das Verfassen der folgenden Arbeit unabdingbar war.

Carina Sonnleitner möchte ich vor allem für ihre Hilfe bei der Durchführung der Replikationsstudie danken. Durch den gemeinsamen Austausch, insbesondere in der Planungsphase, konnten die ersten Hürden leichter genommen werden.

Meinen Eltern danke ich für jegliche Unterstützung, ohne die die Fertigstellung dieser Arbeit nicht möglich gewesen wäre.

Inhaltsverzeichnis

1. Einführung

Das Vertrauen in die Erkenntnisse experimentalpsychologischer Forschung ist in den letzten Jahren erschüttert worden. Es wird von einer *Vertrauenskrise* bzw. *Replizierbarkeitskrise* der Psychologie gesprochen (Pashler & Wagenmakers, 2012), mittlerweile auch von einer Replizierbarkeitsdebatte[1] (z.B. Earp & Trafimow, 2015).

Bem (2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011) publizierte 2011 im *Journal of Personality and Social Psychology* ein Multi-Study-Paper, das die Existenz von Präkognition zeigen wollte. Wohl kaum jemand glaubt, dass diese Ergebnisse tatsächliche Effekte reflektieren. Die Argumentation der Herausgeber beruhte darauf, dass die Vorgehensweise von Bem den formalen Kriterien des Journals entspreche, und das Paper deshalb nicht abgewiesen werden könne (Judd & Gawronski, 2011). Der Fall hat viel Aufmerksamkeit auf sich gezogen, und verdeutlicht, dass die Art und Weise, wie in der Psychologie aus Daten Schlüsse über Effekte gezogen werden, nicht zu vertrauenswürdigen Ergebnissen führt.

In diesem Kontext entstanden verschiedene Bestrebungen vermehrt Replikationen psychologischer Forschung durchzuführen. Das *Reproduciblity Project: Psychology* (R:PP, Open Science Collaboration, 2015) repliziert eine Stichprobe von 100 psychologischen Studien, und zielt so vor allem darauf auf ab, die Replizierbarkeit psychologischer Forschung abzuschätzen. *Many Labs* (z.B. Klein et al., 2014) und *Registered Replication Reports* (Simmons & Holocombe, 2014) wiederholen hingegen ausgewählte Effekte an jeweils mehreren Stichproben, um zu untersuchen, ob die behaupteten Effekte der Prüfung standhalten. Sämtlichen Replikationsinitiativen ist gemeinsam, dass die Replikationen präregistriert werden, und unabhängig von ihrem Ergebnis publiziert werden.

Im Rahmen der vorliegenden Masterarbeit wurde eine Replikation für das Reproducibility Project: Psychology durchgeführt. Bevor das Projekt näher erläutert wird bzw. auch die Ergebnisse der Replikation präsentiert werden, wird zum einen auf Replikationsstudien eingegangen, zum anderen werden Faktoren angesprochen, die das Vertrauen in die Resultate der experimentellen Psychologie schmälern: die Anreizstrukturen

[1] Die vorliegende Arbeit ist als Abschlussarbeit eines Psychologiestudiums entstanden, und konzentriert sich somit auf die Diskussion innerhalb dieses Faches. Es soll aber nicht der Eindruck entstehen, die Psychologie sei als einzige Disziplin von derartigen Überlegungen betroffen. Tatsächlich wird die Replizierbarkeitsdebatte breiter geführt (z.B. Atmanspacher & Maasen, 2016).

in der psychologischen Forschung fördern neue Ergebnisse, aber nicht deren Absicherung; es werden vor allem statistisch signifikante Ergebnisse publiziert; dies führt zu einer verzerrten Literatur; gleichzeitig ist die Aussagekraft eines (signifikanten) *p*-Werts gering, wird aber von den meisten Psychologen überbewertet, und das Konzept der statistischen Teststärke findet weiterhin wenig Beachtung. Weiters werden Möglichkeiten zur Bestimmung der Stichprobengröße von Replikationen angesprochen. Replikationen stellen aber nicht die einzige Möglichkeit dar, Studien auf ihre *Evidentialität* bzw. Vertrauenswürdigkeit hin zu bewerten. Alternativ werden statistische Kennwerte entwickelt, die der Bewertung von Studien dienen. Hier wird nur exemplarisch auf den prominentesten Vertreter – die *p-Curves* – eingegangen (Simonsohn, Nelson, & Simmons, 2014). Abschließend sollen mögliche Lösungsansätze skizziert werden.

2. Replikationsstudien in der psychologischen Forschung

Replikationsstudien („*replication studies*“), sind empirische Studien, deren Design und Methoden nicht neu sind, sondern sich an bereits durchgeführten Studien orientiert. Es handelt sich um Wiederholungsstudien, die es erlauben, die Ergebnisse von Originalstudien zu überprüfen und abzusichern, und verzerrte Befunde zu identifizieren (Döring & Bortz, 2015).

Replikationsstudien können sich, in dem Ausmaß in dem sie der Originalstudie gleichen, unterscheiden. *Direkte Replikationen* sind möglichst exakte Wiederholungen von experimentellen Versuchsanordnungen an einer neuen Stichprobe, und erlauben es falsch positive Ergebnisse aufzudecken (Döring & Bortz, 2015).

Konzeptuelle Replikationen untersuchen eine Hypothese oder ein Ergebnis vorangegangener Forschungsarbeiten mit abweichenden experimentellen Methoden, wie z.B. Design, Operationalisierung und/oder Datenanalyse (Schmidt, 2009). Es wird also nicht die Studie selbst, sondern die der Studie zugrundeliegende Idee wiederholt. Auf den ersten Blick wirken konzeptuelle Replikationen vielversprechend. Lässt sich ein theoretischer Effekt auf unterschiedliche Arten operationalisieren und somit generalisieren, so stärkt das zunächst das Vertrauen in dessen Existenz. Konzeptuelle Replikationen unterliegen allerdings im Prinzip den gleichen systematischen Problemen, denen auch experimentelle Studien ausgeliefert sind, und sind nicht im Stande die gleiche Funktion zu erfüllen, wie

direkte Replikationen (Pashler & Harris, 2012).

Die vorliegende Arbeit beschäftigt sich vornehmlich mit direkten Replikationen bzw. den Hintergründen ihrer Dringlichkeit. Die Problematik von konzeptuellen Replikationen und Originalstudien wird im erst Verlauf des Textes deutlicher. Ist in Folge von Replikationen zu lesen, sind direkte Replikationen gemeint.

2.1. Replikationsstudien werden selten publiziert

Replikationen mit dem Zweck vorangegangene Arbeiten zu bestätigen oder zu widerlegen, werden in der Psychologie selten publiziert. Makel, Plucker und Hegarty (2012) haben die Top 100 psychologischen Fachjournals (gemessen Impact Factor) auf die Anzahl der darin publizierten Replikationen hin durchsucht. In einer Zufallsstichprobe von 500 Publikationen waren nur etwa 1% der Studien direkte oder konzeptuelle Replikationen und wurden auch explizit als Replikation bezeichnet. Davon waren aber 80% konzeptuelle Replikationsstudien, d.h. die Anzahl direkter Replikationen macht gerade einmal 0,2% der psychologischen Publikationen aus.

Es stellt sich die Frage, warum so wenig Replikationen in der Literatur vorhanden sind, obwohl Replikationen für die Überprüfung bzw. Bestätigung vorhandener Ergebnisse zentral erscheinen. Die Anreize eine Replikation durchzuführen sind in der Psychologie jedenfalls sehr gering.

2.2. Die Anreizstrukturen der psychologischen Forschung fördern neue Ergebnisse

Um als Wissenschaftler oder Wissenschaftlerin erfolgreich zu sein, muss man publizieren. Aber nicht jede wissenschaftliche Tätigkeit, die in einem Manuskript mündet, wird auch publiziert – viele der eingereichten Manuskripte werden nicht angenommen. Rotton, Levitt und Foos (1993) beispielsweise berichten Ablehnungsquoten von in etwa 80% für die Journale der American Psychological Association. Forschungsarbeiten, die bestimmte Kriterien erfüllen, haben jedenfalls bessere Karten publiziert zu werden (Nosek, Spies, & Motyl, 2012).

Replikationen scheinen diese Kriterien nicht zu erfüllen. Bereits im APA-Manual wird als einer der möglichen Gründe für die Ablehnung eines Manuskripts angeführt: „... it

is judged as making a limited novel contribution to the field.“ (American Psychological Association, 2010, S. 227). Und auch Herausgeber fördern Replikationen nicht. Neuliep und Crandall (1991) führten eine Umfrage unter 79 Herausgebern und Herausgeberinnen sozialwissenschaftlicher Fachjournale durch, und fanden starke Vorurteile gegenüber der Publikation direkter Replikationen. Während 72% der befragten Herausgeber der Meinung waren, eine Studie, die einen neuen Effekt demonstriere sei wichtiger, als die Replikation eines bereits gezeigten Effekts, gaben nur 6% an, eine Replikation für bedeutsamer zu halten. Replikationen seien langweilig, uninteressant und leisteten keinen neuen Beitrag zum Verständnis von Phänomenen, war der Tenor unter den befragten Herausgebern und Herausgeberinnen.

Neue, noch nicht dagewesene Ergebnisse sind also jene die geschätzt werden. Innovation wird als Basis wissenschaftlichen Fortschritts angesehen. Replikationen sind aber nicht innovativ, sie sind die Wiederholung von etwas bereits Dagewesenem. Das schmälert ihren Wert aber keinesfalls. Die Publikation eines einzigen signifikanten Effekts ist nämlich nicht damit gleichzusetzen, dass dieser auch tatsächlich existiert (vgl. Nosek et. al, 2012).Viel mehr ist die Exaktheit publizierter Ergebnisse, die auf Nullhypothesentests beruhen, kritisch zu sehen. Ein einziges signifikantes Ergebnis ist kein ausreichender Beweis für einen Sachverhalt. Schon Ronald Fisher schrieb:

> In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result. (Fisher, 1971, S. 14)

Fisher betont hier die Bedeutung wiederholter Experimente im Rahmen der Signifikanzprüfung. Das Vertrauen in einen Effekt kann nur zustande kommen, wenn dieser immer wieder gezeigt werden kann. Replizierbarkeit ist also eine (zumindest theoretische) Voraussetzung um Effekte zu demonstrieren.

3. Publikationsbias oder die Ansammlung statistisch signifikanter Effekte in der wissenschaftlichen Literatur

In der empirischen Psychologie hat sich ein signifikantes Ergebnis faktisch als Kriterium für eine Publikation etabliert (Nosek et al., 2012). Ein Ergebnis wird dann als signifikant gewertet, wenn der aus den Daten errechnete *p*-Wert unter dem vorher festgelegten kritischen Signifikanzniveau Alpha liegt. Alpha wird in den Sozialwissenschaften üblicherweise mit $\alpha = .05$ festgelegt, und bezeichnet die Wahrscheinlichkeit einen Fehler 1. Art zu begehen, d.h. die Nullhypothese fälschlicherweise abzulehnen. Der *p*-Wert wiederum gibt die Wahrscheinlichkeit an, die vorliegenden oder extremere Daten zu finden, und zwar unter der Annahme, dass die Nullhypothese stimmt bzw. kein Effekt vorhanden ist (z.B. Gigerenzer, 2004, Kline, 2013).

Psychologie und Psychiatrie sind nach Fanelli (2010) mit 92% positiven Ergebnissen jene Disziplinen mit dem höchsten Anteil an signifikanten Ergebnissen in der wissenschaftlichen Literatur. Dieser Umstand ist nicht neu (Sterling, 1959); der Anteil positiver Ergebnisse an der wissenschaftlichen Literatur stieg über die Jahrzehnte sogar an, und zwar besonders stark in den Sozialwissenschaften (Fanelli, 2012).

Selbst wenn man davon ausgeht, dass nur wahre Hypothesen getestet werden, reicht die in der Psychologie übliche Teststärke (vgl. Kap. 6) nicht aus, um einen derart hohen Prozentsatz statistisch signifikanter Ergebnisse zu erhalten, d.h. die publizierte psychologische Fachliteratur ist nicht vollständig. Forschungsergebnisse, deren Ergebnisse nicht signifikant waren, fehlen in der publizierten Literatur. Es wird auch vom *Publikationsbias* gesprochen (Dickersin & Min, 1993).

Auch wenn die Teststärke in der Psychologie höher wäre, müsste man davon ausgehen, dass einige der geprüften Zusammenhänge nicht existieren. Diese sind für den Wissensbestand eines Faches aber ebenso wertvoll, und sollten Bestandteil der Literatur sein. Es gibt keinen Grund anzunehmen, dass Schätzungen um Null herum weniger Wert hätten, als Schätzungen, die statistisch signifikant von Null abweichen (Greenwald, 1975). Eine von Greenwald durchgeführte Umfrage lässt darauf schließen, dass sich Autoren bei vorliegenden Nullergebnissen eigener Studien nur selten um eine Publikation bemühen. Nullergebnisse tragen aber ebenso zum Wissensbestand einer Disziplin bei, ihr Fehlen stellt einen nicht zu unterschätzenden Verlust dar.

Rosenthal (1979) prägte in diesem Zusammenhang auch den Begriff des *File-Drawer-Problems*. Im ungünstigsten Fall würden ausschließlich nicht existente Zusammenhänge geprüft werden. Dann bestände die Fachliteratur nur aus jenen 5% Studien mit falsch positiven Ergebnissen, während die restlichen 95% der Studien den „file drawer" bildeten.

Der tatsächliche Anteil wahrer Nulleffekte (bzw. der Anteil an Effekten, die so nah an Null sind, dass sie keine praktische Relevanz besitzen) an allen getesteten Hypothesen, ist aber unbekannt. Ebenso ist der Anteil falsch positiver Ergebnisse an der publizierten Literatur unbekannt. Allein die Möglichkeit falsch positiver Ergebnisse müsste ausreichend Motivation für die Überprüfung publizierter Studien bzw. die Durchführung von Replikationen bieten.

4. Das (Miss-)Verständnis von *p*-Werten

Wenn man bedenkt, welchen zentralen Stellenwert der *p*-Wert in der Psychologie einnimmt, ist es überraschend, wie viele Missverständnisse es bezüglich seiner Aussagekraft gibt (Oakes, 1986; Haller & Krauss, 2002; Kline, 2013; Gigerenzer, 2004, Carver, 1978). Die Diskussion um die (wahrgenommene) Bedeutung von *p*-Werten (Nuzzo, 2014) veranlasste die American Statistical Association dieses Jahr zu einem Statement (Wasserstein & Lazar, 2016), in dem in sechs Punkten geklärt wird, was ein *p*-Wert (nicht) ausdrücken kann.

Kline (2013) sieht das verzerrte Verständnis von *p*-Werten sogar als eine der Ursachen der geringen Wertschätzung von Replikationen in der psychologischen Forschung. Psychologen und Psychologinnen sind sich der statistischen Ergebnisse womöglich zu sicher, und halten deshalb Replikationen für nicht notwendig.

Haller und Krauss (2002) setzten einen Fragebogen, den zuvor schon Oakes (1986) verwendet hatte, an sechs deutschen, universitären Psychologieinstituten ein. Sie legten diesen nicht nur Psychologiestudierenden vor, sondern auch akademischen Psychologen und Psychologinnen und jenen, die Statistik in der Psychologie unterrichteten (darunter auch studentische Tutoren und Tutorinnen). Zunächst wurde das Ergebnis einer einfachen Studie berichtet und im Zuge dessen auch ein *p*-Wert ($p = .01$) angegeben. Es folgten 6 Aussagen zum *p*-Wert, die entweder mit richtig oder falsch zu bewerten waren.

Die Aussagen „You have absolutely disproved the null hypothesis.“ bzw. „You have absolutely proved your experimental hypothesis.“ fanden, relativ gesehen, die geringste Zustimmung (Haller & Krauss 2002, S. 5). Etwa 10-30% der Befragten beurteilten diese als „richtig“. Signifikanztests können aber nur Wahrscheinlichkeitsaussagen machen und sind deshalb nicht dazu geeignet, Hypothesen zu widerlegen oder zu beweisen.

Etwa 20-30% der Befragten stimmten folgender Aussage zu: „You have found the probability of the null hypothesis being true“ (Haller & Krauss, 2002, S. 5). Es handelt sich hierbei um den Irrglauben, der *p*-Wert gäbe bei gegebenen Daten die Wahrscheinlichkeit an, dass die Nullhypothese wahr wäre (*inverse probability fallacy*). Hier wird $p(\text{Daten}|H_0)$, also die Wahrscheinlichkeit die vorgefundenen (oder extremere) Daten zu erhalten, wenn die Nullhypothese wahr ist, verwechselt mit $p(H_0|\text{Daten})$, also mit der Wahrscheinlichkeit, dass die Nullhypothese stimmt, wenn die gefundenen Daten vorliegen (Kline, 2013). Der *p*-Wert gibt aber keine Informationen über die Wahrscheinlichkeit der Nullhypothese, sondern nur über die Wahrscheinlichkeit der gefundenen Daten (unter Annahme der Nullhypothese).

Der Aussage „You can deduce the probability of the experimental hypothesis being true.“ stimmten in etwa 30% der wissenschaftlichen Psychologen und Psychologinnen, und etwa 60% der Psychologiestudierenden zu (Haller & Krauss, 2002, S. 5). Damit verfielen sie dem Missverständnis 1-*p* gäbe bei vorliegenden Daten die Wahrscheinlichkeit an, dass die Alternativhypothese wahr wäre (*valid research hypothesis fallacy,* Carver, 1978), d.h. sie verwechselten 1 - $p(\text{Daten}|H_0)$ mit $p(H_1|\text{Daten})$.

