Die Vertrauenswürdigkeit von Replikationen: Der Einfluss von Präregistrierungen auf den Replikationserfolg

The Trustworthiness of Replications: Are They Hacked, Too?

Masterarbeit
Im Studiengang Psychologie
der Fakultät Humanwissenschaften
der Otto-Friedrich-Universität Bamberg

Themenstellerin: Prof. Dr. Astrid Schütz

vorgelegt von:

Name	Leonard David Kaiser
Abgabetermin	15.11.2024

URN: urn:nbn:de:bvb:473-irb-1050139
DOI: https://doi.org/10.20378/irb-105013

Abstract

Theorie: Die Psychologie befindet sich in einer Replikationskrise, die unter anderem auf die „Publish or Perish“ Kultur zurückzuführen ist. Diese Kultur führt dazu, dass in großer Anzahl spannende und signifikante Studien veröffentlicht werden müssen, um im akademischen System Anerkennung zu finden. P-Hacking und andere fragwürdige Forschungspraktiken sind Auswirkungen dieser Kultur. In der vorliegenden Studie wurde untersucht, ob Replikationsstudien ebenfalls von Hacking betroffen sind. Es wurden dazu verschiedene Anreize für das Manipulieren von Replikationen vorgestellt. Besonders wurde auf Null-Hacking eingegangen, das durch die hohe Aufmerksamkeit auf nicht replizierte Befunde, begünstigt werden könnte.

Methodik: Die Methode Präregistrierung gilt als ein direktes Mittel, um Freiheitsgrade von Forschenden einzuschränken und dabei Hacking entgegenzuwirken. Daher untersuchte die vorliegende präregistrierte Studie anhand der bisher umfassendsten Replikationsdatenbank (FORRT Replication Database), ob der Präregistrierungsstatus der Replikationsstudie den Zusammenhang zwischen Originaleffekt und Replikationseffekt moderiert. Nach allen Exklusionen wurden 361 Replikationseffekte berücksichtigt, von denen 124 präregistriert waren. Es wurde darüber hinaus ermittelt, ob die Ähnlichkeit (Closeness) zwischen der Originalstudie und der Replikationsstudie sowie verschiedene Präregistrierungsvorlagen den Zusammenhang zwischen Originaleffekt und Replikationseffekt moderieren. Zudem wurde explorativ untersucht, ob präregistrierte Replikationen sich hinsichtlich des Replikationserfolgs von nicht-präregistrierten Replikationen unterscheiden.

Ergebnis: Der Präregistrierungsstatus der Replikation und die Closeness moderierten den Zusammenhang zwischen dem Originaleffekt und dem Replikationseffekt nicht. Nur eine Präregistrierungsvorlage moderierte den Zusammenhang zwischen Originaleffekt und Replikationseffekt, während alle anderen Vorlagen keinen Effekt auf den Zusammenhang aufwiesen. Zudem zeigten präregistrierte Replikationen keine Unterschiede im Replikationserfolg im Vergleich zu nicht-präregistrierten Replikationen.

Diskussion: Diese Ergebnisse liefern erste Hinweise darauf, dass Replikationen möglicherweise nicht von Hacking betroffen sind. Aufgrund der vorgestellten Limitationen in dieser Studie, braucht es weitere Forschung, die dieses Phänomen untersucht. Denn manipulierte Replikationen erwecken den Anschein von wissenschaftlichem Fortschritt und Qualitätskontrolle, können jedoch langfristig die Entwicklung ganzer Forschungsfelder hemmen.

Der Datensatz und Code sind verfügbar unter: https://osf.io/rkj4x/

Keywords: replication, p-hacking, null-hacking, preregistration, meta-analysis, open science

Abstract

Theory: Psychology is currently facing a replication crisis, which can be attributed in part to the „Publish or Perish“ culture. This culture incentivizes significant, exciting, and numerous publications as a means of gaining rewards within the academic system. P-hacking and other questionable research practices are consequences of this culture. The present study examined whether replication studies are also susceptible to hacking. Various incentives for the manipulation of replications were discussed, with a particular focus on null-hacking, which may be incentivized by the heightened attention on failed replications.

Method: The method of preregistration is considered a direct approach to limit researchers' degrees of freedom and thereby mitigate hacking. Accordingly, this preregistered study utilized the most comprehensive replication database to date (FORRT Replication Database) to examine whether the preregistration status of replication studies moderated the relationship between the original effect and the replication effect. After applying all exclusions, 361 replication effects were analyzed, of which 124 were preregistered. Additionally, the study explored whether the closeness between the original study and the replication study, as well as different preregistration templates, moderated the relationship between the original effect and the replication effect. Moreover, it was examined whether preregistered replications differed in replication success compared to non-preregistered replications.

Results: The preregistration status of replications and the closeness did not moderate the relationship between the original effect and the replication effect. Only one preregistration template moderated the relationship between the original effect and the replication effect, while all other templates had no effect on this relationship. Furthermore, preregistered replications did not exhibit differences in replication success compared to non-preregistered replications.

Discussion: These results provide preliminary evidence suggesting that replications may not be affected by hacking. Given the limitations outlined in this study, further research is needed to

investigate this phenomenon. Hacked replications create the illusion of scientific progress and enhanced transparency, while in reality, they may hinder entire fields of research.

The dataset and code are available at: https://osf.io/rkj4x/

Keywords: replication, p-hacking, null-hacking, preregistration, meta-analysis, open science

Theorie

Die Psychologie sieht sich nicht zum ersten Mal mit einer Vertrauenskrise konfrontiert. Bereits in den 1960er und 1970er-Jahren diskutierten Wissenschaftler*innen über eine Krise in der Psychologie (Bakan, 1970; Elms, 1975; Lakens, 2023). Damals standen insbesondere Versuchsleiter*inneneffekte (Rosenthal, 1966) sowie Zweifel an der Generalisierbarkeit psychologischer Erkenntnisse (Gergen, 1973) im Mittelpunkt.

Fast vier Jahrzehnte später sind erneut Bedenken über eine Vertrauenskrise aufgekommen. Eine aktuelle Umfrage ergab, dass 90 % der Forschenden aus verschiedenen Disziplinen, darunter auch der Psychologie, eine Reproduzierbarkeitskrise in der Wissenschaft wahrnehmen (Baker, 2016). Von den 1576 befragten Wissenschaftler*innen gaben 52 % an, dass eine erhebliche Krise bestehe, während weitere 38 % eine weniger gravierende Krise sahen. Lediglich 3 % verneinten das Vorhandensein einer Krise, während 7 % sich unentschlossen äußerten.

Bereits während der früheren Krise spielte das Thema Replikation eine zentrale Rolle. Es wurde erkannt, dass Replikationsversuche häufig scheiterten (Epstein, 1980; Greenwald, 1976) und dass Befunde, die durch Replikationen gestützt werden, vorrangig publiziert werden sollten (Lubin, 1957).

Sind Reproduzierbarkeit und Replizierbarkeit das gleiche? Reproduzierbarkeit und Replizierbarkeit werden in der Literatur gelegentlich synonym verwendet (z.B. Open Science Collaboration, 2015) und an anderer Stelle auseinanderdividiert (z.B. Schöch, 2023). In dieser Studie wurde der Begriff „Replikation“ verwendet, um sich auf jede Studie zu beziehen, die dieselbe Hypothese wie die Originalstudie testet, gemäß der Definition von Hüffmeier et al. (2016). Diese Definition entspricht derjenigen, die in der FORRT Replication Database (FReD) verwendet wird, welche als Untersuchungsgrundlage dieser Studie diente.

Ein wesentlicher Unterschied zwischen der Krise in den 1960er und 1970er-Jahren und der gegenwärtigen Krise liegt in der Einschätzung der Replizierbarkeit. Damals war unklar, inwieweit Forschungsergebnisse insgesamt replizierbar waren. Ein Grund dafür war die fehlende Präzision in den Methodenabschnitten der Originalstudien, was die Durchführung von Replikationen erschwerte

(Pereboom, 1971). Als eine positive Konsequenz aus der damaligen Krise verbesserten sich die methodischen Beschreibungen, was heute sowohl die Durchführung von Replikationsstudien erleichterte als auch groß angelegte Replikationsprojekte ermöglichte.