Eigentlich hängen diese zwei von Haller und Krauss (2002) untersuchten, falschen Aussagen bezüglich der Wahrscheinlichkeit von H_0 bzw. von H_1 zusammen, da $p(H_1) = 1-p(H_0)$. D.h. wenn eine der Aussagen stimmen sollte, müsste auch die andere Aussage stimmen. Die Antworten der befragten Personen waren aber nicht konsistent. Deutlich mehr Personen glaubten mit Hilfe des *p*-Wertes könne man eine Aussage über die Wahrscheinlichkeit der Alternativhypothese machen.

Etwa zwei Drittel der befragten Personen stimmten folgender Aussage zu: „You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision“ (Haller & Krauss, 2002, S. 5). Es handelt sich hier um die falsche Auffassung, dass wenn aufgrund eines $p < .05$ bei gegebenen Signifikanzniveau $\alpha = .05$ die Nullhypothese verworfen wird, die Wahrscheinlichkeit einen Fehler (1. Art) begangen zu haben kleiner als 5% ist (*local type I error fallacy*). Wird die Nullhypothese verworfen,

entspricht die Wahrscheinlichkeit eine falsche Entscheidung zu treffen, der Wahrscheinlichkeit, dass die Nullhypothese stimmt; denn nur wenn die Nullhypothese stimmt, ist die Entscheidung sie zu verwerfen falsch. Es wird also die Wahrscheinlichkeit die Nullhypothese zu verwerfen, wenn sie wahr ist, also: $\alpha = p(H_0 \text{ verwerfen} | H_0 \text{ wahr})$ verwechselt mit der Wahrscheinlichkeit, dass die Nullhypothese stimmt, wenn sie verworfen wurde (Kline, 2013): $p(H_0 \text{ wahr} | H_0 \text{ verwerfen})$.

Ein häufiger Fehler betrifft auch den Irrtum, dass 1-*p* die Wahrscheinlichkeit angäbe, einen statistisch signifikanten Effekt zu erhalten, sollte die Studie repliziert werden (*replicability fallacy*). Beinahe die Hälfte der befragten Personen gab an die folgende Aussage wäre richtig: „You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions“ (Haller & Krauss, 2002, S. 5).

100% der Psychologiestudierenden, 90% der forschenden Psychologen und 80% der Statistik unterrichtenden Psychologen machten zumindest einen Fehler, indem sie mindestens einer dieser sechs Aussagen zustimmten. Im Durchschnitt machten die Studierenden 2.5 Fehler, die Wissenschaftler 2.0, und die Instruktoren 1.9 Fehler. Wenn sogar, diejenigen, die Statistik unterrichten das Konzept des Nullhypothesentestens nicht verstehen, ist es nicht verwunderlich, dass Missverständnisse so weit verbreitet sind.

Alle sechs dieser Aussagen lassen den *p*-Wert informativer wirken als er ist. Man kann also auch annehmen, dass Psychologen zu viel Vertrauen in die Ergebnisse psychologischer Studien haben, und möglicherweise Replikationen deswegen nicht für notwendig halten (Kline, 2013).

Im Kontext von Replikationen ist der Replizierbarkeitsirrtum (replicability fallacy) besonders interessant: Gäbe 1-*p* tatsächlich die Wahrscheinlichkeit an, dass eine Replikation eines statistisch signifikanten Effekts wiederum zu einem statistisch signifikanten Ergebnis führe, wäre die Notwendigkeit von Replikationen eine weniger dringliche. Vorhersageintervalle von *p*-Werten (*prediction intervals*) verdeutlichen aber, wie weit diese Annahme von der Realität entfernt ist. Ein solches Vorhersageintervall gibt mit einer bestimmten Wahrscheinlichkeit an, innerhalb welcher Grenzen sich der *p*-Wert einer Replikation bewegen wird (Cumming, 2008). Angenommen der *p*-Wert der Originalstudie liegt bei $p = .05$, dann umfasst das einseitige 80% Vorhersageintervall, welches sich bis zum 80. Perzentil der entsprechenden *p*-Wert-Verteilung erstreckt, *p*-Werte bis zu $p = .22$; Das

bedeutet, dass wenn eine Originalstudie einen *p*-Wert von $p = .05$ hat, liegt die Wahrscheinlichkeit, dass eine Replikation einen *p*-Wert kleiner als $p = .22$ ergibt bei 80%[2]. Ein zweiseitiges *p*-Intervall welches sich vom 10. Bis zum 90. Perzentil der entsprechenden Verteilung erstreckt, umfasst *p*-Werte zwischen $p = .00008$, und $p = .44$.

Cumming (2013) spricht in diesem Zusammenhang auch von einem Dance of the p-values. Zufallseffekte, die eine Abweichung vom wahren Wert bedingen, wirken sich stärker auf Veränderungen des p-Werts aus, als auf Veränderungen der Konfidenzintervalle bzw. Effektstärken; p-Werte sind also allein aufgrund des Standardfehlers und der sich daraus ergebenden Variation von einer Stichprobe (bzw. Replikation) zur nächsten, kein zuverlässiger Kennwert zur Bewertung von Ergebnissen.

Konfidenzintervalle sind hinsichtlich des Ergebnisses von Replikationen aussagekräftiger: Die Wahrscheinlichkeit, dass der Mittelwert einer Replikation vom 95% Konfidenzintervall des ursprünglichen Effekts umfasst wird, beträgt 83%3 (Cumming, Williams, & Fidler, 2004). Diese Wahrscheinlichkeit, auch „capture percentage" genannt, erfasst also jenem Anteil an Mittelwerten wiederholter Replikationen, die auf lange Sicht gesehen, innerhalb der Grenzen des ursprünglichen Konfidenzintervalls ausfallen. Doch auch die Variabilität von Konfidenzintervallen wird von Psychologen unterschätzt.

Eine mögliche Ursache warum zu viel Vertrauen in die Aussagekraft von statistischen Ergebnissen gelegt wird, sehen Tversky und Kahneman (1971) in einem zu starken Glauben an die Aussagekraft „kleiner" Zahlen (bzw. Stichproben). Sie sprechen von einem „*belief in the law of small numbers*" (Tversky & Kahneman, 1971, S. 105) in Anspielung an das *Gesetz der großen Zahlen*. Während sehr große Zufallsstichproben eine gute Repräsentation der Populationsparameter bieten, können kleine Stichproben sehr stark von diesen abweichen.

5. P-Hacking, Questionable Research Practices oder warum Nullhypothesentests scheinbar häufiger signifikante Ergebnisse liefern können, als sie sollten

Es gibt Praktiken – oft als *p-Hacking* oder als *Questionable Research Practices* (QRPs, „fragwürdige Forschungspraktiken") bezeichnet, die im Rahmen einer

[2] Allerdings nur, wenn die Variation zwischen den Replikationen nur auf dem Standardfehler beruht.
[3] Auch hier: nur, wenn Messfehler allein aufgrund des Standardfehlers entstanden sind.

wissenschaftlichen Untersuchung angewandt, die Wahrscheinlichkeit eines (falsch) positiven Ergebnisses erhöhen können, und damit auch die Chancen auf eine Publikation steigern (Nosek et al., 2012). Es handelt sich um Vorgehensweisen, die zwar manchen Wissenschaftlern gerechtfertigt erscheinen können, und in der Praxis gebräuchlich sind, aber das Potential besitzen die Anzahl falsch positiver Ergebnisse zu steigern. Die Zuverlässigkeit so gewonnener Ergebnisse erscheint jedenfalls fraglich.

Simmons, Nelson und Simonsohn (2011) sprechen in diesem Zusammenhang von Freiheitsgraden, die Wissenschaftlern und Wissenschaftlerinnen zur Verfügung stehen (*researcher degrees of freedom*), und meinen damit die Flexibilität, die bei der Sammlung und Analyse von Daten in der psychologischen Forschung gegeben ist. Im Zuge der Durchführung einer Studie sind zahlreiche Entscheidungen zu treffen. Es gibt beispielsweise unzählige Möglichkeiten einen vorhandenen Datensatz zu analysieren. Studien werden nicht im Voraus bis ins kleinste Detail geplant, sondern viele Optionen werden offengelassen. Um zu illustrieren, wie einfach es ist, signifikante Ergebnisse zu erhalten, und wie einflussreich das Nutzen bestimmter Analysestrategien ist, haben Simmons et al. (2011) zwei Studien durchgeführt, und mit Hilfe der von ihnen kritisierten Praktiken ausgewertet. In der einen Studie konnten sie zeigen, dass sich das subjektive Alter der Teilnehmer und Teilnehmerinnen erhöhen ließ, indem man ihnen ein Kinderlied vorgespielt hat. In der anderen Studie wurden in einer konzeptuellen Replikation des zuvor demonstrierten Effekts die Teilnehmer und Teilnehmerinnen scheinbar jünger, nachdem ihnen „When I'm Sixty-Four" von den Beatles vorgespielt wurde. Zumindest das Ergebnis der konzeptuellen Replikation ist eindeutig falsch.

In einer Studie mit 15 000 simulierten Stichproben ($N = 40$) verdeutlichen Simmons et al. (2011) den Einfluss von vier gebräuchlichen QRPs auf die Anzahl falsch positiver Ergebnisse (eines *t*-Tests). Beeindruckend an den Ergebnissen ist nicht nur, dass jede einzelne dieser Praktiken die Rate falsch Positiver in etwa verdoppelte, sondern mehr noch, dass die Kombination aller vier untersuchter QRPs gemeinsam, den Anteil falsch Positiver auf über 60% anhob. Es dürfte sich hier dennoch um konservative Schätzungen handeln, da nur ein kleiner Teil der möglichen, problematischen Verhaltensweisen in die Analyse miteinbezogen wurde.

In einer breit angelegten Umfrage unter wissenschaftlich arbeitenden Psychologen und Psychologinnen (John, Loewenstein, & Prelec, 2012) gaben über 90% der Befragten an,

zumindest eine von zehn untersuchten fragwürdigen Praktiken bereits genutzt zu haben. Gefragt wurde nicht nur, ob diese Praktiken von einem selbst bereits ausgeübt worden sind, sondern auch nach einer Einschätzung der Verbreitung dieser Vorgehensweisen unter den Kollegen und Kolleginnen des Faches. Weiters wurde nach einer Einschätzung des Anteils jener Psychologen und Psychologinnen ersucht, die den Einsatz solcher Praktiken eingestehen würden. Mit Hilfe der so gewonnenen Antworten wurde versucht die tatsächliche Verbreitung der QRPs abzuschätzen. Die Prävalenz einiger dieser Forschungspraktiken wurde auf bis zu 100% geschätzt. Im Folgenden wird eine Übersicht der wichtigsten QRPs und ihrer jeweiligen Prävalenz nach (John et al., 2012) gegeben:

5.1. Flexibilität bei der Wahl der abhängigen Variablen

Simmons et al. (2011) fanden, dass die Möglichkeit zwischen zwei abhängigen Variablen (die mit $r = .5$ korrelierten) zu wählen, die Rate falsch Positiver bei $\alpha = .05$ auf fast 10% anhob.[4] Werden also mehrere abhängige Variablen erhoben, aber nur jene abhängigen Variablen berichtet, die „funktionieren", also ein $p < .05$ liefern, wird die Rate falsch positiver Ergebnisse gesteigert. In der von John et al., (2012) durchgeführten Umfrage räumten 67% der Psychologen diese Vorgehensweise ein. Schätzungen zufolge dürfte das Nicht-Berichten von abhängigen Variablen noch häufiger vorkommen (bis zu 100%).

5.2. Flexibilität bei der Auswahl der Versuchsbedingungen

Die Möglichkeit eine von drei Versuchsbedingungen nicht zu berichten, ließ die Rate falsch positiver Ergebnisse auf 13% ansteigen (Simmons et al., 2011). 27% der Psychologen und Psychologinnen geben an dies bereits praktiziert zu haben; die „Dunkelziffer" dürfte bei etwa 70% liegen (John et al., 2012).

[4] Je niedriger die Korrelation zwischen den abhängigen Variablen ausfällt, desto höher ist zudem die Rate falsch positiver Ergebnisse: Bei $r = 1$ sind beide Variablen ident, und somit die Flexibilität geringer; während bei $r = 0$ die Variablen unabhängig sind, und somit die Flexibilität höher ist.

5.3. Flexibilität der Stichprobengröße (*data peeking*)

Werden im Verlauf der Datenakkumulierung wiederholt Signifikanztests durchgeführt, dann steigert das die Rate falsch positiver Testergebnisse erheblich (Armitage, Mcpherson, & Rowe, 1969; Francis, 2012). Wenn man sich also, nachdem man bereits ein nicht signifikantes Ergebnis vorfindet, dennoch entschließt, die Datenerhebung der Studie fortzusetzen, hat das Einfluss auf die Aussagekraft des *p*-Wertes. Bei John et al. bekannten sich 58% der Befragten zu dieser Praktik. Geschätzt wird die Prävalenz auf bis zu 100%. Die Möglichkeit einmal die Anzahl der Versuchspersonen pro Versuchsbedingung um 10 zu erhöhen, sofern zunächst noch kein signifikantes Ergebnis vorliegt, lässt die Rate falsch positiver auf etwa 8% ansteigen. Hat man die Möglichkeit im Verlauf der Datensammlung nach den ersten 10 Beobachtungen pro Versuchsbedingung, bei jeder zusätzlichen Beobachtung pro Bedingung, einen weiteren Signifikanztest durchzuführen, steigt die Wahrscheinlichkeit ein signifikantes Ergebnis zu erhalten auf 22% an (Simmons et al., 2011).

5.4. Flexibilität der Kovarianzanalyse

Simmons et al., (2011) untersuchten außerdem noch die mögliche Auswirkung der statistischen Kontrolle einer Drittvariable auf die Rate falsch Positiver: diese stieg auf beinahe 12% an.

5.5. Flexibilität der Forschungshypothesen oder HARKing

Kerr (1998) versteht unter *HARKing* (Hypothesizing After the Results are Known) Post-Hoc-Hypothesen als A-Priori-Hypothesen darzustellen. D.h. Hypothesen, die erst aufgrund der vorliegenden Datenlage zustande gekommen sind, werden berichtet, als ob sie bereits vor der Sammlung der Daten formuliert worden wären. 35% der Teilnehmer der Umfrage von John et al. (2012) gaben an bereits unerwartete Ergebnisse so dargestellt zu haben, als ob sie diese vorhergesagt hätten; Schätzungen zur Verbreitung dieses Vorgehen belaufen sich auf etwa 90%.

5.6. Flexibilität beim (Ab-)Runden

Abrunden der *p*-Werte scheint eine verbreitete Vorgehensweise zu sein, obwohl es offensichtlich sein sollte, dass so die Rate falsch positiver Ergebnisse beeinflusst wird. Es geht um die Praxis einen *p*-Wert von z.B. $p = .054$ als $p < .05$ darzustellen. 23% geben an dies schon einmal praktiziert zu haben (John et al., 2012); Schätzungen belaufen sich auf über 60%. Hartgerink, van Aert, Nuijten, Wicherts und van Assen (2016) zeigen diese Abrundungstendenz in der psychologischen Literatur, indem sie *p*-Werte, die als $p = .05$ berichtet wurden, aus den berichteten Teststatistiken neu berechneten. 2/3 der rekalkulierten *p*-Werte lagen über .05, während nur 1/3 der neu berechneten *p*-Werte kleiner als .05 ausfielen.

5.7. Selektives Berichten

In der Umfrage von John et al. (2012) bekennen sich 50% (geschätzt 100%) dazu selektiv nur jene Studien in einem Paper zu berichten, die im Endeffekt „funktionierten", also brauchbare (i.e. signifikante) Ergebnisse lieferten.

5.8. Flexibilität beim Datenausschluss

42% (geschätzt 100%) haben schon Daten aus einem Datensatz ausgeschlossen, erst nachdem ihnen die Auswirkung dessen bereits bekannt war.

5.9. Gefälschte Daten

Ob gefälschte Daten als *p*-Hacking angesehen werden sollten, oder ob es sich schlichtweg um Betrug handelt, kann diskutiert werden. Das Problem scheint allerdings in der Psychologie zumindest nicht jenes Ausmaß anzunehmen, wie die anderen hier dargestellten Praktiken. Knapp 2% der befragten Psychologen und Psychologinnen gaben zu bereits einmal Daten fingiert oder manipuliert zu haben (John et al., 2012); ähnliche Schätzungen (2%) findet auch Fanelli (2009). Berühmt geworden als systematischer Datenfälscher ist der Sozialpsychologe Diederik Stapel (Stroebe, Postmes, & Spears, 2012); solche Fälle erregen viel Aufmerksamkeit. Zumindest hat der Fall von Diederik Stapel die

Diskussion um die Notwendigkeit von Replikationen angetrieben (Pashler & Wagenmakers, 2012).

6. Das vernachlässigte Konzept der statistischen Teststärke

Die Berechnung der statistischen Teststärke oder Power wird erst möglich durch die Spezifikation einer Alternativhypothese, wie es Neyman und Pearson vorsehen (Gigerenzer, 2004, Kline 2013). β ist in diesem Kontext die Wahrscheinlichkeit einen Fehler 2. Art zu begehen, also die Wahrscheinlichkeit die Nullhypothese nicht zu verwerfen, obwohl die Alternativhypothese zutrifft. Die Teststärke (1-β) ist die Komplementärwahrscheinlichkeit zu β und bezeichnet somit die Wahrscheinlichkeit die Nullhypothese korrekterweise abzulehnen (bzw. ein statistisch signifikantes Ergebnis zu erhalten), wenn die Alternativhypothese wahr ist (Cohen, 1988, 1992). Die Teststärke ist eine Funktion der Stichprobengröße (*N*), der Effektstärke in der Population (*ES*) und des Signifikanzniveaus (α); sie wächst mit der Stichprobengröße und mit der Größe des Populationseffekts, sinkt aber mit strengerem Signifikanzniveau. Neyman und Pearson sahen ursprünglich vor, bei der Festlegung von α und β, Kosten-Nutzen-Überlegungen miteinfließen zu lassen. Tatsächlich wird in der Psychologie α fast dogmatisch bei α = .05 fixiert. Explizite Überlegungen zu β bzw. der Teststärke (1-β) und Poweranalysen fehlen meist, oder beziehen sich der Konvention folgend sich auf die Empfehlung von Cohen (1988) eine Power von 80% anzustreben.

Das Fehlen von Poweranalysen ist insofern auch problematisch, da die subjektive Einschätzung der Power bei gegebener Effektstärke, Stichprobengröße und α sehr ungenau ist. Die Power für kleine Effekte wird tendenziell überschätzt, während die Teststärke große Effekte zu finden von Psychologen und Psychologinnen oft unterschätzt wird. (Bakker, Hartgerink, Wicherts, & van der Maas, 2016).

6.1. Niedrige Teststärke psychologischer Forschung

Generell ist die Teststärke psychologischer Studien sehr niedrig. Cohen (1962) fand im *Journal of Abnormal and Social Psychology* eine durchschnittliche Power von 48 % für

mittlere Effekte. Sedlmeier und Gigerenzer (1989) untersuchten die Teststärke dieses Journals 24 Jahre später und konnten keine Verbesserung hinsichtlich der Teststärke feststellen. Im Gegenteil, sie kritisieren die Verbreitung der Alphaadjustierung, da dadurch die Power noch weiter sinke. Bakker, van Dijk und Wicherts (2012) schätzen die durchschnittliche Power in der Psychologie, auf Grundlage der typischen Effektstärke (d = 0.5) und der mittleren Stichprobengröße psychologischer Studien (N = 40) auf 35%. Fraley und Vazire (2014) berechnen die Teststärke auf Grundlage der Stichprobengrößen verschiedener sozial- bzw. persönlichkeitspsychologischer Journale. Sie schätzten so die Teststärke der meisten Journale auf etwa 50% einen mittleren (r = .20 bzw. d = 0.41) Effekt zu finden. Jedenfalls bleiben alle Schätzungen weit ab von Cohens Empfehlung von 80%.

6.2. Folgen niedriger Teststärke

Es stellt sich die Frage, inwieweit experimentelle Studien in der Psychologie Ressourcen verschwenden, wenn die Wahrscheinlichkeit Effekte zu finden derart gering bleibt, und somit viele Studien gar nicht in der Lage sind, die Effekte zu finden nach denen sie suchen. Davon abgesehen führt niedrige Teststärke vor allem zu folgenden Problemen: Zunächst wird die Wahrscheinlichkeit einen Effekt zu finden geschmälert; der Anteil falsch Positiver Effekte in der Literatur erhöht sich, und Effekte werden generell in ihrem Ausmaß überschätzt (vgl. Button et al., 2013). Die Kombination mit QRPs verstärkt die Problematik zusätzlich. Insgesamt führt die niedrige Teststärke in der Psychologie zu einer Ansammlung von Studien, deren Zuverlässigkeit zweifelhaft erscheint.

Wahrscheinlichkeit einen vorhandenen Effekt zu entdecken

Das grundlegendste Problem niedriger Teststärke ergibt sich aus der Definition über den Fehler 2. Art (1-β). Studien mit niedrigerer Teststärke haben eine höhere Wahrscheinlichkeit einen Fehler 2. Art zu begehen, also falsch negative Ergebnisse zu liefern. Gleichzeitig sinkt mit der Power die Wahrscheinlichkeit einen vorhandenen Effekt zu finden.

Positiver Vorhersagewert

Studien mit niedriger Power haben aber nicht nur eine geringere Wahrscheinlichkeit einen wahren Effekt zu entdecken; sie erhöhen auch den Anteil falsch positiver Ergebnisse in der Literatur. Je geringer die Teststärke, desto geringer ist die Wahrscheinlichkeit, dass ein signifikantes Ergebnis einen wahren Effekt widerspiegelt (*positive predictive value*), d.h. dass es sich um kein falsch positives, sondern um ein richtig positives Ergebnis handelt (Ioannidis, 2005).

Ioannidis leitet diese Wahrscheinlichkeit folgendermaßen ab: Er definiert R als Verhältnis wahrer zu falscher Zusammenhänge, die innerhalb einer Disziplin geprüft werden; folglich ist $R/(R+1)$ der Anteil wahrer Zusammenhänge an den untersuchen (wahren und falschen) Zusammenhängen, bzw. die Wahrscheinlichkeit, dass eine zu prüfende Hypothese wahr ist. Dieser Quotient unterscheidet sich von Disziplin zu Disziplin und ist jedenfalls unbekannt. Angenommen es werden nun c Zusammenhänge untersucht, dann erhält man ...

$$c(1 - \beta)R/(R + 1) \qquad (1)$$

... wahre Zusammenhänge, die entsprechend als signifikant gefunden werden, und ...

$$c\alpha(R + 1) \qquad (2)$$

... falsche Zusammenhänge, die dennoch als signifikante Ergebnisse gefunden werden. Der Anteil von (1) an der Summe von (1) + (2) ergibt dann die Wahrscheinlichkeit, dass ein signifikantes Ergebnis auch wahr ist. Ioannidis bezeichnet diese Wahrscheinlichkeit als Positiven Vorhersagewert (*positive predictive value*, *PPV*):

$$\text{PPV} = [(1 - \beta)R] / [(1 - \beta)R + \alpha] \qquad (3)$$

Das PPV hängt somit auch von der Teststärke ab: Je geringer diese bemessen ist, desto geringer auch das PPV[5], vorausgesetzt die anderen Faktoren bleiben konstant.

[5] Ist $(1- \beta)R > \alpha$, dann ist es wahrscheinlicher, dass ein Ergebnis falsch ist, als dass es wahr ist.

Ist die Teststärke gering, dann werden nicht nur mehr Fehler 2. Art begangen, also vorhandene Effekte nicht gefunden; es steigt gleichzeitig auch die Anzahl der Fehler 1. Art in der Literatur, also die Anzahl gefundener Effekte, die es nicht gibt. Das liegt daran, dass sich durch geringe Power das Verhältnis falsch Positiver zu richtig Positiver in ungünstiger Weise verändert. Einfach ausgedrückt wird der Anteil richtig positiver Ergebnisse an allen positiven Ergebnissen bei geringerer Teststärke niedriger.

Die Idee des PPV entspricht konzeptuell dem des kritischen Alpha (α^*) bzw. kritischen Beta (β^*) wie es O'Brien und Castelloe (2007) schildern. Hier wird das kritische Alpha (*crucial Type I error rate*) als Wahrscheinlichkeit definiert, dass die Nullhypothese wahr ist, wenn $p < \alpha$ und demzufolge die Nullhypothese abgelehnt wird. Dementsprechend ist das kritische Beta (*crucial Type II error rate*) die Wahrscheinlichkeit, dass die Nullhypothese falsch ist, wenn $p > \alpha$ und die Nullhypothese nicht abgelehnt wird. Bleiben alle anderen Einflussfaktoren konstant, senkt höhere Teststärke beide kritischen Fehlerraten. Folglich sind statistische Ergebnisse zuverlässiger bei hoher Teststärke.

Überschätzte Effektgrößen (Winner's Curse)

Ein weiteres Problem geringer Teststärke ergibt sich dadurch, dass kleine Effekte in ihrer Größe überschätzt werden müssen, um bei geringer Power als statistisch signifikant gefunden zu werden. Wenn Studien mit geringer Teststärke also einen bisher unbekannten signifikanten Effekt finden, ist dieser Effekt wahrscheinlich in seinem Ausmaß überschätzt (Ioannidis, 2008). Ursache für zu hohe Effektgrößen ist neben der geringen Teststärke, auch die Tatsache, dass der Schwellenwert der statistischen Signifikanz ($p < .05$) unterschritten werden muss, um einen Effekt anzunehmen: Kleine Stichproben bringen oft nur genug Teststärke auf, um einen großen Effekt als statistisch signifikant zu finden. Effekte, die eigentlich zu klein sind um mit gegebener Power entdeckt zu werden, findet man nur dann, wenn sie zufällig in der Stichprobe in ihrem Ausmaß überschätzt werden, und somit signifikant werden können. Dieser Effekt wird oft auch mit dem Fluch des Gewinners (*Winner's Curse*) in Verbindung gebracht, wonach bei Auktionen, der Höchstbieter tendenziell zu viel bezahlt, vor allem, wenn der exakte Wert des ersteigerten Gegenstandes nicht bekannt ist (Young, Ioannidis, & Al-Ubaydli, 2008). Die Analogie ist jene, dass die Person, die zufällig den Wert überschätzt, sowohl bei der Auktion als auch beim Signifikanztest zum Zug kommt bzw. „gewinnt".

Niedrige Power + QRPs = besonders viele (falsch) positive Ergebnisse

Kommt zu den rein mathematischen Problemen niedriger Power auch noch die Verwendung fragwürdiger Forschungspraktiken, wie es in der Psychologie nicht unüblich ist (John et al., 2012) hinzu, wird das Problem niedriger Teststärke besonders eklatant. In einer Simulationsstudie konnten (Bakker et al., 2012) zeigen, dass ein signifikanter Effekt häufiger gefunden werden kann, wenn mit einer vorhandenen Stichprobe, nicht nur eine Studie durchgeführt wird, sondern diese Stichprobe aufgeteilt wird, um mehrere kleinere Studien durchzuführen. Hat man also ein Sample von z.B. $N = 200$ findet man eher einen Effekt, wenn man die zur Verfügung stehende Stichprobe auf mehrere Studien (z.B. vier Studien mit $N = 50$) aufteilt, anstatt nur eine Studie durchzuführen. Diese Strategie ist besonders lohnend, wenn die Effektstärken von vorn herein klein sind – die Teststärke also noch weiter sinkt – und zusätzlich auch QRPs genützt werden. D.h. Wissenschaftler kommen eher zu publizierbaren Ergebnissen, wenn die von ihnen durchgeführten Studien eine nur geringe Power aufweisen, und sie zusätzlich QPRs anwenden. Bakker et al. (2012) zeigen aber auch, dass durch die Nutzung solcher Strategien nicht nur die ermittelten Effektstärken überschätzt werden, sondern auch, dass so die Rate falsch positiver Ergebnisse auf bis zu 40% ansteigt.

7. Bestimmung der Stichprobengröße direkter Replikationen

7.1. Stichprobengröße der Originalstudie

Möchte man eine direkte Replikation durchführen, ist es nicht empfehlenswert einfach den Stichprobenumfang der ursprünglichen Studie heranzuziehen. Wird nämlich bei der Replikation die gleiche Stichprobengröße verwendet, wie in der Originalstudie, und war der p-Wert dieser Studie nah am Signifikanzniveau (d.h. $p \sim .05$), dann erreicht man eine Teststärke von etwa 50%, und zwar auch nur dann, wenn der Effekt in der Originalstudie richtig geschätzt wurde (vgl. Button et al., 2013); wurde der Effekt überschätzt, sinkt die Teststärke einer Replikation mit gleichem Stichprobenumfang entsprechend.

7.2. Poweranalyse

Eine etwas bessere Möglichkeit den Stichprobenumfang einer Replikation zu bestimmen, bietet die Durchführung einer Poweranalyse. Als Schätzer für die erwartete Effektstärke wird das Ergebnis der ursprünglichen Studie zugrunde gelegt. Wie bereits gezeigt, muss man davon ausgehen, dass die ursprüngliche Schätzung eines Effekts diesen überschätzt (vgl. *Winner's Curse*), und damit auch die so berechnete Teststärke der Replikation zu optimistisch ist. Simonsohn (2015) zeigt wie sich niedrige Power und das Signifikanzniveau und somit überschätzte Effektgrößen auf die tatsächliche Teststärke von Replikationen auswirken, deren Stichprobengröße mittels Poweranalyse auf Grundlage der ursprünglichen Effektschätzung berechnet wurde. Je niedriger die Teststärke der Originalstudie ausfällt, desto ausgeprägter wird die Teststärke der Replikation in einer Poweranalyse auf Grundlage des ursprünglichen Effekts überschätzt. Wird in der Replikation eine Power von 80% angestrebt, und hatte die Originalstudie 50% Power, dann erzielt man aufgrund einer so durchgeführten Poweranalyse lediglich 51% Power in der Replikation; hatte die Originalstudie nur 30% Power, dann hat die Replikation eine Power von 39%, obwohl in der Poweranalyse 80% Power angestrebt werden.

7.3. Small Telescopes oder Detectability

Simonsohn (2015) schlägt einen anderen Ansatz zur Bestimmung der Stichprobengröße von Replikationen vor, der bei der Bewertung von Replikationsergebnissen ansetzt. Eine häufige und naheliegende Methode um das Ergebnis einer Replikation zu evaluieren, ist zu prüfen ob dieses Ergebnis wieder signifikant ist (und zwar in die gleiche Richtung wie das Originalergebnis). Darauf zielt auch die Durchführung einer Poweranalyse ab. Man berechnet eine Stichprobengröße, die es der Replikation ermöglichen soll, mit einer bestimmten Wahrscheinlichkeit ein signifikantes Ergebnis vorzufinden, sofern dieses vorhanden ist. Dieses Vorgehen muss jedoch vor allem im Kontext von Replikationen kritisch gesehen werden. Ein signifikantes Ergebnis einer Replikation bestätigt zwar die Originalstudie, und stärkt das Vertrauen in den gezeigten Effekt. Ein nicht signifikantes Ergebnis einer Replikation widerlegt diesen aber nicht; es kann in Einklang mit der ursprünglichen, signifikanten Schätzung eines Effektes stehen; z.B.

wenn der Effekt in die gleiche Richtung deutet, wie in der Originalstudie, aber nicht groß genug ist, um mit der vorliegenden Stichprobengröße signifikant zu werden.

Eine Signifikanztestung hat das Ziel zu zeigen, dass ein Effekt signifikant von Null verschieden ist; das Konfidenzintervall des Effekts die Null also nicht umfasst. Eine Replikation will aber zeigen, sofern ein Effekt nicht vorhanden ist, dass dieser in etwa Null ist. Das ist mit der klassischen Signifikanzprüfung nicht möglich.

Simonsohn (2015) wählt daher einen anderen Ansatz. Anstatt zu zeigen, dass ein Effekt nicht vorhanden ist, soll gezeigt werden, dass die ursprüngliche Studie nicht dazu geeignet war, den Effekt zu untersuchen, weil dieser Effekt dazu zu klein ist. Um diese Idee zu verdeutlichen, illustriert er eine Analogie zu Teleskopen: Ein Astronom, der nur ein kleines Teleskop zur Verfügung hat, behauptet einen Planten entdeckt zu haben. Daraufhin versucht ein anderer Astronom mit einem größeren Teleskop, den neuen Planeten zu sichten, findet diesen aber nicht. Zwar widerlegt er damit die Existenz dieses Planeten nicht, aber zumindest steht das Nichtfinden des Planeten mit einem größeren Teleskop im Widerspruch zur ursprünglichen Sichtung. Ein Planet der mit einem kleinen Teleskop gefunden werden kann, sollte auch mit einem größeren Teleskop gefunden werden können.

Simonsohn (2015) definiert nun jene Effektgröße, für welche die Originalstudie mit der ursprünglichen Stichprobengröße eine Power von 33% erzielt hätte, als $d_{33\%}$. Es wird angenommen, dass eine Studie nicht in der Lage ist, einen derart kleinen Effekt mit der genutzten Stichprobengröße bedeutsam zu untersuchen. Wenn die Replikation nun zeigen kann, dass sich der untersuchte Effekt signifikant von $d_{33\%}$ unterscheidet, kann angenommen werden, dass die ursprüngliche Studie zu klein war, um den Effekt sinnvoll zu erfassen.

Um die Stichprobengröße einer Replikation zu bestimmen, kann nach Simonsohn (2015) ebenso eine Poweranalyse durchgeführt werden. Allerdings unterscheiden sich die Null- und Alternativhypothese dieses Ansatzes. Man möchte mit einer bestimmten Wahrscheinlichkeit zeigen können, dass der Effekt der Replikation sich von $d_{33\%}$ unterscheidet (H_0), und erwartet, dass der tatsächliche Effekt gleich Null ist (H_1). Simonsohn (2015) zeigt aber, dass eine Replikation immer etwa 2.5-mal die Anzahl der Beobachtungen der Originalstudie braucht, um eine Power von 80% zu erzielen, die Nullhypothese von $d_{33\%}$ abzulehnen, wenn der wahre Effekt gleich Null ist. Somit ist die Durchführung einer Poweranalyse obsolet. D.h. nach Simonsohn (2015) sollte die Stichprobengröße der

Replikation 2.5-mal größer sein als jene der Originalstudie, sofern man zeigen möchte, dass diese nicht in der Lage war den gezeigten Effekt bedeutsam festzustellen.