Im Jahr 2015 veröffentlichte die Open Science Collaboration einen umfangreichen Replikationsversuch (Open Science Collaboration, 2015). Ziel war es, 100 psychologische Experimente zu replizieren, die in drei renommierten Fachzeitschriften veröffentlicht worden waren. Dabei konnten 36 von 100 Befunden erfolgreich repliziert werden, wobei die Effektstärken im Durchschnitt etwa halb so groß ausfielen, wie in den Originalstudien. In der Zusammenschau von zahlreichen groß angelegten Replikationsprojekten aus verschiedenen Disziplinen (z. B. Camerer et al., 2016, 2018; Klein et al., 2014, 2018; Open Science Collaboration, 2015) wurde eine durchschnittliche Replikationsrate von 46 % festgestellt (Röseler et al., 2022). Die bislang umfangreichste Replikationsdatenbank, FReD, umfasst aktuell 2348 Replikationseinträge (Stand 9. Juli 2024), vorwiegend aus der Psychologie, und verzeichnet eine Replikationserfolgsrate von 54 % (Röseler et al., 2024).

Die Wahrnehmung der in der Umfrage von Baker (2016) befragten Wissenschaftler*innen, dass eine Krise besteht, wird daher durch empirische Daten gestützt und ist insbesondere in der Psychologie gut dokumentiert. Die Replikationsraten „suggest that there is room to improve reproducibility in psychological science“ (Open Science Collaboration, 2015, p.15). Gleichzeitig ist die Durchführung von Replikationsstudien in der Psychologie dennoch keine gängige Praxis. So betrug der Anteil der direkten Replikationen in den 100 einflussreichsten Fachzeitschriften in der Psychologie im Zeitraum 2010 bis 2021 nur 0,2% (Clarke et al., 2024).

Gründe für die Replikationskrise

Die Ursachen der Replikationskrise sind vielfältig. Die vorliegende Studie konzentriert sich auf die „Publish or Perish“-Kultur (Fanelli, 2010) und akademische Belohnungssysteme, die auch während der früheren Krise als beitragende Faktoren identifiziert wurden (Elms, 1975; Mahoney, 1979). Weitere Faktoren, die zur aktuellen Replikationskrise beitragen, sind die geringe statistische Power in den Originalstudien (Schimmack, 2020), Kontextsensitivität (Van Bavel et al., 2016), die

Theoriekrise (Oberauer & Lewandowsky, 2019) sowie Bedenken hinsichtlich der Messgenauigkeit und Validität von Instrumenten (Flake et al., 2017).

Das Verständnis akademischer Belohnungssysteme erfordert Einblicke in die Entscheidungsprozesse der Fachzeitschriften. Diese stehen vor der Aufgabe, aus einer Vielzahl von Einreichungen Studien zur Veröffentlichung auszuwählen, wobei häufig Studien bevorzugt werden, die als „newsworthy" gelten (Rothstein et al., 2005). Was macht eine Studie berichtenswerter als eine andere?

Studien scheinen berichtenswerter, wenn sie positive Ergebnisse präsentieren, die Hypothesen bestätigen. Zunächst wirken positive Befunde möglicherweise ansprechender, da sie neue Erkenntnisse zum Forschungsfeld beitragen. Nullbefunde leisten jedoch ebenfalls einen bedeutenden Beitrag zum Wissensfortschritt und sind daher genauso wichtig. Dennoch werden Studien mit signifikanten Ergebnissen häufiger veröffentlicht (Cooper et al., 1997; Dickersin et al., 1987; Sterling, 1959).

Darüber hinaus ist Forschenden heutzutage bewusst, dass hochrangige empirische Fachzeitschriften spektakuläre, unerwartete und interessante Ergebnisse erwarten (Giner-Sorolla, 2012; Miller & Bamberger, 2016; Nosek et al., 2012). Diese Ergebnisse gelten aufgrund dem Wunsch nach Neuheit und Sensationalismus als interessanter (Berinsky et al., 2021). Dieser Bias und dessen Konsequenzen sind umfangreich in der Literatur dokumentiert und er wird als Publikationsbias bezeichnet (Franco et al., 2014; Pfeiffer et al., 2011; Rothstein et al., 2005).

Antonakis (2017) identifizierte fünf Krankheiten in der Wissenschaft, die wissenschaftlichen Fortschritt beeinträchtigen. Dazu gehören „significosis, an inordinate focus on statistically significant results", „neophilia, an excessive appreciation for novelty" und „disjunctivitis, a proclivity to produce large quantities of redundant, trivial, and incoherent works", die wesentlich zum Publikationsbias beitragen.

Als Folge des Publikationsbias werden Nullbefunde – Studien, in denen Hypothesen nicht unterstützt werden – metaphorisch betrachtet im „File-Drawer" abgelegt (Rosenthal, 1979). Diese

Nullbefunde werden als nicht ausreichend „newsworthy“ angesehen (Miller & Bamberger, 2016; Nosek et al., 2012).

Für Wissenschaftler*innen hat der Publikationsbias zur Folge, dass sie Studien einreichen müssen, die signifikante Ergebnisse liefern und viel Aufmerksamkeit erregen (Bakker et al., 2012). Gleichzeitig sind sie gezwungen, häufig und in hochrangigen Fachzeitschriften zu publizieren, um ihre akademischen Karrieren voranzutreiben (Bakker et al., 2012; Martin, 1992; Nosek et al., 2012; Sovacool, 2008). Diese Entwicklung resultiert aus der stark wettbewerbsorientierten Natur der Wissenschaft, in der Kennzahlen wie die Anzahl der Publikationen und die Anzahl der Zitationen zentrale Erfolgsindikatoren darstellen (Fanelli, 2010). Diese Kennzahlen werden auch als die Währung der Wissenschaft bezeichnet (Röseler, in preparation). „Publication influences hiring, salary, promotion, tenure, and grant decisions” (Nosek et al., 2012, p. 2). Infolgedessen sind Wissenschaftler*innen dem sogenannten „Publikationsdruck“ ausgesetzt (Nosek et al., 2012).

Darüber hinaus sind die Verträge von Wissenschaftler*innen, insbesondere in Deutschland, in der Regel befristet und von kurzer Dauer. Um diesen prekären Arbeitsbedingungen zu entkommen und unbefristete Verträge sowie Forschungsförderungen zu erhalten, wird häufiges Veröffentlichen in hochrangigen Fachzeitschriften als die effektivste Strategie angesehen (Röseler, in Vorbereitung). Daher werden Wissenschaftler*innen belohnt, die Quantität ihrer Veröffentlichungen über deren Qualität zu stellen (Larivière & Costas, 2016; Nosek et al., 2012). Die Publikation zahlreicher kurzer Artikel wird als vorteilhaft erachtet (Ellemers, 2013), da sie die Wahrscheinlichkeit erhöhen, dass einige Publikationen häufig zitiert werden (Larivière & Costas, 2016).

Zusammenfassend lässt sich festhalten, dass Studien, die signifikante Ergebnisse berichten und erhebliches Interesse wecken, eine höhere Wahrscheinlichkeit haben, veröffentlicht zu werden. Gleichzeitig werden Wissenschaftler*innen dazu motiviert, häufig zu veröffentlichen, was sie in Versuchung führen kann, ihre Forschung zu manipulieren (Nosek et al., 2012).

Freiheitsgrade

Können Wissenschaftler*innen ihre eigene Forschung manipulieren? Die Antwort könnte in den Worten des Nobelpreisträgers Ronald Coase liegen: „If you torture the data long enough it will confess“ (Good, 1972). Wissenschaftler*innen haben bei der Planung, Durchführung und Auswertung ihrer Studien „Researchers' Degrees of Freedom“ (Simmons et al., 2011). Wicherts et al. (2016) identifizierten 34 solcher Freiheitsgrade, die sie in der psychologischen Forschung für möglich halten. Wenn Forschende diese Freiheitsgrade nutzen, um ihre Ergebnisse in Richtung signifikanter Ergebnisse zu manipulieren, spricht man von „P-Hacking“ (Simmons et al., 2013). P-Hacking ist ein Teilbereich der sogenannten „questionable research practices“ (QRPs).