7.4. Safeguard Power

Ein weiteres Konzept, das bei der Bestimmung der Stichprobengröße von Replikationen das Ziel verfolgt, diese vor überschätzten Originaleffekten und somit unterschätzter Teststärke abzusichern, ist die *Safeguard Power* von Perugini, Gallucci und Costantini (2014). Anstatt einer Poweranalyse die Effektgröße der Originalstudie zugrunde zu legen, wird ein entsprechendes Konfidenzintervall, bzw. dessen untere Grenze als Effektschätzer in der Poweranalyse herangezogen. Die Idee ist folgende: Die Originalstudie bietet nur eine Punktschätzung eines Effekts, der wahre Wert ist unbekannt. Wird der wahre Wert in der Originalstudie unterschätzt und auf Grundlage der Schätzung die Stichprobengröße anhand einer Poweranalyse bestimmt, hat das keine gravierenden Folgen für die tatsächliche Power. In diesem Fall wäre die tatsächliche Teststärke höher als die in der Poweranalyse angestrebte Teststärke. Wird der wahre Wert in der Originalstudie aber überschätzt, dann sind die Konsequenzen problematisch, denn dann ist die tatsächliche Teststärke niedriger als die angestrebte. Um dem entgegenzuwirken wird als Schätzer der Effektgröße, nicht der Effekt der Originalstudie (d_0) eingesetzt; es wird zunächst ein Konfidenzintervall um d_0 bestimmt und dann die untere Grenze dieses Konfidenzintervalls (d_s) als Grundlage für die Poweranalyse eingesetzt. Perugini et al. empfehlen hier die Berechnung eines 60% Konfidenzintervalls, damit das Risiko, dass der tatsächliche Wert kleiner ist als die untere Grenze des Konfidenzintervalls auf 20% minimiert wird. So wird in 80% der Fälle zumindest die angestrebte Teststärke mittels der in der Poweranalyse errechneten Stichprobengröße erreicht.

8. Reproducibility Project: Psychology

8.1. Einführung und Ziele des Reproducibility Project: Psychology

Replikationen sind also für wissenschaftliches Arbeiten von zentralem Wert, auch weil sie in der Lage sind die Evidenz von Effekten zu überprüfen. Ergebnisse, welche

möglicherweise nur zufällig zustande gekommen sind, bergen Gefahren für die Wissenschaft den falschen Fährten zu folgen bzw. Forschungsstränge zu eröffnen, die ins Leere führen. Replikationen werden jedoch nur selten publiziert und wohl auch selten durchgeführt. Die Anreizstrukturen fördern Innovation, nicht Konfirmation. Niedrige Teststärke und *p*-Hacking lassen an der Aussagekraft psychologischer Studien, die auf Nullhypothesentests beruhen Zweifel aufkommen. Aus solchen Überlegungen heraus entstand das *Reproducibilty Project: Psychology* (Open Science Collaboration, 2012, 2015).

Die grundlegende Idee des Reproducibility Project: Psychology (RP:P) ist simpel: Um die Replizierbarkeit psychologischer Studien empirisch abschätzen zu können, wird eine quasi-zufällige Stichprobe von psychologischen Studien ($N = 100$) tatsächlich repliziert. Ein Vorhaben, das für einzelne Personen oder Institutionen kaum realisierbar erscheint, und somit im Rahmen eines kollaborativen Projekts mit über 270 Beteiligten durchgeführt wurde.

Publiziert wurden die 100 replizierten Studien ursprünglich im Jahr 2008 in drei prominenten psychologischen Fachjournalen, namentlich dem *Journal of Personality and Social Psychology*, *Journal of Experimental Psycholgy: Learning, Memory and Cognition* und *Psychological Science*. Die Auswahl der Fachjournale sollte ein Spektrum psychologischer Disziplinen abdecken, um diese auf Unterschiede in der Replizierbarkeit untersuchen zu können. Auch das Publikationsjahr 2008, wurde mit Bedacht gewählt. Die Originalstudien sollten nicht zu lange in der Vergangenheit liegen, um nach Möglichkeit auf die ursprünglichen Materialien und Instruktionen zurückgreifen zu können. Andererseits wollte man ein Set von Studien replizieren, welches in der Anzahl der Zitationen bereits variierte, um den Impact der Studien als möglichen Prädiktor der Replizierbarkeit in die Analysen miteinzubeziehen.

Somit verfolgte das Projekt primär zwei Ziele: Erstens sollte eine empirische Abschätzung der Replizierbarkeitsrate psychologischer Forschung erfolgen, und zweitens wollte man mögliche Variablen, die Einfluss auf die Replizierbarkeit psychologischer Studien haben, identifizieren.

8.2. Konkreter Ablauf der Replikationen

Der Ablauf jeder Replikation wurde durch ein detailliertes Replikationsprotokoll festgelegt, und die Replikationsteams wurden bei dem Prozess von zwei

Projektkoordinatorinnen begleitet. Am Beginn jeder Replikation stand die Auswahl der zu replizierenden Studie. Um die Möglichkeit eines Selektionsbias zu minimieren, wurden zu replizierende Studien nur schrittweise zur Auswahl freigegeben. Für jede Studie wurde im Voraus ein Schlüsseleffekt identifiziert. Dieser bezog sich, sofern sich ein Artikel aus mehreren Studien zusammensetzte, auf die letzte Studie, welche auch die zu Replizierende darstellte (Open Science Collaboration, 2015).

Die Originalautoren und -autorinnen der Studien wurden von Beginn an in die Replikation miteinbezogen; der erste Kontakt fand statt, um zu eruieren, ob die Möglichkeit bestand, die Originalmaterialien der Studien zu erhalten, sofern diese noch vorhanden waren. Auch um Hinweise hinsichtlich der Durchführung der Replikation wurden die Originalautoren gebeten. Details der Studie, die nicht aus dem publiziertem Artikel hervorgehen, wurden abgeklärt und eventuelle Unklarheiten ausgeräumt.

Ein Replikationsbericht („Replication Report“) wurde bereits vorab angefertigt, und enthielt nicht nur ein Protokoll der durchzuführenden Studie, sondern auch einen konkreten statistischen Analyseplan. Zum einen wurden die Replikationen somit präregistriert, zum anderen konnten Änderungen, die erst im Nachhinein vorgenommen wurden, einfach verfolgt werden. Der Replikationsbericht wurde geprüft, und auf der Projekthomepage veröffentlicht. Auch die Materialien bzw. das Experimentalprogramm wurden hier bereits vor der eigentlichen Durchführung der Replikation online archiviert (osf.io/ezcuj).

Im Anschluss an die Datensammlung wurden die Daten und die konfirmativen Analysen veröffentlicht; ebenso ein um die Ergebnisse ergänzter Replikationsbericht. Diese Vorgehensweise der frei zugänglichen Archivierung der Daten, Materialien und der Auswertung maximiert nicht nur die Transparenz, sondern auch die Verantwortlichkeit jedes und jeder am Projekt beteiligten. Zur Qualitätssicherung wurden die statistischen Analysen von einem an der Replikation unbeteiligten Dritten reproduziert[6]. Der Replication Report wurde den Originalautoren und Originalautorinnen zur Durchsicht übermittelt, zudem auch innerhalb des Projekts überprüft, und gegebenenfalls adaptiert. Insgesamt wurden auf diese Weise 100 Replikationen durchgeführt.

[6] Asendorpf und Conner (2013) werten die Reproduzierbarkeit eines Resultats aus dem Datenset als Voraussetzung für dessen Replizierbarkeit. Damit ist gemeint, dass ein Auswerter B, die exakt gleichen Ergebnisse erhält, die zuvor Auswerter A berichtet hat, wenn beide dasselbe Datenset mit den gleichen statistischen Methoden auswerten.

8.3. Teststärke der Replikationen

Teststärkenberechnungen wurden auf Grundlage der jeweiligen Schlüsseleffekte der Originalstudien durchgeführt. Minimale Anforderung an Replikationen waren 80% Power. Einige Studien erzielten jedoch 90% bzw. 95% Power. Die mittlere Teststärke der Replikationsstudien des RP:P belief sich auf 92%.

8.4. Ergebnisse des Reproducibility Project: Psychology

Es existiert keine Standardmethode zur Evaluation von Replikationsergebnissen. Um die Ergebnisse der Replikationen des RP:P zu beurteilen, wurden verschiedene Ansätze in die Analysen miteinbezogen (Open Science Collaboration, 2015): Erstens wurde der Anteil signifikanter Ergebnisse ($p < .05$) unter den Replikationen mit dem Anteil signifikanter Ergebnisse unter den Originalstudien verglichen. Zweitens wurden die Effektstärken von Original- und Replikationsstudien gegenübergestellt. Dazu wurden alle relevanten Effektstärken auf ein gemeinsames Maß (Korrelationskoeffizient *r*) gebracht. Die Effektgrößen der Originalstudien wurden durchgängig positiv kodiert, die Effektgrößen der Replikationsstudien wurden negativ kodiert, wenn der Effekt in die entgegengesetzte Richtung der Originalstudie deutete. Zusätzlich interessierte der Anteil an originalen Effektgrößen, die vom 95%-Konfidenzintervall der Effektgrößen der Replikation umfasst wurden. Drittens wurden die meta-analytischen Kombinationen von Original- und Replikationsstudien auf ihre Signifikanz geprüft. Und zuletzt wurde der Erfolg der Replikationen auch subjektiv von den Replikationsteams bewertet.

Um mögliche Korrelate der Replizierbarkeit zu identifizieren, wurden neben dem Journal und der Subdisziplin zahlreiche weitere potenzielle Indikatoren der Replizierbarkeit in den Analysen berücksichtigt. Zunächst sechs Maße, die sich jeweils auf die Originalstudie beziehen: *p*-Wert, Effektgröße, Freiheitsgrade des statistischen Tests, Bedeutung des Effekts, eine Bewertung wie überraschend der Effekt war, und die Expertise der Originalautoren und Autorinnen.

Außerdem fanden sieben Charakteristika der Replikationsstudie Berücksichtigung: *p*-Wert, Effektgröße, Power (die Berechnung erfolgte auf Grundlage der ursprünglichen Effektgrößenschätzung und der Stichprobengröße der Replikation), Freiheitsgrade des statistischen Tests, eine Bewertung dessen, wie herausfordernd sich die Replikation

darstellte, eine Bewertung der Qualität der Replikation durch die Replikationsteams, und die Expertise der Replikationsteams.

Anteil der signifikanten Ergebnisse

Die bekannteste Methode das Ergebnis einer Replikation zu bewerten, ist zu prüfen, ob die Replikation wieder einen signifikanten Effekt ($p < .05$) in die gleiche Richtung wie die Originalstudie zeigt (*vote-counting*). 97 der 100 Originalstudien (97%) hatten positive Ergebnisse. Unter der Annahme, dass alle Originaleffekte akkurat geschätzt wurden, sind auf Grundlage der durchschnittlichen Teststärke der Replikationsstudien ($M = .92$) unter den Replikationen 89 statistisch signifikante Ergebnisse zu erwarten; tatsächlich waren es nur 35 (36%; 95%-KI = [26.6%, 46.2%]). Das Problem dieser Vorgehensweise ist aber, dass womöglich einige der Ergebnisse das Signifikanzkriterium aufgrund zu niedriger Teststärke verfehlt haben (Open Science Collaboration, 2015).

Effektgröße der Replikation im Vergleich zur Effektgröße der Originalstudie

Der Vergleich der Höhe der Effektstärken von Original- und Replikationsstudie ermöglicht eine Bewertung der Replikationen ohne auf deren *p*-Werte angewiesen zu sein. Die originalen Effekte ($M = 0.40$, $SD = 0.19$) waren statistisch signifikant größer als die Effekte der Replikationen ($M = 0.20$, $SD = 0{,}26$); Wilcoxon's $W = 7137$, $p < .001$).

Eine weitere Möglichkeit das Ergebnis einer Replikation zu evaluieren, ist es das Konfidenzintervall der Effektstärke der Replikation zu bestimmen, und zu eruieren ob dieses Konfidenzintervall die Effektstärke der Originalstudie überdeckt. Insgesamt war das bei den Replikationsstudien des RP:P in 47% der Fall.

Kumulative Evidenz

Die meta-analytische Kombination von Original und Replikationseffekten ermöglicht eine Bewertung der kumulativen Evidenz, wie auch eine Bewertung der Präzision der Evidenz. 51 von 75 (68%) der Ergebnisse bleiben signifikant, d.h. das 95%-Konfidenzinterfall umfasst die Null nicht (Fixed-Effects-Modell).

Subjektive Bewertung der Replikationsteams

Neben den quantitativen Maßen des Replikationserfolgs, erfolgte eine subjektive Bewertung der Replikation durch die Replikationsteams. Die Frage „Did your results replicate the original effect?“ (Open Science Collaboration, 2015, S. aac4716-4) wurde von 39 Replikationsteams bejaht.

Korrelate der Replizierbarkeit

Es zeigte sich eine bessere Replizierbarkeit kognitivpsychologischer als sozialpsychologischer Forschung. Es konnten 21 von 42 (50%) der kognitionspsychologischen Effekte repliziert werden, aber nur 14 von 55 (25%) der sozialpsychologischen Effekte, wenn als Kriterium die erneute Signifikanz ($p < .05$) gewählt wurde. Eine mögliche Erklärung ist, dass die Originaleffekte in der Sozialpsychologie (v.a. jene in JPSP) schwächer waren, als die ursprünglichen kognitionspsychologischen Effekte.

Der statistische Test mit dem der Effekt gefunden wurde, spielte eine Rolle für den Replikationserfolg. 23 von 49 (47%) der Studien, die einen einfachen oder Haupteffekt testeten, konnten repliziert werden, während nur 8 von 37 (22%) der Interaktionseffekte repliziert werden konnten ($p < .05$ als Kriterium).

Replikationserfolg und originaler *p*-Wert korrelieren negativ miteinander ($r_s = -.33$). Zwei Drittel der ursprünglichen Studien mit $p < .001$ zeigten in der Replikation wieder signifikante p-Werte. Ebenso waren größere originale Effektgrößen mit höherer Wahrscheinlichkeit in den Replikationen signifikant ($r_s = .30$), aber gleichzeitig zeigten jene Replikationen deren originale Effektgrößen groß waren, eine höhere Differenz zwischen originaler und replizierter Effektgröße ($r_s = .28$). Auch die Teststärke der Replikation korrelierte positiv mit dem Replikationserfolg ($r_s = .37$; aber nicht mit der Differenz der Effektgrößen zwischen Original- und Replikationsstudien.

Überraschende Effekte ($r_s = -.24$) und herausfordernde Replikationen ($r_s = -.22$) korrelierten negativ mit dem Replikationserfolg.

Kein Zusammenhang zeigte sich zwischen der Expertise der Originalautoren und Originalautorinnen ($r_s = -.07$) bzw. der Expertise der Replikationsteams ($r_s = -.10$) und der Signifikanz der Replikationsstudien.

Zusammenfassung

Insgesamt folgerten die Autoren (Open Science Collaboration, 2015), dass kein einziger Indikator die Ergebnisse der Replikationen ausreichend beschreiben kann. Alle fünf gewählten Maße der Replizierbarkeit korrelierten aber positiv untereinander; r_s bewegte sich zwischen .22 bis .96, Mdn_r = .57. Und gemeinsam deuten sie darauf hin, dass der Großteil der Replikationen eine schwächere Evidenz für die Originalergebnisse erbrachte.

Die korrelativen Tests zeigen, dass die Stärke der originalen Evidenz den besten Prädiktor des Replikationserfolgs darstellte.

8.5. Diskussion

Das Projekt erhielt breite mediale Aufmerksamkeit, auch außerhalb psychologischer Fachkreise. Die Berichterstattung reichte von renommierten Printmedien wie der New York Times, dem Guardian über den Spiegel bis zu namhaften Onlineformaten wie wired.com oder vox.com (siehe Altmetric, 2016).

Die wohl lauteste Kritik erschien, wie das Projekt selbst, in Science und kam von dem Sozialpsychologen Daniel Gilbert und einigen Kollegen, die drei statistische Fehler des RP:P gefunden haben wollen, und so zeigen wollten, dass die Replizierbarkeit psychologischer Forschung sogar eher hoch sei; gleichzeitig erschien eine Gegenreplik einer Teilgruppe der Open Science Collaboration (Anderson et al., 2016); es folgten zudem unzählige Blogbeiträge, die die statistischen Fehler von Gilbert et al. aufzeigten (z.B. Srivastava, 2016; Gelman, 2016; Lakens, 2016; Simonsohn, 2016)

Erwähnenswert erscheint vor allem der Beitrag von Simonsohn (2016), der auf Probleme der verschiedenen Möglichkeiten der Bewertung von Replikationsergebnissen hindeutet. Ein Ansatz des RP:P betraf den Vergleich des Konfidenzintervalls der Replikation mit der Effektgröße der Originalstudie. Gilbert et al. (2016) präferieren das Vorgehen, das Konfidenzintervall des Originaleffekts zu bestimmen, und mit der Effektgröße der Replikation zu vergleichen. Beide Vorgehensweisen können, wie Simonsohn zeigt, zu verzerrten Schlussfolgerungen führen. Interpretiert man nur jene Replikationen, deren Konfidenzintervalle die Originaleffekte umfassen als erfolgreich, erhält man die meisten nicht geglückten Replikationen, wenn die Teststärke besonders hoch ist. Dann ist nämlich das Konfidenzintervall des Replikationseffekts besonders präzise bzw. eng, und umfasst in vielen Fällen die womöglich überschätzten Effekte der Originalstudien nicht mehr.

Die andere Methode, also die Frage, ob das Konfidenzintervall der Originalstudie den Replikationseffekt umfasst, führt allerdings zu folgender paradoxen Situation: Ist eine Originalstudie besonders unpräzise, und reicht ihr Konfidenzintervall bis nahe an Null, dann wird es besonders schwierig eine Replikation zu finden, die der Originalstudie widerspricht, auch wenn der tatsächliche Effekt bei etwa Null liegt. In diesem Fall liefern nämlich etwa die Hälfte der Replikationen Effekte, die sich innerhalb dieses Konfidenzintervalls bewegen, auch wenn dieser Effekt nicht existiert. Diese Vorgehensweise belohnt quasi unzuverlässige bzw. ungenaue Schätzungen der Originalstudien.

Diese Überlegungen zeigen vor allem, dass es nicht trivial ist den Erfolg einer Replikation zu bewerten. Je nachdem für welche Methode man sich entscheidet, erhält man unterschiedliche Schätzungen der Replizierbarkeitsquote. Die Ergebnisse können je nachdem wie sie dargestellt werden, bzw. statistisch „geframed“ werden, zu unterschiedlichen Schlussfolgerungen führen. Das liegt auch daran, dass viele der Ergebnisse des RP:P nicht eindeutig sind. Die Ergebnisse selbst ändern sich dadurch aber nicht, nur das Bild welches erzeugt wird.

Der zentrale Wert des RP:P liegt aber sicherlich darin, die Diskussion um problematische Forschungspraktiken und nicht zufriedenstellende Replizierbarkeit psychologischer Ergebnisse an die breite Masse getragen zu haben. Das Projekt verhalf die Thematik ins Zentrum des psychologischen Interesses zu rücken. Die Aufmerksamkeit, die das Projekt erfahren hat, ist etwa am Altmetric-Score ablesbar, der im Oktober 2016 bei 2970 liegt (Altmetric, 2016). Altmetric reiht das Paper des RP:P im Vergleich mit Publikationen ähnlichen Alters im Oktober 2016 auf Platz 3, und im Vergleich mit sämtlichen Publikationen auf Platz 46. Das breite Interesse der Psychologie, auch im deutschsprachigen Raum, ist aber auch daran erkennbar, wie gut besucht der Vortrag von Brian Nosek, dem Initiator des RP:P, am diesjährigen Kongress der Deutschen Gesellschaft für Psychologie war.

Zu diesem Erfolg verhalf sicherlich auch die Tatsache, dass vor allem jener Anteil an Replikationen berichtet wurde, die wieder statistisch signifikant wurden. Dieser Wert ist besonders plakativ, da so der Anteil replizierter Studien besonders gering erscheint. Signifikanz stellt aber nach wie vor, die Bedingung einer Publikation dar, und ist für viele Psychologen der Kennwert der Wahl, wenn es um die Bewertung von Studien geht. Deshalb ist es womöglich für Psychologen am einfachsten und intuitivsten verständlich, dass es ein

Problem in der Art und Weise gibt, wie Daten erhoben, ausgewertet und interpretiert werden, wenn nur 36% der publizierten Studien wieder zu einem signifikanten Effekt führen. Der größte Nutzen des RP:P kann darin gesehen werden, die theoretischen Überlegungen, die zum Vertrauensverlust geführt haben, einer breiteren Masse greifbarer gemacht zu haben, und eine empirische Evidenz dafür gefunden zu haben.

Bezieht man nun die zuvor geschilderten Probleme der niedrigen Teststärke psychologischer Forschung, der überschätzter Effektstärken, des Publikationsbias oder auch der QRPs in die Überlegungen mit ein, ist das Ergebnis des RP:P weit weniger überraschend. Man kann natürlich kritisieren, dass die Berechnung der Power der Replikationen nicht adäquat erfolgte. Man weiß theoretisch, dass durch die niedrige Teststärke der ursprünglichen Studien und das Signifikanzniveau, die Power einer Replikation überschätzt wird. Aber genau diese Probleme thematisiert das RP:P. Es geht nicht primär darum die Originalergebnisse zu überprüfen. Es war keinesfalls das Ziel so viele Studien wie möglich zu bestätigen bzw. zu diskreditieren. Das ist mit einzelnen Replikationen ebenso wenig möglich, wie eine einzige Studie einen Effekt beweisen kann. Es gibt andere Replikationsprojekte, die den Fokus verstärkt darauflegen. Es geht vielmehr darum, die theoretischen Überlegungen auch empirisch darzulegen. Wenn also Methoden Anwendung fänden, deren Ziel es ist, beispielsweise auf die niedrige Power der Originalstudien Bezug zu nehmen, und diese zu korrigieren (etwa Safeguard Power[7]), dann wird es schwierig die Auswirkungen eben genau dieser Probleme (niedrige Teststärke) zu zeigen. Powerüberlegungen stellen in Originalstudien eine Seltenheit dar, das RP:P führte immerhin klassische Poweranalysen durch. Diese Vorgehensweise ist durchaus angemessen, wenn es, wie hier, um die Bewertung der Methodik der Originalstudien geht. Hätten die Autoren den Mehraufwand auf sich genommen, und noch größere Stichprobengrößen angestrebt, wären womöglich ein paar Replikationen mehr wieder signifikant, oder zumindest wären viele der Replikationen klarer einer Seite (repliziert vs. nicht repliziert) zuordenbar. Dann wäre aber auch etwas an der Kritik des RP:P verloren gegangen, und diese wäre nicht so einfach, so unmittelbar erfahrbar gewesen. Die Forderung nach adäquater Power, um informative Ergebnisse zu erhalten, muss bereits bei den Originalstudien ansetzten, und nicht erst bei

[7] Die Idee der Safeguard Power (Perugini et al., 2014) wurde allerdings erst 2014 publiziert. Zu diesem Zeitpunkt war das Projekt bereits weit vorangeschritten.

den Replikationen. Die Unsicherheit, der die Ergebnisse der Replikationen ausgesetzt sind, ergibt sich in diesem Fall erst aus der Unsicherheit der ursprünglichen Ergebnisse.

Außerdem kann man die Tatsache, dass das RPP Replikationen auch am *p*-Wert bewertet hat, durchaus auch als Kritik an jenem selbst lesen. Diese zeigt, dass eine Bewertung allein am *p*-Wert nicht ausreichend Informationen einbezieht um eine adäquate Einschätzung zu ermöglichen. Die Überlegung ist eine ähnliche, wie bei der Korrektur der Poweranalyse. Wenn man die Probleme normativer Forschungspraktiken zeigen möchte, gelingt dies am einfachsten, wenn man durch die Anwendung eben jener selbst, die vorhandenen Missstände demonstriert. Die Power der Replikationen mag problematisch sein, die Auswertung jener am *p*-Wert auch, aber diese Punkte waren bei den Originalstudien nicht weniger fragwürdig. Man kann sagen, dass in dieser scheinbaren Schwäche des RP:P auch eben seine Stärke liegt.

Die Vorbildwirkung des RP:P setzt dabei an anderer Stelle an. Sämtliche Informationen, Daten, Materialien und Hintergrundinformationen werden frei zur Verfügung gestellt. Zusätzliche Analysen durchzuführen ist somit jedem und jeder möglich. Alle durchgeführten Replikationen wurden präregistriert und unabhängig vom Ergebnis veröffentlicht. Das RP:P zeigt hier einen Weg vor, der zukunftsweisend sein könnte, und zwar nicht nur für Replikationsstudien, sondern viel mehr noch für Originalstudien. Solche Bestrebungen die Wissenschaft zu öffnen werden unter dem Begriff *Open Science* zusammengefasst (vgl. Kap. 10).

Insgesamt zeigt das RP:P jedenfalls, dass Replizierbarkeit psychologischer Forschung nicht zufriedenstellend ist, und auch nicht viel Vertrauen in die Ergebnisse einzelner experimenteller Studien gelegt werden sollte. Replikationen sind ein möglicher Weg dieses Vertrauen zu stärken.

9. Direkte Replikation von van Dijk et al. (2008) im Rahmen des Reproducibilty Project: Psychology

Die folgende Darstellung der direkten Replikation von van Dijk, van Kleef, Steinel und van Beest (2008) orientiert sich zweckmäßigerweise in Teilen an dem im Zuge des RP:P veröffentlichten Replication Report (Slowik & Voracek, 2015). Dieser wurde, wie im Rahmen des RP:P vorgesehen, vor der Datensammlung frei zur Verfügung gestellt, und

wurde anschließend um Ergebnisse und Diskussion ergänzt. D.h. die vorliegende Replikation wurde präregistiert; ebenfalls sind die Daten und Analyseskripte der Replikation archiviert (https://osf.io/xtsq6). Allerdings werden die verwendeten Materialien, auf Bitte des Originalautors hin, nicht zur Verfügung gestellt.

9.1. Theoretischer Hintergrund

Die Studie von (van Dijk et al., 2008) untersucht den Einfluss von Emotionen im Rahmen des Ultimatumspiels. Das *Ultimatumspiel* ist eine klassische Versuchsanordnung um das Verhalten von Personen in Entscheidungssituationen zu untersuchen. Eine Person X hat die Aufgabe ein Gut (z.B. einen bestimmten Geldbetrag) zwischen sich selbst und einem Mitspieler oder einer Mitspielerin Y aufzuteilen. Der Mitspieler oder die Mitspielerin Y hat daraufhin die Möglichkeit das Angebot anzunehmen oder abzulehnen. Nimmt Y das Angebot an, dann erhalten die beiden Personen die Geldbeträge entsprechend der von X vorgeschlagenen Aufteilung. Wenn der Mitspieler oder die Mitspielerin Y das Angebot ablehnt, bekommen beide Personen nichts (Güth, Schmittberger, & Schwarze, 1982). Eine Variation des Ultimatumspiels bietet das *Deltaspiel* (Suleiman, 1996). Im Falle einer Ablehnung des Angebots durch Spieler Y, wird die von Spieler X vorgeschlagene Aufteilung mit einem Faktor Delta ($0 \leq \delta \leq 1$) multipliziert. Ist Delta gleich Null ($\delta = 0$), entspricht die Variation dem klassischen Ultimatumspiel.

Van Dijk et al. (2008) möchten die Rolle von Emotionen im Kontext des Ultimatumspiels untersuchen. Zunächst gehen sie davon aus, dass Emotionen eine soziale Funktion erfüllen, dass sie bei sozialen Interaktionen Informationen über das Gegenüber liefern. Van Kleef, De Dreu und Manstead (2004) zeigten, dass in Verhandlungssituationen das Limit von Spielern und Spielerinnen, die sich ärgerten, als höher wahrgenommen wurde, als jenes von Spielern und Spielerinnen, die sich freuten. Personen, die sich ärgerten, wurden höhere Zugeständnisse gemacht. Ärger zu vermitteln, scheint sich demzufolge in Verhandlungssituationen auszuzahlen.

Van Dijk et al. (2008) vermuten aber, dass dies differenzierter betrachtet werden muss. Ärger zahle sich für den einen Spieler nur dann aus, wenn es für den anderen Spieler keinen anderen Ausweg gebe, als nachzugeben. In Experiment 3, welches im Rahmen des RP:P repliziert wurde, variieren die Autoren die Verteilung der Macht zwischen zwei

Spielern eines Deltaspiels, bei dem 100 Münzen zwischen zwei Spielern aufgeteilt werden sollen. Unter der Macht eines Spielers oder einer Spielerin verstehen sie den Einfluss, den diese auf den gegenseitigen Ausgang des Spiels haben. Die Variation der Macht wird erreicht über die Variation von Delta. Ist Delta hoch, dann sind die Konsequenzen einer Ablehnung des Angebots durch Y weniger drastisch für X und somit verschiebt sich die Machtverteilung zugunsten von X.

Zusätzlich zur Variation von Delta ($\delta = 0$ vs. $\delta = .9$) bzw. der Machtverteilung zwischen den Spielern, wurde auch die Emotion des Gegenspielers (Ärger vs. Freude) manipuliert. Sind die Konsequenzen niedrig, wird einem Gegenspieler der sich ärgert, ein geringeres Angebot gemacht, als einem Gegenspieler, der sich freut – so die Hypothese von van Dijk.

Tatsächlich zeigte das Ergebnis von Dijk et al. (2008) den vermuteten Effekt: Die Teilnehmer und Teilnehmerinnen boten Gegenspielern bzw. Gegenspielerinnen, die Ärger kommunizierten, durchschnittlich 12.6 Münzen weniger an, als sich freuenden Mitspielern und Mitspielerinnen, wenn die Konsequenzen einer Ablehnung des Angebots niedrig waren $F(1,99) = 16.62$, $p < .0001$. Dieser Effekt wurde als Schlüsseleffekt im Rahmen des RP:P betrachtet und wurde sowohl zur Berechnung der Stichprobe, als auch zur Evaluation der direkten Replikation herangezogen.

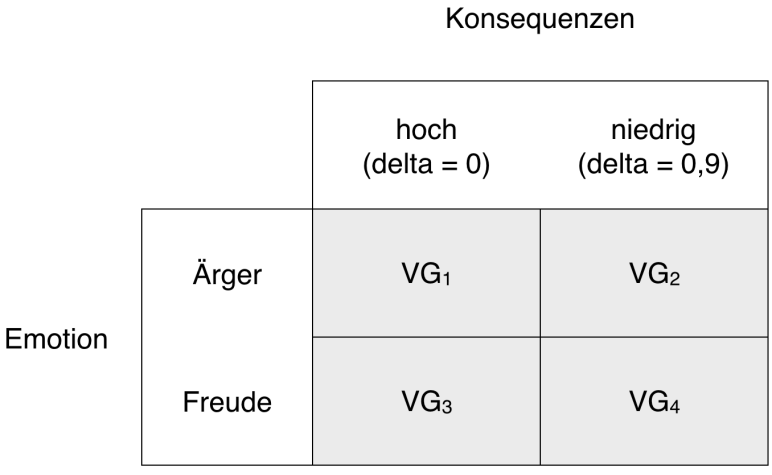

Abbildung 1: Design von van Dijk et al. (2008) bzw. der Replikation (2015): Als unabhängige Variablen dienten zum einen die Konsequenzen eines abgelehnten Angebots, zum anderen die Emotion des Gegenspielers bzw. der Gegenspielerin.

9.1.Methoden

Design

Van Dijk et al. entsprechend wurde ein 2 (Emotion von Y: Ärger vs. Freude) x 2 (δ= 0 vs. δ = .09) Between-Subject Design herangezogen (Abb. 1).

Stichprobe

Die Stichprobe der Replikation bestand aus 83 Studierenden der Universität Wien (27 Männer und 56 Frauen) im Alter von 18 bis 55 Jahren (M = 24.5, SD = 7.35). Die Versuchspersonen wurden über das LABS (Laboratory Administration for Behavioral Sciences) der Fakultät für Psychologie zur Teilnahme an der Studie eingeladen. Mit dem Ziel erfahrene Teilnehmer und Teilnehmerinnen, die sich der Möglichkeit einer in der Studie implementierten Täuschung bewusst sind, von der Teilnahme auszuschließen, wurden explizit nur Studienanfänger und Studienanfängerinnen angesprochen. Der Großteil der Versuchspersonen waren Psychologiestudierende im ersten Semester.

In der Einladungsmail wurde darauf hingewiesen, dass eine Teilnahme an der Studie bezahlt werde; ein genauer Betrag wurde aber nicht genannt. Außerdem wurde betont, dass die Durchführung der Studie nur möglich sei, wenn vier Personen gleichzeitig teilnehmen. Dabei handelt es sich um die minimale Anzahl an Versuchspersonen, die sich während der Studie im Raum befinden müssen, um diese schlüssig und damit auch glaubwürdig wirken zu lassen. Der Versuchsaufbau vermittelt den Anschein jeweils zwei Teilnehmer bzw. Teilnehmerinnen seien miteinander verbunden. Die Versuchsperson erfährt aber nicht, wer ihr vermeintlicher Gegenspieler oder ihre vermeintliche Gegenspielerin ist.

Materialien

Eine englischsprachige Übersetzung der im Original niederländischen Instruktionen wurde von Eric van Dijk zur Verfügung gestellt. Die experimentelle Manipulation erfolgte über Open Sesame 2.9 (Mathôt, Schreij, & Theeuwes, 2012).

Prozedur

Die in der Originalstudie beschriebene Prozedur wurde beibehalten; fehlende Details konnten über den Kontakt mit den ursprünglichen Autoren in Erfahrung gebracht. Die Versuchsreihen fanden in den Experimentalräumlichkeiten der Fakultät für Psychologie der Universität Wien statt. Zunächst erhielten die Versuchspersonen die Information, dass sie an einer Studie zum Verhalten in Verhandlungssituationen teilnehmen und, dass sie gerade zufällig mit einer der weiteren Person im Raum verbunden werden. Allen Teilnehmern und Teilnehmerinnen wurde mitgeteilt, dass ihnen die Rolle X zugewiesen werde, während die mit Ihnen verbundene Person die Rolle Y erhalte. Es folgten sechs allgemeine Aussagen zu Verhandlungssituationen, wie etwa: „Strategie spielt eine wichtige Rolle bei Verhandlungen". Das Ausmaß der Zustimmung musste jeweils angegeben werden. Erst im Anschluss daran wurde den Teilnehmern und Teilnehmerinnen mitgeteilt, dass ihre Antworten nun ihrem Gegenüber Y zugeschickt werden. Danach wurde den Teilnehmern und Teilnehmerinnen ihre Rolle (100 Münzen im Wert von € 0.08 aufzuteilen, bzw. ein Angebot an Y zu machen), und die Rolle ihres Gegenübers Y (das Angebot anzunehmen oder abzulehnen) erklärt. Sollte Y das Angebot annehmen, werden die Münzen entsprechend verteilt. In den Versuchsgruppen, in denen die Konsequenzen hoch (bzw. niedrig) waren, wurden die Personen informiert, dass im Falle einer Ablehnung des Angebots, dieses mit 0 (bzw. 0.9) multipliziert werde, bevor es verteilt werde. D.h. in den Versuchsgruppen mit hohen Konsequenzen, würden die Teilnehmer ohne Remuneration aussteigen, sollte Y ihr Angebot ablehnen. In den Versuchsgruppen mit niedrigen Konsequenzen, würden sowohl X und Y mit 90% des Angebots aussteigen, sollte Y dieses ablehnen.