„QRPs are intentional behaviors that capitalize on the gray area of acceptable scientific behavior or exploit the Degree of Freedom, which can contribute to the irreproducibility of results by increasing the probability of false positive results“ (“Replication crisis,” 2024). Zu den weiteren Formen von QRPs gehören HARKing (Hypothesizing After the Results are Known), selektive Berichterstattung, das Weglassen von Daten aus Pilotstudien sowie die nachträgliche Datenveränderung nach Überprüfung der Signifikanz (Kravitz & Mitroff, 2020).

Gegenmaßnahme zu P-Hacking

Die direkteste Methode zur Einschränkung der Freiheitsgrade von Wissenschaftler*innen ist die Präregistrierung. Bei diesem Ansatz werden alle möglichen Entscheidungen im Voraus festgelegt, wodurch die Freiheitsgrade eingeschränkt werden, die zu verzerrten Ergebnissen führen können. Die Vorteile der Präregistrierung wurden in der Literatur klar benannt (Lakens, 2019; Nosek et al., 2018). Infolgedessen wird die Präregistrierung als primäre Methode zur Bekämpfung von P-Hacking empfohlen (Simonsohn, 2023). Vor zehn Jahren handelte es sich bei der Präregistrierung noch um eine relativ unbekannte Methode, doch Daten aus vier renommierten Fachzeitschriften zeigten, dass 43 % der Studien aus dem Jahr 2022 präregistriert waren (Simonsohn, 2023).

Die Evaluation von Präregistrierungen umfasst zwei zentrale Kriterien (Heirene et al., 2021; Röseler et al., 2022). Erstens sollte eine Präregistrierung alle möglichen Entscheidungen im Voraus

antizipieren und dokumentieren. Zweitens sollte das Format der Präregistrierung eine schnelle und effiziente Überprüfung durch die Evaluator*innen ermöglichen.

Die Wirksamkeit von Präregistrierungen zur Reduzierung von P-Hacking wurde bislang empirisch wenig untersucht. Van den Akker et al. (2023a) fanden keinen Unterschied in der Häufigkeit signifikanter Ergebnisse zwischen präregistrierten und nicht-präregistrierten Studien und schlossen daraus, dass die Präregistrierung P-Hacking nicht wirksam einschränkt. Allerdings unterschieden sie nicht zwischen Präregistrierungen mit und ohne festgelegtem Analyseplan. Brodeur et al. (2024) zeigten, dass Präregistrierungen nur dann effektiv sind, wenn sie einen festgelegten Analyseplan enthalten, der angibt, welche Hypothesen getestet und wie sie analysiert werden. In der Psychologie sind festgelegte Analysepläne ein Standardbestandteil von Präregistrierungen, während in der Ökonomie Präregistrierungen von RCTs, die im American Economic Association Registry hochgeladen werden, lediglich die Option bieten, einen festgelegten Analyseplan hinzuzufügen. Ähnlich zeigten Scheel et al. (2021), dass der Anteil positiver Ergebnisse in Registered Reports (44 % positive Ergebnisse), in denen detaillierte festgelegte Analysepläne erforderlich sind, signifikant niedriger war als in einer zufälligen Stichprobe von konfirmatorischen Studien aus der Standardliteratur (96 % positive Ergebnisse). Die Einbeziehung des Analyse-Codes in die Präregistrierung erhöht zusätzlich die Transparenz des Verfahrens weiter und reduziert die Freiheitsgrade der Forschenden. Zwei weitere zentrale Faktoren zur Bewertung der Wirksamkeit von Präregistrierungen sind die „producibility“ und die „consistency“ (van den Akker et al., 2024). Producibility bezieht sich auf das Ausmaß, in dem die Studie basierend auf den Informationen der Präregistrierung durchgeführt werden kann. Consistency hingegen gibt an, inwieweit die durchgeführte Studie mit dem präregistrierten Plan übereinstimmt.

Van den Akker et al. (2024) entdeckten, dass in präregistrierten Studien häufig Abweichungen von den Präregistrierungen vorliegen, die nicht berichtet werden. Dies steht im Einklang mit den Ergebnissen von Claesen et al. (2021), die in der Zeitschrift *Psychological Science* herausfanden, dass 90% aller präregistrierten Studien Abweichungen enthielten. In einer detaillierten Untersuchung spezifischer Abweichungen von Hypothesen (van den Akker et al., 2023b), wurde in

einem Datensatz von 459 präregistrierten Studien festgestellt, dass die Hälfte der tatsächlich durchgeführten Studien Hypothesen ausließen, die ursprünglich präregistriert waren. Etwa 20% der präregistrierten Studien änderten sogar die Richtung der Hypothese in der durchgeführten Studie. Trotz der hohen Prävalenz und Relevanz von Abweichungen von der Präregistrierung werden Abweichungen im Begutachtungsprozess selten überprüft (Syed, 2023). Dies könnte zum Teil daran liegen, dass Forschende oft keine präzisen Links zu ihren Präregistrierungen bereitstellen (Simonsohn, 2023). Beispielsweise stellen sie möglicherweise nur einen Link zu einem OSF-Ordner zur Verfügung, wo das Finden der spezifischen Präregistrierungsvorlage zeitaufwendig sein kann.

Insgesamt verdeutlichen die Abweichungen von der Präregistrierung die entscheidende Bedeutung des Kriteriums der Vollständigkeit von Präregistrierungen. Wenn einige Freiheitsgrade uneingeschränkt bleiben, kann eine effektive Eindämmung von P-Hacking und anderen fragwürdigen Forschungspraktiken nicht gewährleistet werden.

Umso viele Freiheitsgrade wie möglich zu berücksichtigen, ist die Wahl der Präregistrierungsvorlage ebenso entscheidend. Van den Acker et al. (2024) stellen fest, dass die Auswahl einer detaillierten Präregistrierungsvorlage die Producibility verbessern kann, da die Vorlage eine größere Bandbreite an Entscheidungen erfasst.

Abweichungen von präregistrierten Plänen sind nicht immer problematisch und können im Rahmen von Forschungsarbeiten sogar unvermeidlich sein (Nosek et al., 2019). Dennoch ist eine transparente Berichterstattung und Rechtfertigung solcher Abweichungen von wesentlicher Bedeutung. Abweichungsvorlagen, wie sie Heirene et al., (2021, https://osf.io/6fk87) vorschlagen, können Forschenden dabei helfen, Abweichungen von ihren präregistrierten Plänen zu dokumentieren und zu erklären.

Belohnungsstrukturen in Replikationen

Während viel Aufmerksamkeit auf fragwürdige Forschungspraktiken und deren Gegenmaßnahmen innerhalb von Originalliteratur gefallen ist, wurden fragwürdige

Forschungspraktiken in Replikationsstudien bisher kaum beachtet. Die Frage kommt auf: Wie sind die Belohnungsstrukturen innerhalb Replikationsliteratur aufgebaut?

Schon in der Krise in den 1960er und 1970er-Jahren war den Forscher*innen bewusst, dass Replikationsstudien schwieriger zu veröffentlichen waren. “Some journals explicitly state that they do not accept replication studies in principle, while others implicitly follow a similar policy; it is not surprising that few are ever published” (Fishman & Neigher, 1982, p. 539).

Das hatte sich in den Anfängen der zweiten Krise in 2011 noch nicht verändert, als Replikationsstudien von Bem’s pre-cognition Studie im Journal JPSP nicht veröffentlicht wurden mit der Begründung: „This journal does not publish replication studies, whether successful or unsuccessful” and “We don’t want to be the Journal of Bem Replication” (Aldhous, 2011).

In der Konsequenz wurden Replikationsstudien nicht in peer-reviewed Zeitschriften veröffentlicht, sondern in grauer Literatur wie zum Beispiel Online-Blogs (z.B., datacolada.org, openmkt.org), Open Science Framework Registries, studentischen Projekten (Boyce et al., 2023), und preprint Archiven (z.B. psyarxiv) und vielen weiteren. Die Belohnungen Replikationsstudien zu veröffentlichen sind daher historisch gesehen eher gering.

Sollten Replikationen durchgeführt werden, so wären die überzeugendsten Gründe dafür idealistischer Natur und nicht von äußeren Anreizen bestimmt. Der Fortschritt wissenschaftlicher Erkenntnisse durch die Anwendung strenger methodischer Standards erscheint in diesem System sinnvoller als das Streben nach hohen Zitationszahlen. Daher wäre es überraschend, wenn Forschende, die Replikationen durchführen, fragwürdige Forschungspraktiken anwenden würden, um ihre Ergebnisse zu manipulieren.