Nachdem die Rollen von X und Y geklärt waren, erhielten die Versuchspersonen die angebliche Reaktion des Gegenspielers Y („Ich ärgere mich." vs. „Ich freue mich."), auf die zuvor an Y weitergeleiteten Antworten des Fragebogens präsentiert. Dabei wurde betont, dass Y nicht wisse, dass X diese Information erhalte.

Anschließend wurden die Versuchspersonen aufgefordert Y ein Angebot zu machen. Zusätzlich wurden Manipulation Checks implementiert und die Versuchspersonen sollten das Ausmaß ihrer Emotionen und jenes des Gegenspielers (Ärger und Freude) beurteilen. Außerdem wurde gefragt, wie hoch das Angebot wohl sein müsste, damit Y dieses annehme, und wie wahrscheinlich es sei, dass Y das tatsächlich gemachte Angebot

annehme. Am Ende der Studie wurden die Teilnehmer über den Zweck der Studie aufgeklärt und mit € 8.00 entlohnt.

Unterschiede zur Originalstudie

Die Stichprobe der Originalstudie stammt aus den Niederlanden, die Stichprobe der Replikation hingegen aus Österreich.

In Absprache mit den Originalautoren wurden minimale Anpassungen der Instruktion vorgenommen, die als Adaption von van Dijk et al. (2008) in das Jahr 2015 bzw. als Modernisierung angesehen werden können. In der ursprünglichen Fassung wurde eine Sequenz präsentiert, die das Ziel hatte, die Teilnehmer davon zu überzeugen, sie seien tatsächlich mit einem anderen Teilnehmer im Raum verbunden. Es wurde die Suche nach einer Verbindung zu einem Server vorgetäuscht. Diese Sequenz wurde in der Replikation gestrichen, da die Rechengeschwindigkeit in den letzten Jahren derart zugenommen hat, dass eine längere Ladezeit nicht mehr realistisch erscheint. Außerdem wurden in der Originalstudie die Versuchspersonen aufgefordert das Angebot an ihren Gegenspieler nicht nur einzutippen, sondern zuvor auch noch auf ein Blatt Papier zu schreiben, welches dann angeblich der mit ihnen verbunden Person Y überreicht wurde.

Die Teilnehmer und Teilnehmerinnen der Replikation waren im Durchschnitt älter (M_{Alter} = 24.5 Jahre, SD_{Alter} = 7.35) als jene der Originalstudie (M_{Alter} = 20.8 Jahre). Zwei Punkte scheinen dazu beizutragen: Ersten Studienanfänger im Fach Psychologie sind an der Universität Wien möglicherweise bereits älter, als jene an anderen Universitäten. Im Studienjahr 2004/05 jedenfalls waren die Studierenden im ersten Semester im Durchschnitt bereits M_{Alter} = 22.9 (SD_{Alter} = 5.9) Jahre alt (Voracek et al., 2010). Zweitens wird der Mittelwert in der Stichprobe der Replikation durch Ausreißer verzerrt, da vier Personen, die an der Replikation teilnahmen, älter als 40 Jahre waren. Der Median beträgt 22 Jahre, der Modalwert 19 Jahre. Somit wird der geringe Altersunterschied als vernachlässigbar angesehen.

9.2. Ergebnisse

Anzahl der angebotenen Münzen im Ultimatumspiel

Eine 2 (Emotion von Y) x 2 (Konsequenzen eines abgelehnten Angebots) ANOVA der Anzahl der angebotenen Münzen ergab einen statistisch signifikanten Haupteffekt der Konsequenzen einer Ablehnung $F(1, 79) = 12.92$, $p < .01$, $\eta_p^2 = .14$. Die Teilnehmer und Teilnehmerinnen boten ihrem Gegenüber mehr Münzen in der Bedingung mit hohen Konsequenzen ($M = 46.95$, $SD = 6.50$) an, als in der Bedingung mit niedrigen Konsequenzen ($M = 36.90$, $SD = 16.84$).

Ein Levene Test deutete auf heterogene Varianzen der Höhe des Angebots in den vier Versuchsbedingungen hin $F(3, 79) = 10.32$, $p < .001$. Anstatt eines F-Tests wurde ein t-Test durchgeführt, um die beiden interessierenden Zellen (Ärger vs. Freude von Y in der Bedingung mit niedrigen Konsequenzen einer Ablehnung) zu vergleichen. Hier zeigte der Levene Test keine heterogenen Varianzen an $F(1, 38) < .01$, $p > .99$. Die Anzahl der angebotenen Münzen unterschied sich in den beiden Bedingungen nicht $t(38) = -0.26$, $p = .80$, $d = -0.08$ (Abb. 2).

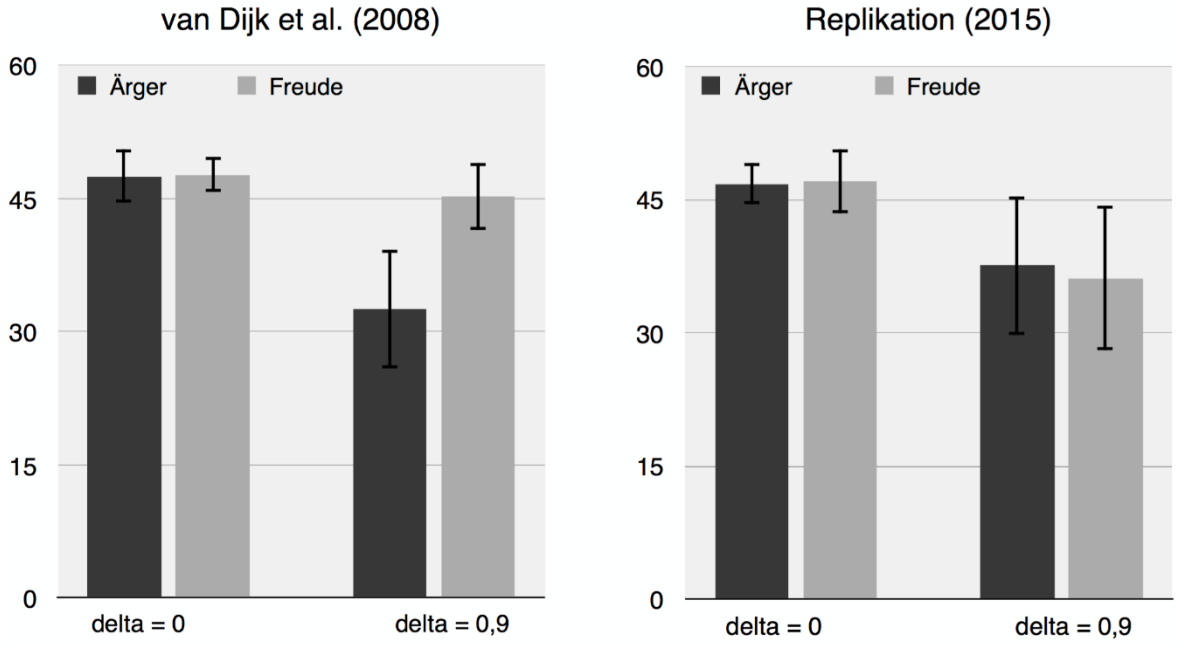

Abbildung 2: Anzahl der gebotenen Münzen durch die Teilnehmer in den vier Versuchsbedingungen bei van Dijk et al (2008) und der Replikation. Die Fehlerbalken kennzeichnen jeweils das 95%-KI.

Zusätzlich unterschied sich in der Bedingung mit hohen Konsequenzen die Anzahl der Münzen, die einem sich ärgerndem Y (M = 46.82, SD = 5.26) angeboten wurde nicht von der Anzahl der Münzen, die einem sich freuenden Y (M = 46.82, SD = 5.26) angeboten wurde, $t(41) = 0.14$, $p = .89$, $d = 0.04$. Ein Levene Test (Ärger vs. Freude von Y in der Bedingung mit hohen Konsequenzen) zeigte keine heterogenen Varianzen an, $F(1, 41) = 0.41$, $p = .52$.

Manipulationschecks

Eine 2 (Emotion von Y) x 2 (Konsequenzen eines abgelehnten Angebots) ANOVA des wahrgenommenen Ärgers des Gegenspielers bzw. der Gegenspielerin zeigte einen statistisch signifikanten Haupteffekt für die Emotion von Y, $F(1, 79) = 84.99$, $p < .001$, $\eta_p^2 = .52$. Das deutet darauf hin, dass die Personen in der Bedingung mit sich ärgerndem Gegenüber, ihr Gegenüber als verärgerter wahrnahmen (M = 3.69, SD = 1.05), als Personen in der Bedingung mich sich freuendem Gegenüber (M = 1.56, SD = 1.05). Auch eine 2 x 2 ANOVA der wahrgenommen Freude des Gegenspielers bzw. der Gegenspielerin zeigte nur einen statistisch signifikanten Effekt für die Emotion von Y, $F(1, 79) = 172.36$, $p < .001$, $\eta_p^2 = .69$. Personen in der Bedingung mit sich ärgerndem Gegenüber schätzten das Ausmaß der Freude ihres Gegenüber als geringer ein (M = 2.90 , SD = 0.96) als das die Personen in der Bedingung mit einem sich freuendem Gegenüber taten (M = 5.32, SD = 0.76).

Die Teilnehmer und Teilnehmerinnen wurden nach den Folgen einer Ablehnung ihres Angebots durch den Gegenspieler bzw. die Gegenspielerin gefragt. Alle Personen haben diese Frage den zuvor präsentierten Instruktionen entsprechend beantwortet.

Außerdem wurde nach dem relativen Einfluss gefragt, den X bzw. Y auf die mögliche Verteilung der Münzen haben. Eine 2 x 2 ANOVA ergab nur einen Haupteffekt der Konsequenzen der Ablehnung $F(1, 79) = 21.14$, $p < .001$, $\eta_p^2 = .21$. Personen in der Bedingung mit hohen Konsequenzen (M = 3.70, SD = 1.68) nahmen den relative Einfluss von Y als höher wahr, als Personen in der Bedingung mit niedrigen Konsequenzen (M = 2.20, SD = 1.27). Nach Dijk et al. (2008) zeigen diese Ergebnisse, dass die Manipulation über die Emotion (Ärger vs. Freude) von Y bzw. die Konsequenzen einer Ablehnung ($\delta = 0$ vs. $\delta = 0.9$) erfolgreich war.

Wahrscheinlichkeit, dass der Gegenspieler das Angebot annehmen wird

Teilnehmer und Teilnehmerinnen wurden gefragt für wie wahrscheinlich sie es halten, dass ihr Gegenspieler bzw. ihre Gegenspielerin das Angebot annehmen werde. Eine 2 x 2 ANOVA zeigte einen statistisch signifikanten Haupteffekt für die Konsequenzen einer Ablehnung $F(1, 79) = 19.75$, $p < .001$, $\eta_p^2 = .20$. Die Personen in der Bedingung mit niedrigen Konsequenzen einer Ablehnung ($M = 4.25$, $SD = 2.17$) glaubten, verglichen mit den Personen in der Bedingung mit hohen Konsequenzen ($M = 5.95$, $SD = 1.27$), es sei weniger wahrscheinlich, dass Y das Angebot annehmen werde. Es wurde kein Haupteffekt der Emotion von Y gefunden, $F(1, 79) = 2.38$, $p = .127$, $\eta_p^2 = .03$. Ebenso zeigte sich keine Interaktion $F(1, 79) = 1.13$, $p = .29$, $\eta_p^2 = .01$.

Bedeutung der Möglichkeit einer Ablehnung des Angebots

Die Teilnehmer und Teilnehmerinnen sollten angeben, inwieweit die Möglichkeit von Y das Angebot abzulehnen, ihre Entscheidung beeinflusst hat. Eine 2 x 2 ANOVA zeigte einen statistisch signifikanten Haupteffekt nur für den Faktor der Konsequenzen der Ablehnung $F(1, 79) = 14.74$, $p < .001$, $\eta_p^2 = .16$. Die Tatsache, dass das Gegenüber ablehnen konnte, war also wichtiger, wenn die Konsequenzen einer möglichen Ablehnung hoch waren ($M = 4.91$, $SD = 1.72$), als wenn diese niedrig waren ($M = 3.35$, $SD = 1.94$).

Wahrgenommenes Limit

Die Teilnehmer und Teilnehmerinnen sollten einschätzen, welches Angebot Y zumindest bekommen müsste, um dieses auch anzunehmen. Dieses Maß wurde als wahrgenommenes Limit des Gegenspielers interpretiert. Eine 2 x 2 ANOVA zeigte den erwarteten Haupteffekt des Faktors Emotion von Y nicht $F(1, 79) = 0.06$, $p = .81$, $\eta_p^2 < .01$. Die Einschätzung eines verärgerten Gegenübers ($M = 42.40$, $SD = 13.17$) unterschied sich nicht von der Einschätzung eines sich freuenden Gegenübers ($M = 43.17$, $SD = 10.41$).

Emotion der Studienteilnehmer und Studienteilnehmerinnen

Schließlich wurden die Teilnehmer und Teilnehmerinnen gefragt, ob sie sich geärgert oder gefreut haben. Eine 2 x 2 ANOVA der Ärgerratings ergab einen Haupteffekt nur für die Emotion der Empfänger $F(1, 79) = 77.77$, $p < .001$, $\eta_p^2 = .47$. Personen in der

Bedingung mit verärgerten Y gaben an sich in einem stärkeren Ausmaß zu ärgern, (M = 3.00, SD = 1.15) als die Personen, die in der Bedingung mich sich freuendem Y waren (M = 1.24, SD = 0.58).

Entsprechend zeigte eine 2 x 2 ANOVA der Bewertung der ihrer Freude durch die Versuchspersonen, einen Haupteffekt des Faktors der Emotion von Y $F(1, 79) = 97.90$, $p < .001$, $\eta_p^2 = .55$. Personen, welche sich in der Bedingung mit verärgertem Y befanden gaben in geringerem Ausmaß an sich zu freuen (M = 2.31, SD = 1.05) als die Personen, die sich in der Bedingung mit sich freuendem Gegenüber taten (M = 4.37, SD = 0.86). Ein Haupteffekt des Faktors der Konsequenzen einer Ablehnung auf die Emotion der teilnehmenden Personen wurde nicht gefunden $F(1, 79) = 3.89$, $p = .052$, $\eta_p^2 = .05$.

Zusammenfassung

Es wurden keine Hinweise gefunden, dass die Teilnehmer und Teilnehmerinnen ihren Gegenspielern und Gegenspielerinnen in der Bedingung mit niedrigen Konsequenzen ein geringeres Angebot machten, sofern diese sich ärgerten. Das Experiment 3 der Studie von van Dijk et al. (2008) konnte somit nicht repliziert werden.

Ebenso konnte kein Effekt der Emotion auf die Höhe des Angebotes in der Bedingung der hohen Konsequenzen gefunden werden. Allerdings fand van Dijk et al. (2008) diesbezüglich auch keine Unterschiede. Ebenso wurden in der Replikation ähnliche Varianzen gefunden: die Angebote in der Bedingung der hohen Konsequenzen variierten wie bei van Dijk et al. (2008) in einem geringeren Ausmaß, als das bei den Angeboten in der Bedingung mit niedrigen Konsequenzen der Fall war.

Außerdem konnten eine Reihe von Effekten, die allerdings nicht zentral für die Hypothesen der Originalautoren waren, bestätigt werden. Ähnliche Effekte wurden bei den Manipulations-Checks gefunden; zudem spielte die Möglichkeit, dass der Gegenspieler bzw. die Gegenspielerin das Angebot ablehnen konnte, je nachdem ob die Konsequenzen hoch oder niedrig waren, eine unterschiedliche Rolle. Ebenso wurden die Effekte auf die Emotion der der Teilnehmer und Teilnehmerinnen gefunden werden, auch wenn diese in der Replikation schwächer als in der Originalstudie ausfielen.

Der Haupteffekt der Emotion auf das wahrgenommene Limit wurde nicht gefunden. Angesichts der Höhe des Effekts bei van Dijk et al. ($\eta_p^2 = .045$), war die Teststärke der Replikation um diesen Effekt festzustellen, mit einer Stichprobengröße (N = 83)

allerdings niedrig (50%).

9.3. Evaluation des Schlüsseleffekts der Replikation

Im Folgenden wird das Ergebnis des Schlüsseleffekts der Replikation hinsichtlich weiterer Möglichkeiten bewertet. Es soll auch die zuvor geschilderte Methode der Detectability bzw. der Small Telescopes demonstriert werden, und jene Stichprobengröße berechnet werden, die entsprechend der Safeguard Power für die Replikation nötig gewesen wäre.

Signifikanz

Wie bereits dargelegt, zeigte die Replikation kein signifikantes Ergebnis bezüglich des Schlüsseleffekts des Experiments 3 von van Dijk et al. (2008). Die interessierenden Bedingungen unterschieden sich nicht statistisch signifikant $t(38) = -0.26$, $p = .80$, $d = -0.08$.

Effektgröße der Replikation im Vergleich zur Effektgröße der Originalstudie

Das Konfidenzintervall der Replikation (95% KI = [-0.70, 0.54]) umfasst den Effekt der Originalstudie nicht. Ebenso liegt der Effekt der Replikation außerhalb des Konfidenzintervalls der Originalstudie (95%-KI = [0.36, 1.53]) (vgl. Abb. 3).

Meta-Analyse

Nach der meta-analytischen Kombination der Replikation mit der Originalstudie auf Grundlage eines Fixed-Effect-Modell (in Anlehnung an die Auswertungsstrategien des RP:P) bleibt der ursprüngliche Effekt weiterhin statistisch signifikant (Abb. 3). Der meta-analytische Schätzer beläuft sich dann auf $d = 0.46$ (95%-KI = [0.04, 0.89], $p = .03$). Aber auch direkte Replikationen sind nie vollkommen ident mit der Originalstudie (z.B. bezogen auf die Stichprobe) und somit ist es sinnvoll von einem Random-Effects-Modell auszugehen. In diesem Fall erhält man einen meta-analytischen Schätzer von $d = 0.44$ (95%-KI = [-0.57, 1.44], $p = .39$); so ist der Effekt nicht mehr statistisch signifikant.

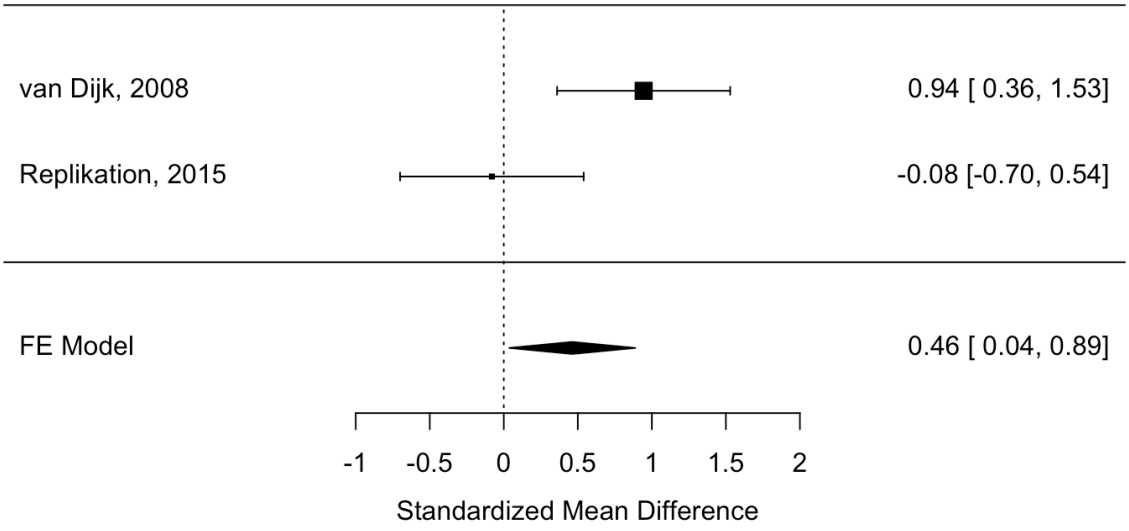

Abbildung 3. Resultat der Meta-Analyse der ursprünglichen Studie von van Dijk et al. (2008) und der Replikation. Der Effekt bleibt statistisch signifikant (p = .03) unter Annahme eines Fixed-Effect-Modells.

Small Telescopes

Nach Simohnsohn (2015) kann das Ergebnis einer Replikation danach bewertet werden, ob der in der Replikation gefundene Effekt signifikant verschieden von $d_{33\%}$ ist (vgl. Kap. 7). So kann gezeigt werden, dass die Originalstudie nicht in der Lage gewesen ist, diesen Effekt bedeutsam zu untersuchen. Simonsohn (2015) empfiehlt die ursprüngliche Stichprobe mit 2.5 zu multiplizieren, um für diesen Test eine Power von 80% zu erhalten; tatsächlich hatte die vorliegende Replikation eine Stichprobe die kleiner war als die ursprüngliche Stichprobe, und eine Power von nur 38% um zu zeigen, dass der Effekt statistisch signifikant verschieden von $d_{33\%}$ ist. Es handelt sich um ein noch kleineres „Teleskop“ als das welches in der Originalstudie Verwendung fand. Dennoch soll das Verfahren hier demonstriert werden.

Die Effektstärke, für welche die Originalstudie eine Teststärke von 33% aufwies, beläuft sich auf $d_{33\%}$ = 0.43. Die Replikation zeigte einen Effekt von d = -0.08. Das 90%-Konfidenzintervall für den Effekt der Replikation umfasst Werte zwischen d = -0.603 und d = 0.438. Somit umfasst dieses Konfidenzintervall $d_{33\%}$ und die Replikationsstudie kann nicht zeigen, dass die ursprüngliche Studie zu klein war, um diesen Effekt zu untersuchen.

Safeguard Power

Entsprechend dem Konzept der Safeguard Power (Perugini et al., 2014), wäre die Stichprobe der Replikation bedeutend größer gewesen. Die Originalstudie konnte einen Effekt von $d_0 = 0.94$ zeigen. Die untere Grenze des 60%-Konfidenzintervalls beläuft sich auf $d_s = 0.71$. Auf Grundlage dieser Grenze berechnet, benötigt man bei einem Zweigruppenvergleich für 80% Power Zellen mit einer Stichprobengröße von $N_s = 33$, anstatt der $N = 20$ pro Zelle der Replikation. Insgesamt wäre also eine Stichprobengröße von $N = 132$ Personen anzustreben gewesen. Für 95% Power auf Grundlage von $d_s = 0.71$ wäre eine Zellengröße von 53 Personen bzw. insgesamt eine Stichprobengröße von $N = 212$ benötigt worden.

Zusammenfassung

Insgesamt sprechen die in der Replikation gefundenen Daten nicht für den von van Dijk et al. (2008) gefundenen Effekt. Es kann letztlich aber nicht ausgeschlossen, dass unbekannte Moderatoren, oder die geringfügigen Unterschiede zum Verschwinden des Effekts beigetragen haben. Zugleich ist im Zusammenhang mit diesen Möglichkeiten aber auch festzuhalten, dass die Originalstudie solche Effektmoderatoren weder testete, noch thematisierte. Die Sensitivität eines Effekts gegenüber nur geringfügigen Implementierungsvarianten würde jedenfalls gegen die Robustheit und Generalisierbarkeit („Bandbreite“) eines solchen Effektes sprechen.

10. Mögliche Lösungsansätze – hin zu reliablen Effekten

Neben aus den zuvor geschilderten Problemen, direkt ableitbaren Verbesserungsmöglichkeiten, wie etwa höherer Teststärke bzw. größerer Stichproben, der Publikation von sämtlichen untersuchten Effekte oder der Durchführung von Replikationen, werden vor allem folgende Punkte diskutiert:

10.1. Effektstärken und Konfidenzintervalle

Cumming (2013) propagiert das Berichten von Effektstärken und Konfidenzintervallen, anstatt der Anwendung von Nullhypothesentestens.

Konfidenzintervalle können die Variabilität vorgefundener Daten verdeutlichen. Vielen Psychologen ist wahrscheinlich nicht bewusst mit welcher Unsicherheit die von ihnen erzielten Ergebnisse verbunden sind. Möglicherweise kann eine explizite Veranschaulichung dieser Unsicherheit auch dazu verhelfen, diese Problematik mehr ins Bewusstsein der Agierenden zu rücken, so die Power psychologischer Ergebnisse erhöhen, und damit die Unsicherheit begrenzen bzw. Konfidenzintervalle schmälern. Allerdings sind diese Empfehlungen bereits im aktuellen APA-Manual (American Psychological Association, 2010) verankert.

Wird primär das Ziel Effekte präzise zu schätzen verfolgt; d.h. möglichst schmale Konfidenzintervalle zu erhalten, und nicht nur signifikante Ergebnisse, also Konfidenzintervalle, welche die Null nicht umfassen, ist die klassische Poweranalyse nur bedingt zur Bestimmung der Stichprobengröße geeignet. Ein Ansatz der hier als Alternative dienen kann, ist die *Accuracy in Parameter Estimation* (AIPE; Maxwell, Kelley, & Rausch, 2008), die darauf abzielt die Weite eines Konfidenzintervalls zu begrenzen.

10.2. Präregistrierung

Direkte Replikationen sind deswegen so gut geeignet, um falsch positive Ergebnisse zu identifizieren, weil sie die Entscheidungen, die in der Originalstudie getroffen wurden festhalten, und somit jegliche Flexibilität im Zuge der Durchführung der Studie unterbunden wird. Präregistrierung meint, dass bereits im Vorhinein genau festgelegt wird, wie in der Studie vorgegangen wird. Es werden also Design und Analyse der Studie bereits dargelegt bevor Daten gesammelt und ausgewertet werden. Im Prinzip entspricht somit eine direkte Replikation einer Präregistrierung der Originalstudie im Nachhinein, da beide die gegebene Flexibilität reduzieren. Aus diesem Grund sind auch konzeptuelle Replikationen nur bedingt geeignet falsch positive Ergebnisse zu identifizieren. Im Gegensatz zu exakten Replikationen ist man bei der Durchführung konzeptueller Replikationen nicht an einen bestimmten Pfad der Auswertung gebunden, sondern hat wie im Falle einer nicht präregistrierten Studie unzählige Möglichkeiten vorzugehen. Die Präregistrierung einer Studie bietet also, ebenso wie die direkte Replikation, eine gute Möglichkeit das Problem der vorherrschenden Freiheitsgrade zu unterbinden.

10.3. Open Science und Badges

Im Kontext der Präregistrierung wurde die Verwendung von *Badges* vorgeschlagen (Kidwell et al., 2016). Autoren und Autorinnen von Psychological Science haben seit Anfang 2014 die Möglichkeit ihre Studien mit drei verschiedenen Badges „zertifizieren" zu lassen. Neben einem Badge für die Präregistrierung einer Studie, kann man auch für das Teilen der Daten bzw. der Materialien, jeweils eigene Badges erhalten. Wie Kidwell et al., (2016) zeigen, konnte mit diesem Anreiz der Anteil an „offenen" Datensätzen und „offenen" Materialien deutlich gesteigert werden.

Diese Bestrebungen stehen in engem Zusammenhang mit dem Begriff der *Open Science.* Darunter versteht man Bemühungen in der Wissenschaft, diese frei verfügbar zu machen. Zum besseren Verständnis könnte man hier eine Analogie zur digitalen Open Source Bewegung sehen, welche Software schaffen möchte, die allen zu Verfügung steht, und deren Entstehungsprozess einfach einsehbar ist. Analog will Open Science, eine Wissenschaft forcieren, die offen ist, eine breitere Masse erreicht, und nicht nur hinter verschlossenen Türen stattfindet. Es soll zudem nicht nur das Endergebnis präsentiert werden, sondern auch der Prozess dahin offengelegt werden.

10.4. Änderung der Publikationspraktiken

Im Kontext der Replizierbarkeitsdebatte, wird auch immer wieder eine Änderung der Publikationspraktiken gefordert. Nicht nur werden derzeit vor allem Anreize für positive und neue Ergebnisse, und kaum Anreize für reliable Effekte gesetzt (Nosek et al., 2012). Das System der wissenschaftlichen Publikationen wurde noch nicht in ausreichendem Maße an neue Technologien adaptiert; die Einbindung digitaler Formate erstreckt sich nur um das klassische Publikationsformat herum, anstatt die technischen Möglichkeiten auszunutzen. Nosek und Bar-Anan (2012) nennen hier z.B. Seitenanzahlbegrenzungen in Journalen, Verzögerungen zwischen dem Zeitpunkt, zu dem ein Paper akzeptiert wird und dem Zeitpunkt der Publikation; bis auf Retractions (die v.a. Betrugsfälle betreffen), Errata und publizierte Kommentare gibt es kaum Möglichkeiten, nach der Publikation etwa widerlegte Effekte als solche zu kennzeichnen. Außerdem scheint der viel gerühmte Peer-Review nicht sonderlich objektiv zu sein (Peters & Ceci, 1982).

11. Alternativen zur Bewertung der Evidentialität von Forschungsergebnissen am Beispiel der *p*-Curves

Man kann das Reproducibility Project: Psychology als Ansatz aufgreifen, die *Evidentialität* bzw. Vertrauenswürdigkeit psychologischer Forschung zu bewerten. Replikationen im großen Stil stellen dafür aber nicht die einzige Möglichkeit dar. Aktuell werden zahlreiche statistische Kennwerte zur Bewertung der Evidentialtät von Forschungsergebnissen entwickelt. Dieser Ansatz soll hier am Beispiel der *p*-Curves (Simonsohn et al., 2014) demonstriert werden. *P*-Curves untersuchen die Verteilung von *p*-Werten. Unter der Annahme der Nullhypothese sind die *p*-Werte einer Reihe unabhängiger Tests gleichverteilt auf einem Intervall von Null bis Eins. D.h. wenn kein Effekt vorhanden ist, dann tritt jeder *p*-Wert mit gleich hoher Wahrscheinlichkeit auf, und ist somit in 5% der Fälle $p < .05$. Ist aber ein Effekt vorhanden, dann ist die Verteilung der *p*-Werte rechtsschief, d.h. man findet viele kleine *p*-Werte und weniger große *p*-Werte. Je Stärker die Power der statistischen Tests, desto schiefer wird die Verteilung der *p*-Werte. Um Forschungsergebnisse zu bewerten, schließt man nur jene Werte in die Analyse mit ein, die $p < .05$, da quasi nur diese publiziert werden. Simonsohn et al. (2014) zeigen in Simulationsstudien, dass die Verteilung der *p*-Werte linksschief ist, sofern die Effekte nur auf *p*-Hacking zurückgehen. Ist eine Verteilung rechtsschief, dann kann davon ausgegangen werden, dass evidentialer Wert vorhanden ist, *p*-hacking kann aber dennoch nicht ausgeschlossen werden. Die Autoren stellen zudem ein einfach zu bedienendes Onlineprogramm zur Verfügung. Die Nützlichkeit solcher Verfahren demonstrieren die Autoren selbst, z.b. an dem Verlauf der *p*-Curve zum Power Posing (Simmons & Simonsohn, in press.).

12. Ausblick und Diskussion

Zahlreiche bekannte psychologische Effekte werden in jüngster Vergangenheit angezweifelt, und ihre Evidenz neu geprüft. Ein gutes Beispiel stellt der Effekt des Power Posing dar. *Power-Posing* (Carney, Cuddy & Yap, 2010) zufolge hat das Einnehmen einer dominanten Körperhaltung nicht nur Auswirkungen auf das Verhalten, sondern zeigt auch auf hormoneller Ebene Wirkung. Der Effekt wurde sehr populär; der TED-Talk zum Power

Posing von Amy Cuddy gehört mit über 37 Millionen Views zu den meistgesehenen überhaupt. Das liegt sicherlich auch daran, dass dieser einfach nachvollziehbar ist, und ein Nutzen bzw. eine Anwendbarkeit für jeden einzelnen direkt abgeleitet werden kann: Wenn ich meine Körperhaltung selbstbewusst und raumeinnehmend ausrichte, dann werde ich tatsächlich selbstbewusster, und vielleicht so auch noch erfolgreicher sein. Der ursprünglich gezeigte Effekt basiert allerdings auf einer sehr kleinen Stichprobe ($N = 42$). Außerdem ist die Hypothese, dass die Körperhaltung Auswirkungen hat, nicht besonders spezifisch. Eine präregistrierte Replikation mit einem fünfmal größerem Sample (Ranehill et al., 2015) konnte den Effekt nicht bestätigen. Simmons analysierten die Evidenz zunächst mittels *p*-Curves auf ihrem Blog (Simmons & Simonsohn, 2015). Hier zeigten sie auch mittels ihres Ansatzes der Detectability, dass die ursprüngliche Studie nicht geeignet war den Effekt zu untersuchen; eine Publikation dieser *p-Curves* folgt demnächst in Psychological Science (Simmons & Simonsohn, in press). Die Verteilung der *p*-Curves von 33 Studien entspricht dem, was man erwartet, wenn kein Effekt vorhanden ist. Während Dana Carney, die Erstautorin der Originalstudie sich mittlerweile von dem Effekt distanziert, und auch die Anwendung von QRPs im ursprünglichen Paper einräumt (Carney, 2016), blieb Amy Cuddy 2015 noch unbeeindruckt (Simmons & Simonsohn, 2015). „I respectfully disagree with the interpretations and conclusions of Simonsohn et al.“ ist ihr Kommentar auf die entsprechenden *p*-Curves. Es überrascht nicht sonderlich, dass Amy Cuddy nicht kampflos aufgibt; ihren Erfolg und ihre Karriere hat sie vor allem diesem Effekt zu verdanken.

Amy Cuddys Doktormutter Susan Fiske, Sozialpsychologin an der Princeton University, verfasste ein Kommentar für den APS Observer, das vorab im Internet verbreitet wurde (Fiske, in press). Es bieten sich Einblicke in die Sichtweise einer Psychologin, die das Problem fehlender Replizierbarkeit in der Psychologie nicht sehen möchte. Fiske kritisiert hier erstens, dass Diskussion und Austausch über die sozialen Medien stattfinden, und zweitens, dass so akademische Karrieren zerstört werden.