Doch spätestens seit der Veröffentlichung der Open Science Collaboration in 2015 kam es zu erheblichen Transformationen. Diese Publikation, die in „Science“ veröffentlicht wurde, hat zum derzeitigen Zeitpunkt 9597 Zitationen auf Google Scholar (3. Oktober, 2024) erhalten und hat viele weitere groß angelegte Replikationsprojekte nach sich gezogen. Seitdem hat sich die Anzahl einzelner Replikationsstudien signifikant vervielfacht (Clarke et al., 2023) und es haben sich viele neue methodologische Richtlinien entwickelt (z.B, Brandt et al., 2014; Moreau & Wiebels, 2023;

Schauer & Hedges, 2021). Replikationsforschung hat zudem auch in der breiten Öffentlichkeit Beachtung gefunden, beispielsweise in Artikeln der *New York Times* (Carey, 2015, 2016). Es entstanden Open-Science-Zentren (z. B. Center for Open Science; Open Science Center LMU), spezielle Förderprogramme für Replikationsvorhaben wurden eingerichtet, und neue Karrierewege für Open-Science-Befürworter*innen bildeten sich heraus (z. B. Institute for Replication). Wichtige wissenschaftliche Fachzeitschriften fördern mittlerweile aktiv die Einreichung von Replikationsstudien (*In praise of replication studies and null results*, 2020; *Registered replication reports*, n.d.), und Replikationen werden auch in renommierten Zeitschriften veröffentlicht (z. B. Hagger et al., 2016; Rohrer et al., 2015). Zusammengefasst steht die Replikationsforschung seit 2015 vermehrt im Fokus der Aufmerksamkeit.

Dies wirft die Frage auf: Inwieweit verändert diese erhöhte Aufmerksamkeit die Belohnungsstrukturen für Replikationsforschung?

Gotcha Bias & Null-Hacking

Bisher fand der Publikationsbias in Replikationsstudien nur geringe Beachtung. Berinsky et al. (2021) gingen diesem Forschungsthema in einem Vignettenexperiment nach. Sie erstellten Vignetten, die sowohl Originalstudien als auch Replikationsstudien beschrieben, und variierten dabei, ob die Replikation die Originalstudie unterstützte oder infrage stellte. Nach der Sichtung der Ergebnisse wurden die Teilnehmenden gebeten zu beurteilen, ob sie die jeweilige Studie veröffentlichen würden. Die Stichprobe bestand aus Politikwissenschaftsprofessor*innen, die an Institutionen in den USA lehrten. Die Ergebnisse zeigten, dass Replikationen, die die Originalstudie infrage stellten, eine höhere Wahrscheinlichkeit hatten, veröffentlicht zu werden, als solche, die sie unterstützten. Diese Verzerrung bezeichneten die Autor*innen als „Gotcha-Bias“, der eine „Gotcha, your original study is wrong“-Haltung impliziert. Ioannidis und Trikalinos (2005) beschrieben ein ähnliches Phänomen und stellten fest, dass extrem gegenteilige Ergebnisse häufiger veröffentlicht werden, was sie als „Proteus-Phänomen“ bezeichneten.

In der Vergangenheit erwies es sich als schwierig, Replikationsstudien zu veröffentlichen, während heutzutage gescheiterte Replikationen in hochrangigen Fachzeitschriften erscheinen (z. B.

Camerer et al., 2018; Gerber et al., 2016; Open Science Collaboration, 2015; Ranehill et al., 2015; Rohrer et al., 2015). Ist dies darauf zurückzuführen, dass Replikationen inzwischen „angesagt“ sind, oder darauf, dass gescheiterte Replikationen Aufmerksamkeit erzeugen? Wahrscheinlich auf beides. Berinsky et al. (2021) erklären den Gotcha-Bias mit dem Bedürfnis nach Neuartigkeit und Sensationalismus. Diese Situation erinnert an falsch-positive Befunde aus Originalstudien: Je aufregender eine Studie, desto größer die Wahrscheinlichkeit, dass sie veröffentlicht wird. Es lässt sich vermuten, dass, wenn eine bedeutende und viel zitierte Originalstudie nicht repliziert werden kann, ein erheblicher Teil der ursprünglich auf die Originalstudie gerichteten Aufmerksamkeit auf die Replikation übergehen könnte.

Sowohl Publikationsbias als auch Gotcha-Bias priorisieren die „newsworthiness“ über wissenschaftliche Klarheit. Für den wissenschaftlichen Fortschritt sollten Replikationen mit positiven Ergebnissen, die die Originalstudien unterstützen, ebenso häufig veröffentlicht werden wie solche, die sie infrage stellen.

Wenn Replikationen, die die Originalstudien infrage stellen, häufiger veröffentlicht werden und Forscher*innen auf Publikationen angewiesen sind, schafft diese Situation erneut einen Anreiz zur Manipulation. Ergänzend zum P-Hacking wird dieses Phänomen als Null-Hacking bezeichnet (Bryan et al., 2019).

Genauso wie Originalforscher*innen haben auch Replikationsforscher*innen Freiheitsgrade beim Planen, Durchführen und Auswerten von Replikationen. Bryan et al. (2019, S. 1) argumentieren, dass die Freiheitsgrade von Replikationsforscher*innen „make it far too easy to obtain and publish false negative replication results, even while appearing to adhere to strict methodological standards“. Zum Beispiel müssen Forschende lediglich eine Reihe hoch korrelierter Kovariaten hinzufügen, wodurch das Ergebnis nicht mehr signifikant ist. In ihrer Studie widerlegen sie eine Replikationsstudie (Gerber et al., 2018), die im Vergleich zur Originalstudie (Bryan et al., 2011) eine hohe Anzahl von Kovariaten hinzufügte, den Zeitraum der Datenerhebung variierte und Interaktionseffekte einbezog, was zu einem nicht signifikanten Ergebnis führte. Bei der erneuten Analyse der Daten der Replikationsstudie, bei der die Kovariaten entfernt und der Zeitraum

angepasst wurde, stellte sich heraus, dass die Daten der Replikationsstudie, genauso wie die der Originalstudie, signifikante Ergebnisse lieferten. Interessanterweise hielt sich die Replikationsstudie an gängige Standards für Replikationsstudien: 1. Die Stichprobengröße der Replikation war signifikant größer als die der Originalstudie. 2. Die Replikation führte Robustheitstests durch. 3. Die Replikation verwendete nahezu identische experimentelle Materialien wie die Originalstudie. Dennoch scheinen diese Standards unzureichend zu sein, um Null-Hacking zu verhindern.

Open Data erleichtert ebenfalls Null-Hacking, da Null-Hacker die Daten der Originalstudie untersuchen und zahlreiche Modelle ausprobieren können, bis die Ergebnisse nicht mehr signifikant sind (Protzko, 2018). Sie können dann dieses nicht-signifikante Modell hervorheben. Angesichts der kleinen Effektgrößen in der Psychologie (Szucs & Ioannidis, 2017) ist es wahrscheinlich, dass viele Modelle nicht signifikante Ergebnisse zeigen. Die Frage ist, ob man der Replikation vertrauen sollte, die viele Modelle mit einem Nullbefund zeigt, oder der Originalstudie, die ein einzelnes Modell präsentiert, das einen Effekt zeigt.

Ökonomische Interessen können ebenfalls Null-Hacking hervorrufen. Ein prominentes Beispiel ist die Tabakindustrie, die über Jahrzehnte hinweg Wissenschaftler*innen engagierte, um wissenschaftliche Unsicherheiten über die negativen Auswirkungen des Rauchens zu erzeugen. Wiederholte Reanalysen stellten den wissenschaftlichen Konsens infrage und verzögerten politische Maßnahmen gegen das Rauchen (Michaels & Monforton, 2005).

Die Auswirkungen von Null-Hacking sind gravierend und betreffen nicht nur die öffentliche Gesundheit, sondern auch die Wissenschaft. Replikationsforschung fördert wissenschaftliche Klarheit und Fortschritt, während null-gehackte Replikationsliteratur, obwohl sie ähnlich erscheint, in Wirklichkeit Forschungsergebnisse untergräbt. Falsch-positive Originalliteratur wird durch falsch-negative Replikationsliteratur ergänzt.

Das Ausmaß des Null-Hackings ist unklar, da dieses Phänomen erst kürzlich in den wissenschaftlichen Diskurs aufgenommen wurde und sich nur schwer untersuchen lässt. Dennoch zeigen Bryan et al. (2019) in einem realen Beispiel, dass Null-Hacking tatsächlich vorkommt.

P-Hacking in Replikationsstudien

Es gibt keine vorhandene Literatur, die sich mit P-Hacking in Replikationsstudien befasst. Dennoch gibt es mehrere mögliche Motivationen, warum Forschende Replikationen P-Hacken. Forscher*innen könnten Replikationsstudien nutzen, um ihre ursprünglichen Ergebnisse zu untermauern. Dies würde ihre Originalstudien vertrauenswürdiger erscheinen lassen und andere dazu ermutigen, auf ihrer Arbeit aufzubauen. Forschende, die P-Hacken, könnten dadurch höhere Zitationsraten erzielen.

Zusätzlich könnten Abhängigkeiten zwischen Forschenden Anreize zur Manipulation von Replikationsstudien schaffen. Wenn beispielsweise eine Laborleiterin ihren Doktoranden bittet, ihre Forschung zu replizieren, könnte die bestehende Abhängigkeit dazu führen, dass der Doktorand durch die Nutzung seiner Freiheitsgrade eine erfolgreiche Replikation begünstigt. Die prekären Arbeitsbedingungen in der Wissenschaft verstärken dabei die Abhängigkeiten. Zudem könnte der Ruf der Institution durch erfolgreiche Replikationen gefestigt werden.

Es wäre daher interessant zu untersuchen, ob eine Netzwerkanalyse aufzeigen könnte, ob die Erfolgsrate von Replikationen innerhalb eines Forscher*innen-Netzwerks höher ist als zwischen verschiedenen Netzwerken. Forscher*innen innerhalb eines Netzwerks könnten Anreize haben, die Forschung ihres Netzwerks durch Replikationsstudien zu unterstützen, wodurch das spezifische Forschungsfeld robuster erscheinen würde.

Sowohl im Fall von P-Hacking als auch bei Null-Hacking von Replikationen nutzen Replikationsforscher*innen ihre Freiheitsgrade, um den Erfolg der Replikation zu manipulieren.

Präregistrierung bei Replikationsstudien

Ähnlich wie bei Originalstudien, bei denen die Präregistrierung darauf abzielt, die Freiheitsgrade zu reduzieren, sollte die Präregistrierung von Replikationsstudien ebenfalls darauf abzielen, die Freiheitsgrade der Replikationsforscher*innen zu verringern. Die Freiheitsgrade von Replikationsforscher*innen ermöglichen P-Hacking oder Null-Hacking von Replikationsergebnissen. Durch die Minimierung der Freiheitsgrade durch die Präregistrierung sollte theoretisch weniger P-Hacking und Null-Hacking auftreten. Freiheitsgrade verschleiern im

Wesentlichen die Ergebnisse von Replikationen, und die Präregistrierung sollte diese Freiheitsgrade einschränken.

Die vorliegende Untersuchung

Das Hauptziel dieser Studie besteht darin, das mögliche Auftreten von Hacking in Replikationsstudien zu untersuchen, wobei sowohl Null-Hacking als auch P-Hacking berücksichtigt werden. Die Analysen basieren auf der umfangreichsten verfügbaren Replikationsdatenbank, FReD.

No-Hacking Hypothese

Traditionell sah sich die Veröffentlichung von Replikationsstudien erheblichen Herausforderungen gegenüber. Die akademischen Belohnungen für Replikationsforschung waren gering, sodass es wahrscheinlich ist, dass Replikationsstudien mit dem Ziel der Weiterentwicklung der Forschung durchgeführt wurden. Unter diesen Umständen ist die Manipulation von Replikationen fragwürdig. Replikationsforscher*innen würden daher nicht ihre Freiheitsgrade nutzen, um ihre Ergebnisse zu manipulieren. Folglich sollte der Status der Präregistrierung der Replikationsstudie die Beziehung zwischen dem originalen Effekt und dem Replikationseffekt nicht moderieren, da die Freiheitsgrade nicht ausgenutzt werden.

Die No-Hacking-Hypothese besagt, dass der Präregistrierungsstatus einer Replikation keinen Einfluss auf die Beziehung zwischen dem originalen Effekt und dem Replikationseffekt hat.

Hacking Hypothese

Ob Forschende die Replikationen durchführen im Vergleich zu Forschenden, die Originalstudien durchführen, methodisch strengere Praktiken befolgen, bleibt bisher unerforscht. Dennoch ist es bemerkenswert, dass lediglich etwa 20 % der 2348 Replikationseffekte in der FReD Replikationsdatenbank mit einer Präregistrierung verknüpft sind. Die Methode Präregistrierung hat sich bereits in den frühen Tagen der aktuellen Replikationskrise als eine der bedeutendsten Gegenmaßnahmen herauskristallisiert. Daher sollten Replikationsforscher*innen über die Wichtigkeit der Präregistrierung gut informiert sein. Die relativ niedrige Präregistrierungsrate deutet somit nicht auf rigoroses methodisches Arbeiten mit dem Ziel des wissenschaftlichen Fortschritts hin. Darüber hinaus war die Verbreitung von fragwürdigen Forschungspraktiken in

Originalstudien erheblich hoch (Singh Chawla, 2021). Warum sollte dies bei Replikationen anders sein?

Replikationen sind momentan im Fokus der Aufmerksamkeit. Gescheiterte Replikationen werden in hochrangigen Fachzeitschriften veröffentlicht. Null-Hacking ist relativ einfach und bereits in der Literatur dokumentiert. Das Replizieren eigener Arbeiten oder der Arbeiten von Kolleg*innen trägt zur Wahrnehmung der Robustheit der Forschung bei und erhöht damit die Wahrscheinlichkeit von Zitationen. Sowohl die Anzahl von Publikationen als auch die Anzahl von Zitationen sind wichtige Faktoren für den akademischen Aufstieg. Dies deutet darauf hin, dass es durchaus Anreize gibt, Replikationsstudien zu manipulieren.

Wenn Replikationsforscher*innen ihre Forschung manipulieren wollen, könnten sie ihre Freiheitsgrade nutzen, um die Ergebnisse der Replikation zu verzerren. Eine Präregistrierung würde dazu beitragen, dies einzuschränken. Daher sollte, wenn Hacking in Replikationsstudien vorkommt, der Präregistrierungstatus der Replikationsstudie die Beziehung zwischen dem Originaleffekt und dem Replikationseffekt beeinflussen.

Hacking-Hypothese: Der Präregistrierungsstatus der Replikationsstudie moderiert den Zusammenhang zwischen dem Originaleffekt und dem Replikationseffekt.

Ob es sich vermehrt um Null-Hacking oder P-Hacking handelt, bleibt unklar. Diese Unklarheit erschwert es, die Richtung der Effekte vorherzusagen. Es werden zwei Hypothesen vorgestellt: die Null-Hacking-Hypothese und die P-Hacking-Hypothese, welche beide auf plausiblen Argumenten basieren, jedoch aufgrund der begrenzten verfügbaren Literatur nicht ausreichend untersucht sind.