Sie meint wohl jene Wissenschaftskarrieren, die auf jenen spektakulären Ideen aufbauen, die sich gut vermarkten lassen, deren Datenlage aber dürftig erscheint, die womöglich sogar auf einer besonders freien Auslegung der Möglichkeiten der Datenauswertung basieren. Es stellt ein Problem dar, wenn zweifelhafte Forschungspraktiken mit Ruhm belohnt werden, solide Arbeiten aber nicht. Forschung darf nicht allein auf der Grundlage einer guten Idee bewertet werden, die zugrundeliegende

Evidenz muss ebenso überzeugend sein. Das verdeutlichen die bereits angesprochenen Anreize, die in der Wissenschaft vorherrschen. Es ist nicht das Finden von Tatsachen, das erfolgreiche akademische (v.a. Sozial-)psychologen auszumachen scheint; worauf es ankommt sind möglichst neue, interessante, spektakuläre Befunde, die in der Originalstudie Signifikanz erreichten. Eine Bestätigung der Ergebnisse ist nicht notwendig.

Facebook, Twitter und die Blogs etablierter Wissenschaftler ermöglichen einen Austausch, der auf traditionellem Wege nicht möglich ist. Daher ist es kaum überraschend, dass diese Möglichkeiten auch genutzt werden. Innerhalb der klassischen akademischen Publikationswege gibt es kaum Raum für Diskussionen, also wird dieser außerhalb geschaffen. Fiske spricht von „self-appointed data police“ und „methodological terrorism“, wenn Unzulänglichkeiten von Forschungsarbeiten offen in sozialen Medien diskutiert werden. Andrew Gelman findet auf seinem Blog folgende Worte: „Methodological terrorism is when you publish a paper in a peer-reviewed journal, its claim is supported by a statistically significant t statistic of 5.03, and someone looks at your numbers, figures out that the correct value is 1.8, and then posts that correction on social media“ (Gelman, 2016). Die Verfechter und Verfechterinnen der Replizierbarkeit sind auch jene, die sich eine offene Wissenschaft wünschen. Es ist für sie vollkommen natürlich sich offen über Ergebnisse und Inhalte auszutauschen. Sie sehen Wissenschaft als in Bewegung an, eine Publikation ist nicht in Stein gemeißeltes Wissen. Das Aufgreifen von Fehlern und Missverständnissen, der Austausch darüber kann zu neuen Ideen verhelfen und so die Wissenschaft vorantreiben. Die traditionellen Wege, lassen kaum Kommunikation über Inhalte zu. Susan Fiske stößt sich aber daran, dass diese Äußerungen keinem Peer-Review ausgesetzt sind. Dabei kann man diese Diskussionen selbst als eine Art Peer-Review ansehen, mit dem großen Unterschied, dass sie keine direkten Konsequenzen auf die besprochenen Arbeiten haben. Offene Kritik an wissenschaftlichen Arbeiten muss selbstverständlich sein. Akademische Psychologen müssen damit leben, dass auch sie Fehler machen, und vor allem, dass auch sie aus diesen lernen können.

Was Fiske in ihrem Kommentar, wohl eher unfreiwillig gelingt, ist zwei zentrale Punkte zu thematisieren, deren Änderung wesentlich für die Verbesserung der psychologischen Forschung erscheinen. Zunächst ist das die Frage, was akademischen Erfolg ausmacht. Lange Zeit waren das Publikationen mit signifikanten Ergebnissen, die neuartige und überraschende Effekte belegen (Nosek et al., 2012). Die Ängste Fiskes, dass

Karrieren etablierter Psychologen auf dem Spiel stehen, zeigen, dass diese Kriterien einer Änderung unterworfen ist – hin zu replizierbaren Effekten.

Außerdem missfallen Susan Fiske die Ausmaße der Diskussion in sozialen Medien. Diese verdeutlichen aber vor allem, dass die traditionellen, akademischen Kommunikationswege für in einer digitalen Welt lebende Wissenschaftler unzureichend sind. Die Etablierung von einfachen Alternativen wird in die eigene Hand genommen, es werden zunächst die einfachsten Mittel gewählt, wie z.B. die Erstellung von Facebookgruppen (*PsychMAP* von Michael Inzlicht oder *Psychological Methods Discussion Group* von Ulrich Schimmack). Brian Nosek widmet sich in *Scientific Utopia I: Opening Scientific Communication* der Diskussion der Problematik akademischer Kommunikation und schlägt mögliche Lösungen vor (Nosek & Bar-Anan, 2012).

Die Ursachen der Replizierbarkeitskrise sind also nicht nur in den besprochenen statistischen Problemen zu suchen, die Probleme liegen tiefer, sind systematischer Natur und somit auch im gesamten akademischen Publikationssystem zu suchen. Doch es sind bereits erste Änderungsansätze zu vernehmen. Nicht nur hat die Thematik mittlerweile breite Aufmerksamkeit bekommen, dieses Interesse hat auch Veränderungen angestoßen. *Basic and Applied Social Psychology* (*BASP*) beispielsweise verzichtet seit Anfang 2015 auf *p*-Werte (Trafimow & Marks, 2015). *Comprehensive Results in Social Psychology* (CRSP) ist das erste sozialpsychologische Journal, das nur präregistrierte Studien veröffentlicht (Jonas & Cesario, 2015). Die breit geführte Debatte um die Vertrauenswürdigkeit psychologischer Forschung kann auch als ein Neuanfang gesehen und genützt werden. Die Methoden und Publikationspraktiken der psychologischen Forschung sind jedenfalls einem grundlegenden Wandel unterworfen, der nicht mehr aufzuhalten ist. Die Art und Weise wie Daten erhoben, ausgewertet, interpretiert und publiziert werden, wird in wenigen Jahren eine andere sein.

Literaturverzeichnis

Altmetric. (2016). Estimating the reproducibility of psychological science Overview of attention for article published in Science, August 2015. Abgerufen 22. Oktober 2016, von https://www.altmetric.com/details/4443094#score

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Anderson, C., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C., … Zuni, K. (2016). Response to comment on "Estimating the reproducibility of psychological science". *Science*, *351*, 1037. doi: 10.1126/science.aad9163

Armitage, P., Mcpherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*, *132*, 235–244. doi: 10.2307/2343787

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119. doi: 10.1002/per.1919

Atmanspacher, H., & Maasen, S. (Eds.). (2016). *Reproducibility : principles, problems, practices, and prospects*. John Wiley & Sons.

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069–1077. doi: 10.1177/0956797616647519

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. doi: 10.1177/1745691612459060

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. doi: 10.1037/a0021524

Button, K. S., Ioannidis, J. P. , Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi: 10.1038/nrn3475

Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, *21*(10), 1363–1368. doi: 10.1177/0956797610383437

Carney, D. (2016). My position on "power poses". Abgerufen am 18. Oktober 2016 von http://faculty.haas.berkeley.edu/dana_carney/pdf_My%20position%20on%20power%20poses.pdf

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378–399. doi: 10.17763/haer.48.3.t490261645281841

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145–153. doi: 10.1037/h0045186

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi:10.1037/0033-2909.112.1.155

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*(4), 286–300. doi: 10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. London, England: Routledge.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, *3*(4), 299–311. doi: 10.1207/s15328031us0304_5

Dickersin, K., & Min, Y. (1993). Publication bias: The problem that won't go away. *Annals of the New York Academy of Sciences*, *703*(1), 135–148. doi: 10.1111/j.1749-6632.1993.tb26343.x

Döring, N., & Bortz, J. (2015). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (5. Aufl.). Berlin: Springer.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621. doi: 10.3389/fpsyg.2015.00621

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLOS ONE*, *4*, e5738. doi: 10.1371/journal.pone.0005738

Fanelli, D. (2010). „Positive" results increase down the hierarchy of the sciences. *PLOS ONE*, *5*, e10068. doi: 10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. doi: 10.1007/s11192-011-0494-7

Faul, F., Erdfeld, E., Lang, A. G., & Buchner, A. (2007). A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. doi: 10.3758/BF03193146

Fisher, R. (1971). *The Design of Experiments* (8th ed.). New York: Hafner.

Fiske, S. T. (in press). Mob rule or wisdom of crowds? *APS Observer*. Abgerufen von https://www.dropbox.com/s/9zubbn9fyi1xjcu/Fiske presidential guest column_APS Observer_copy-edited.pdf

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. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–91. doi: 10.3758/s13423-012-0322-y

Gelman, A. (3. März 2016). More on replication crisis [Blog post]. Abgerufen 16. Oktober 2016, von http://andrewgelman.com/2016/03/03/more-on-replication-crisis/

Gelman, A. (20. September 2016). "Methodological terrorism" [Blog post]. Abgerufen 16.

Oktober 2016, von http://andrewgelman.com/2016/09/20/methodological-terrorism/

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*(5), 587–606. doi: 10.1016/j.socec.2004.09.033

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20. doi: 10.1037/h0076157

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, *3*(4), 367–388. doi: 10.1016/0167-2681(82)90011-7

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online*, *7*(1), 1–20.

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Distributions of p-values smaller than .05 in psychology: What is going on? *PeerJ*, *4*, e1935. doi:10.7717/peerj.1935

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. doi: 10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Med*, *2*, e124. doi: 10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. doi: 10.1177/0956797611430953

Jonas, K. J., & Cesario, J. (2015). How can preregistration contribute to research in our field? *Comprehensive Results in Social Psychology*, 1–7. doi: 10.1080/23743603.2015.1070611

Judd, C. M., & Gawronski, B. (2011). Editorial comment. *Journal of Personality and Social Psychology*, *100*(3), 406. doi: 10.1037/0022789

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. doi: 10.1207/s15327957pspr0203_4

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., … Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, *14*, e1002456. doi: 10.1371/journal.pbio.1002456

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A „many labs“ replication project. *Social Psychology*, *45*(3), 142–152. doi: 10.1027/1864-9335/a000178

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*, 863. doi: 10.3389/fpsyg.2013.00863

Lakens, D. (6. März 2016). The statistical conclusions in Gilbert et al (2016) are completely invalid [Blog post]. Abgerufen 26. Oktober 2016, von

http://daniellakens.blogspot.com/2016/03/the-statistical-conclusions-in-gilbert.html

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. doi: 10.1177/1745691612460688

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. doi: 10.1146/annurev.psych.59.103006.093735

Neuliep, J. W., & Crandall, R. (1991). Editorial bias against replication research. In J. W. Neuliep (Hrsg.), *Replication research in the social sciences* (S. 85–90). Newbury Park, CA: Sage.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, *23*(3), 217–243. doi: 10.1080/1047840X.2012.692215

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. doi: 10.1177/1745691612459058

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, *506*(7487), 150–152. doi: 10.1038/506150a

O'Brien, R. G., & Castelloe, J. (2007). Sample size analysis for traditional hypothesis testing: concepts and issues. In A. Dmitrenko, C. Chuang-Stein, & R. B. D'Agostino (Eds.), *Pharmaceutical statistics using SAS: A practical guide* (S. 237–271). Cary, NC: SAS Institute.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657–660. doi: 10.1177/1745691612462588

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

Pashler, H., & Harris, C. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 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*(6), 528–530. doi: 10.1177/1745691612465253

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*(3), 319–332. doi: 10.1177/1745691614528519

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, *5*(2), 187–195. doi: 10.1017/S0140525X00011183

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance

in a large sample of men and women. *Psychological Science*, *26*(5), 653–656. doi: 10.1177/0956797614553946

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. doi:10.1037/0033-2909.86.3.638

Rotton, J., Levitt, M. J., & Foos, P. (1993). Citation impact, rejection rates, and journal value. *American Psychologist*, *48*(8), 911–912.

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*(2), 90–100. doi: 10.1037/a0015108

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309–316. doi: 10.1037/0033-2909.105.2.309

Simmons, D., & Holocombe, A. (28. Februar 2014). Registered replication reports - A new article type at perspectives on psychological science. Abgerufen 25. Oktober 2016 von https://www.psychologicalscience.org/publications/observer/2014/march-14/registered-replication-reports.html

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. doi: 10.1177/0956797611417632

Simmons, J. P., & Simonsohn, U. (in press). Power posing: p-curving the Evidence. *Psychological science*.

Simmons, J. P., & Simonsohn, U. (5. August 2015). Power posing: Reassessing the evidence behind the most popular TED talk [Blog post]. Abgerufen 25. Oktober 2016, von http://datacolada.org/37

Simonsohn, U. (2015) Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. doi: 10.1177/0956797614567341

Simonsohn, U. (3. März 2016). Evaluating replications: 40% full ≠ 60% empty [Blog post]. Abgerufen 1. Oktober 2016, von http://datacolada.org/47

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. doi: 10.1037/a0033242

Srivastava, S. (3. März 2016). Evaluating a new critique of the Reproducibility Project [Blog post]. Abgerufen 25. September 2016, von https://hardsci.wordpress.com/2016/03/03/evaluating-a-new-critique-of-the-reproducibility-project/

Sterling, T. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34.

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, *7*(6), 670–688. doi: 10.1177/1745691612460687

Suleiman, R. (1996). Expectations and fairness in a modified ultimatum game. *Journal of*

Economic Psychology, *17*, 531–554. doi: 10.1016/S0167-4870(96)00029-3

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1–2. doi: 10.1080/01973533.2015.1012991

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. doi: 10.1037/h0031322

van Dijk, E., van Kleef, G. A, Steinel, W., & van Beest, I. (2008). A social functional approach to emotions in bargaining: when communicating anger pays and when it backfires. *Journal of personality and social psychology*, *94*(4), 600–614. doi: 10.1037/0022-3514.94.4.600

van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2004). The interpersonal effects of emotions in negotiations: A motivated information processing approach. *Journal of Personality and Social Psychology*, *87*(4), 510–528. doi: 10.1037/0022-3514.87.4.510

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. doi: 10.1037/a0022790

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *American Statistician*, *1305*(March), 00–00. doi: 10.1080/00031305.2016.1154108

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

Abstract

Die psychologische Fachliteratur besteht de facto nur aus positiven, d.h. statistisch signifikanten Ergebnissen. Die niedrige statistische Teststärke und Verbreitung von Questionable Research Practices bzw. *p*-hacking in der psychologischen Forschung lässt vermuten, dass der Anteil falsch positiver Ergebnisse höher ausfällt, als bisher vermutet. Replikationen stellen eine Möglichkeit dar empirische Studien auf ihre Evidentialität bzw. Vertrauenswürdigkeit zu prüfen, und können dazu beitragen falsch positive Ergebnisse zu identifizieren. Bislang stellten Replikationen in der Psychologie eine Seltenheit dar. Das Reproduciblity Project: Psychology war ein erster Versuch die Rate der Replizierbarkeit psychologischer Forschung empirisch abzuschätzen, indem eine Stichprobe von 100 psychologischen Studien wiederholt wurde. Das Projekt erhielt breite mediale Aufmerksamkeit, wohl auch weil nur 36% der replizierten Studien wieder zu signifikanten Ergebnissen führten. Im Zuge der vorliegenden Masterarbeit wurde eine Replikation von van Dijk et al. (2008, JPSP) im Rahmen des Reproducibility Project: Psychology durchgeführt. Trotz einer Power von 95% die ursprüngliche Effektstärke wieder zu finden, konnte das Originalergebnis nicht repliziert werden. Es werden verschiedene Methoden zur Bestimmung der Stichprobengröße von Replikationen skizziert, und Ansätze der Bewertung von Replikationsergebnissen, sowie ihre Limitationen diskutiert.

Schlagwörter:
Replikation, Reproducibility Project: Psychology, *p*-Hacking, QRPs, Teststärke, falsch Positive

Abstract

The field of psychology is heavily biased towards publishing studies with positive, i.e., statistically significant results. Low statistical power and the now better known prevalence of Questionable Research Practices (QRPs) or *p*-hacking both suggest a higher rate of false positives than previously assumed. Direct replications provide a way to assess the evidential value of empirical studies and can help to identify false positives. To date replications have been rarely conducted in the field of psychology. The Reproducibility Project: Psychology constitutes a first effort to estimate the reproducibility of psychological science, by replicating a sample of 100 psychological studies. It received widespread attention in the media, assumingly also due to the fact that only 36% replications resulted in significant outcomes. Within the present master thesis, a replication of van Dijk et al. (2008, JPSP) was conducted as part of the Reproducibility Project: Psychology. Methods to estimate the required sample size for replications are discussed, as are problems regarding the evaluation of replication results.

Keywords:
replication, Reproducibility Project: Psychology, *p*-hacking, QRPs, statistical power, false positives

Curriculum Vitae

Name	Agnieszka Malgorzata Slowik
Geburtsdatum	20. April 1986
Sprachkenntnisse	Deutsch, Englisch (fließend), Polnisch (fließend)

Publikationen

2016	Kidwell, M. C., Lazarevic, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, 14, e1002456
2015	Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716.

Tagungsbeiträge

2016	Stieger, S., Slowik A., Sonnleitner, C., Kuhlmann T. & Voracek M. (2016). *Replikationen im großen Stil: Erfahrungen aus dem Reproducibility Project: Psychology am Beispiel von drei exakten Replikationen*. Beitrag präsentiert auf der 12. Tagung der Österreichischen Gesellschaft für Psychologie, 31.3. - 2.4. 2016, Universität Innsbruck.

Ausbildung

2015	Abschluss Bachelor of Science, Studienfach Psychologie
2007 – 2009	Studium der Architektur an der Technischen Universität Wien
seit 2004	Studium der Psychologie an der Universität Wien
06/2004	Matura mit ausgezeichnetem Erfolg am Bundesrealgymnasium Franklinstraße 21, 1210 Wien

Praktische Erfahrungen

seit 10/206	Studienassistentin am Institut für Psychologische Grundlagenforschung und Forschungsmethoden, Arbeitsbereich Forschungsmethoden
03/2013 – 09/2016	Standard Service GmbH der Tageszeitung „Der Standard“.
04/2015 - 07/2015	Praktikum am Institut für Angewandte Psychologie, Universität Wien: Mitarbeit am Projekt Arbeit im Wandel