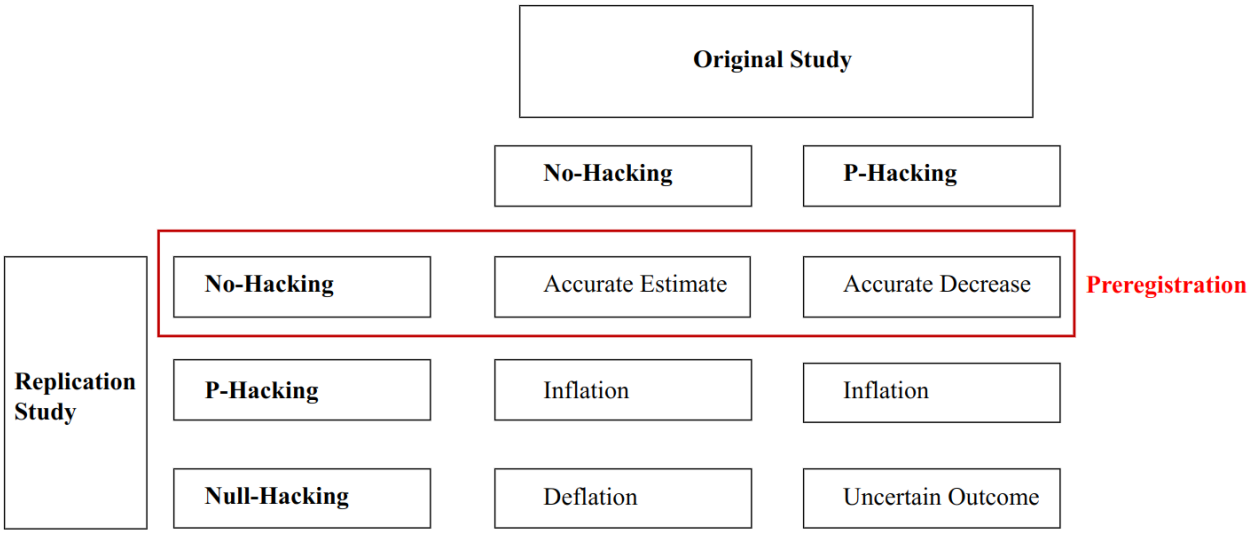
Null-Hacking-Hypothese: Der Präregistrierungsstatus der Replikation moderiert den Zusammenhang zwischen dem Originaleffekt und dem Replikationseffekt, wobei *präregistrierte* Replikationsstudien einen höheren Zusammenhang aufzeigen.

P-Hacking-Hypothese: Der Präregistrierungsstatus der Replikation moderiert den Zusammenhang zwischen dem Originaleffekt und dem Replikationseffekt, wobei *nicht-präregistrierte* Replikationsstudien einen höheren Zusammenhang aufzeigen.

In Abbildung 1 sind die unterschiedlichen Auswirkungen von Hacking in Originalstudien und Hacking in Replikationsstudien im Hinblick auf das Replikationsergebnis graphisch veranschaulicht.

Abbildung 1

Einfluss von Hacking auf Ergebnisse in Paarungen von Original- und Replikationsstudien



Hinweis: Diese Grafik veranschaulicht die Auswirkungen von Hacking auf die Ergebnisse von Replikationsstudien im Hinblick auf Ergebnisse der Originalstudien. Es wird zwischen Hacking in Originalstudien und in Replikationsstudien unterschieden. Das rote Feld zeigt Bereiche an, in denen eine Präregistrierung Hacking beeinflusst.

Methode

Um Hacking in Replikationsstudien zu untersuchen, wurde eine präregistrierte hierarchische Analyse anhand der Replikationsdatenbank FReD durchgeführt. Alle Abweichungen von der Präregistrierung wurden hervorgehoben und sind im Diskussionsteil diskutiert. Nicht-präregistrierte Analysen werden als explorative Analysen gekennzeichnet.

Datenbank FReD

Zur Analyse wurde die FORRT Replication Database verwendet. FReD ist ein gemeinschaftsbasiertes, öffentlich zugängliches Projekt, das darauf abzielt, die Replizierbarkeit von Studien systematisch zu dokumentieren und sie leicht auffindbar und zugänglich zu machen. Die Datenbank besteht aus groß angelegten Replikationsstudien (z. B. Klein et al., 2014; Open Science

Collaboration, 2015; Soto, 2019) sowie aus einer hohen Anzahl individueller Replikationsstudien (192 Studien bis zum 7. Oktober 2024). Unveröffentlichte Replikationen sind ebenfalls enthalten, um den Publikationsbias entgegenzuwirken. Die Datenbank ist eine lebendige Ressource, was bedeutet, dass sie regelmäßig aktualisiert und erweitert wird. Zum Beispiel können Forschende über ein Formular Replikationen eintragen, die sie durchgeführt haben oder ihnen bekannt sind. Um die Qualitätssicherung zu gewährleisten, werden diese Einträge, wie alle anderen Einträge in der Datenbank auch, manuell von Forschenden des Projekts validiert.

Das Ziel ist es, eine möglichst umfassende Datenbank zu erhalten, daher sind die Inklusionskriterien eher liberal gestaltet. So können alle Studien aus den Sozialwissenschaften und aus der Medizin aufgenommen werden. Zurzeit ist der größte Teil der Studien dennoch aus der Psychologie. Darüber hinaus wurde eine relativ breite Definition von Replikationen genutzt: Jede Studie, die die gleiche Hypothese wie die Originalstudie testet, wird als Replikation betrachtet (Hüffmeier et al., 2016). Es gibt in der Datenbank zusätzlich die Möglichkeit die Ähnlichkeit der Replikation zur Originalstudie einzuschätzen.

Die Datenbank ist so strukturiert, dass jede Zeile einem Replikationseffekt entspricht. Die Variablen „ref_original“ und „ref_replication“ zeigen die Zuordnung zu bestimmten Artikeln an. Diese Struktur ermöglicht die Darstellung komplexer Kombinationen von Original- und Replikationsstudien, zum Beispiel eine Replikation von zwei verschiedenen Originalstudien.

FReD wurde für diese Studie ausgewählt, da es sich derzeit um die umfassendste Replikationsdatenbank handelt und somit die höchstmögliche statistische Power für die Analysen bietet. Alle Einträge in der Datenbank wurden bis zur Version vom 9. Juli 2024 einbezogen (https://osf.io/jvrd8). Es ist zu beachten, dass weitere Einträge sowie mögliche Korrekturen und Aktualisierungen, die nach diesem Datum vorgenommen wurden, nicht in die Analyse einfließen.

Analyseplan

Exklusionen

Einträge mit Werten in der Variable „exclusion“ wurden ausgeschlossen. Zudem wurden Einträge entfernt, bei denen die Variable „notes_validation“ Begriffe wie „duplicate“ oder „No

actual replication conducted“ enthielt. Replikationsstudien, die vor 2011 durchgeführt wurden, wurden ausgeschlossen, da vor diesem Jahr Präregistrierungen unbekannt waren. Nur Einträge mit einem Validierungsstatus von 1 (validated with everything correct) oder 2 (validated with errors highlighted, corrected, and noted in notes_validation) wurden einbezogen. Darüber hinaus wurden Einträge mit fehlenden Werten für die Pflichtvariablen ausgeschlossen, sodass nur Einträge mit dem Status == 2 (exhaustive entry with all required outcomes coded) berücksichtigt wurden. Schließlich schloss die Analysemethode automatisch Einträge aus, die keine Werte für die relevanten Variablen aufwiesen.

Berechnung von Werten

Alle möglichen Effektstärken der Originalstudien und der Replikationsstudien wurden in Bravais-Pearson-Korrelationen umgewandelt. Cramer’s V, Bayes Factors, Hazard Ratios, Cohen’s q und Risk Ratios konnten nicht konvertiert werden. Die Effektstärken der Originalstudien wurden als positiv kodiert, während die Replikationseffekte nur dann als positiv kodiert wurden, wenn sie mit der Richtung der originalen Effekte übereinstimmten. Eine Funktion wurde erstellt, um den Präregistrierungsstatus der Replikationsstudie zu kodieren, wobei Präregistrierungslinks in binäre Werte umgewandelt wurden. Dieser Prozess wurde manuell überprüft. Darüber hinaus wurde eine zufällige Stichprobenkontrolle durchgeführt, um potenzielle Einträge zu identifizieren, die tatsächlich präregistriert waren, jedoch nicht als solche erfasst wurden.

Die Hypothesen bezüglich Hacking, Null-Hacking und P-Hacking wurden mithilfe einer Mehrebenenanalyse untersucht, bei der die Effektstärken der Replikationen innerhalb der Replikationsartikel hierarchisch strukturiert wurden. Das Modell ermittelte, wie die Effektstärke der Originalstudie und der Präregistrierungsstatus der Replikationsstudie die Effektstärke der Replikation vorhersagten. Daher war die Analyse als Moderationsanalyse strukturiert, um zu bestimmen, ob der Präregistrierungsstatus die Beziehung zwischen der Effektgröße der Originalstudie und der Effektgröße der Replikation signifikant moderiert. Dieser Ansatz schien vorteilhaft, da keine Entscheidung über ein spezifisches Kriterium für den Replikationserfolg erforderlich war. Explorativ wurde dieses Verfahren auch auf nicht-validierte Einträge angewendet.

Darüber hinaus wurde eine zusätzliche explorative Analyse durchgeführt, um zu untersuchen, ob die Präregistrierung von Replikationsstudien den Replikationserfolg beeinflusst. Hierbei wurde ein spezifisches Kriterium für den Replikationserfolg verwendet. Eine Replikation wurde als erfolgreich angesehen, wenn sowohl der originale Effekt als auch der Replikationseffekt statistisch signifikant waren und in die gleiche Richtung zeigten. Dieses Kriterium wurde aufgrund seiner weitverbreiteten Verwendung in der Literatur gewählt.

Da Forscher*innen, die Replikationen einreichen, im Einreichungsformular angeben konnten, ob das Ergebnis als „success“, „informative failure to replicate“, „inconclusive“ oder „practical failure to replicate“ klassifiziert wurde, standen einige Einträge für die Variable „result“ vor der Analyse zur Verfügung. Da die Ergebnisse von Replikationen in einigen Studien nicht immer leicht berechnet werden können, wurden diese bereitgestellten Ergebnisse herangezogen; jedoch wurden alle berechenbaren Ergebnisse auch tatsächlich berechnet. Die Signifikanz der Originaleffekte und der Replikationseffekte wurde mithilfe von Konfidenzintervallen ermittelt. Wenn die Signifikanz für entweder die Originalstudie oder die Replikationsstudie nicht berechnet werden konnte (z. B. aufgrund fehlender Daten oder nicht konvertierbarer Effektgrößen), wurde der Eintrag nicht berechnet und der vorhandene Wert in der Variablen „result“ verwendet. Einträge, bei denen sowohl die Originaleffekte als auch die Replikationseffekte Signifikanzwerte hatten, wurden auf Übereinstimmungen in den Signifikanzmustern überprüft. Bei einem Widerspruch zwischen dem berechneten Ergebnis und dem vorher angegeben Ergebnis, wurde das berechnete Ergebnis gewählt. Für Einträge ohne Wert in der Variable „result“ wurde ebenfalls versucht, ein Ergebnis zu berechnen. Nur die Variablen „success“ und „informative failure to replicate“ wurden in der finalen Analyse berücksichtigt, während „practical failure to replicate“ und „inconclusive“ ausgeschlossen wurden. Ein generalisiertes lineares gemischtes Modell mit Zufallseffekten wurde verwendet, um den Einfluss des Präregistrierungsstatus auf den Replikationserfolg zu bestimmen.

Neben der Untersuchung des Effekts der dichotomen Präregistrierungsvariable wurde ein genauerer Blick darauf geworfen, ob verschiedene Arten von Präregistrierungsvorlagen die Beziehung zwischen der Originaleffektgröße und der Replikationseffektgröße beeinflussen, da

einige Vorlagen detaillierter oder umfassender sind als andere. Dies wurde unter Verwendung eines generalisierten linearen gemischten Modells mit Zufallseffekten durchgeführt. Der Vorlagentyp wurde ermittelt, indem ein API-Link aus den OSF-Links erstellt und dann die Vorlageninformationen aus der Variablen „registration_supplement“ extrahiert wurden. Einige Links wurden korrigiert, wenn sie fälschlicherweise zum OSF-Ordner anstelle der Präregistrierungsseite führten. Links von AsPredicted wurden direkt als „Preregistration Template from AsPredicted.org“ kategorisiert.

Eine explorative Analyse untersuchte außerdem, ob die Nähe der Replikationsstudie zur Originalstudie die Beziehung zwischen der Originaleffektstärke und der Replikationseffektstärke moderiert. Die Variable Closeness war bereits in der Datenbank kodiert. Für diese Analyse wurden die Werte von 3, 2, 1 (different, close, exact) in 3, 2, 1 (exact, close, different) umkodiert. Ein lineares gemischtes Modell mit Zufallseffekten wurde für diese Moderationsanalyse verwendet.

Die Berechnung von Werten, Visualisierung, Hypothesentests und explorativen Analysen wurde mit GNU R Version 4.4.1 durchgeführt. Die folgenden Pakete wurden verwendet: openxlsx (Schauberger & Walker, 2021), lmerTest (Kuznetsova et al., 2017), ggplot2 (Wickham, 2016), ggExtra (Attali & Baker, 2023), pwr (Champely et al., 2018), knitr (Xie, 2015), kableExtra (Zhu, 2021), dplyr (Wickham et al., 2018), httr (Wickham & Bryan, 2019), jsonlite (Ooms, 2014), lme4 (Bates et al., 2015), viridis (Garnier, 2018) und broom.mixed (Carter et al., 2020). Das R-Skript wurde präregistriert (https://osf.io/3d4yk). Das finale Analyseskript ist online verfügbar (https://osf.io/6sq72), und alle Änderungen sind hervorgehoben.

Ergebnisse

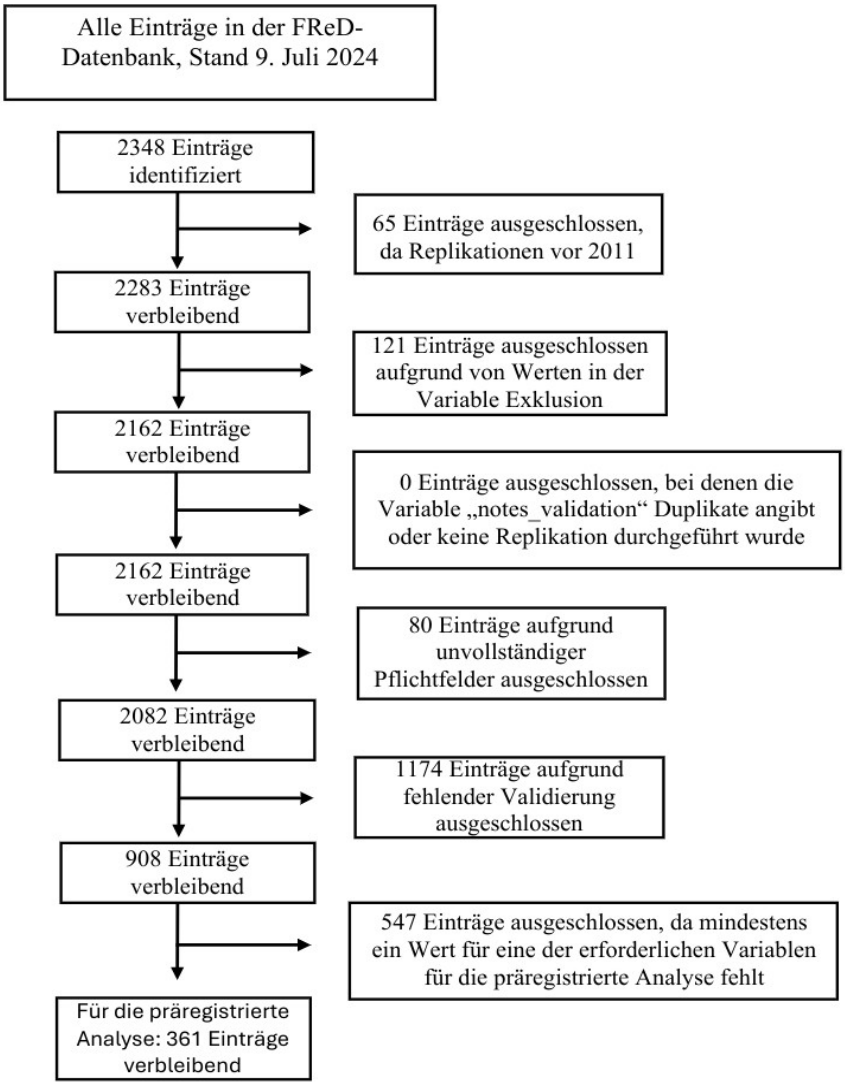
Überblick des Datensatzes

Nach den Exklusionen (siehe Abbildung 2) blieben insgesamt 908 Einträge übrig. Davon berichteten 376 Einträge (41,41 %) über eine originale Effektgröße. Der Mittelwert dieser originalen Effektgrößen betrug r = .33 (SD = .199). 363 Einträge (39,98 %) berichteten über eine Replikationseffektgröße. Der Mittelwert der Replikationseffektstärken betrug r = .16 (SD = .218). Die in den Replikationsstudien beobachteten Effektstärken waren signifikant kleiner als die in den

Originalstudien, T(1) = 10.738, p < .001, Δr = -0.164, 95% CI [0.134, 0.195], k = 363. Es ist wichtig zu beachten, dass diese Analyse die Mehrebenenstruktur der Daten nicht berücksichtigt hat. Weitere Einzelheiten zu den deskriptiven Statistiken sowohl für die originalen als auch für die Replikationseffektstärken sind in Tabelle 1 zu finden.

Abbildung 2

Exklusionen in der FReD Datenbank



Hinweis: Das Flussdiagramm zeigt die Reduktion der Einträge der FReD-Datenbank von 2348 auf 908 Einträge durch eine Reihe von Exklusionen. Für die präregistrierte Analyse bleiben 361 Replikationseffekte übrig.

Tabelle 1: Deskriptive Statistiken für Originaleffektstärken und Replikationseffektstärken

	Deskriptive Statistiken						
Variable	**N**	**Fehlend**	**Mittelwert**	**Median**	**SD**	**Min**	**Max**
Original Effect Size	376	532	0.329	0.274	0.199	0.005	0.961
Replication Effect Size	363	545	0.165	0.090	0.218	-0.450	0.990

Hinweis: Die Tabelle fasst die Anzahl der Beobachtungen (N), die Anzahl fehlender Werte (Fehlend), den Mittelwert, Median, die Standardabweichung (SD), sowie Minimal- und Maximalwerte für die Originaleffektstärken und die Replikationseffektstärken zusammen. Alle Werte sind als Bravais-Pearson's r angegeben.

Hypothesentests

Nach dem zusätzlichen Ausschluss der Zeilen mit mindestens einem fehlenden Wert für eine der erforderlichen Variablen für die präregistrierte Analyse (originale Effektstärke, Replikationseffektstärke, Präregistrierungsstatus der Replikation oder Zugehörigkeit zum Replikationsartikel) verblieben 237 nicht-präregistrierte Replikationseffektstärken (26,10%) und 124 präregistrierte Replikationseffektstärken (13,66%) (siehe Tabelle 2). Dies führte zu 216 unabhängigen Artikeln für Originaleffektstärken und 72 unabhängigen Artikeln für Replikationseffektstärken in der präregistrierte Analyse (siehe Tabelle 3).

Tabelle 2: Verteilung der präregistrierten Replikationseffekte in den präregistrierten Analysen

	Verteilung der Präregistrierungen
Kategorie	**Anzahl**
Präregistrierte Replikationseffekte	124
Nicht-präregistrierte Replikationseffekte	237

Hinweis: Die Tabelle zeigt die Verteilung des Präregistrierungsstatus bei Einträgen, bei denen alle notwendigen Variablen für die präregistrierten Analysen (Originaleffektstärke, Replikationseffektstärke, Präregistrierungsstatus der Replikation und Zugehörigkeit zum Replikationsartikel) ohne fehlende Werte vorlagen.

Tabelle 3: Anzahl der unabhängigen Artikel in den präregistrierten Analysen

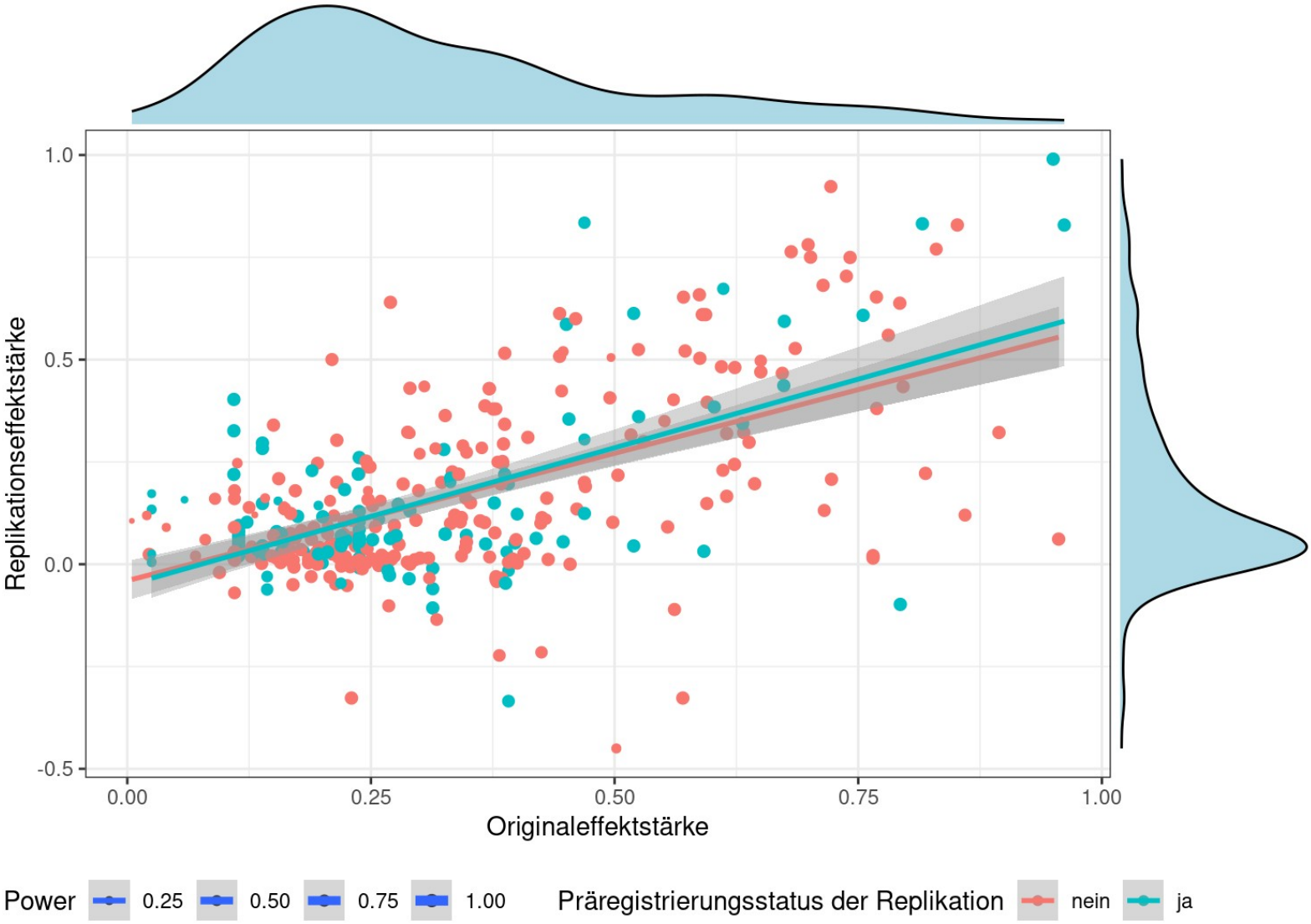
	Unabhängige Artikel
Artikeltyp	**Anzahl**
Unabhängige Originalartikel	216
Unabhängige Replikationsartikel	72

Hinweis: Die Tabelle zeigt die Anzahl der unabhängigen Artikel bei Einträgen, bei denen alle notwendigen Variablen für die präregistrierten Analysen (Originaleffektgröße, Replikationseffektgröße, Präregistrierungsstatus der Replikation und Zugehörigkeit zum Replikationsartikel) ohne fehlende Werte vorlagen.

Insgesamt gingen daher 361 von 908 Einträgen (39.76%) in die Analyse ein (siehe Abbildung 3). Das lineare gemischte Mehrebenenmodell zeigte einen signifikanten Zusammenhang zwischen den Originaleffekten und den Replikationseffektstärken, $r = 0{,}618$, $SE = 0{,}059$, $t = 10{,}45$, $p < .001$. Der Präregistrierungsstatus der Replikation hatte keinen signifikanten Einfluss auf die Replikationseffektstärken, $r = -0{,}065$, $SE = 0{,}050$, $t = -1{,}30$, $p = .194$. Entgegen der Hypothese moderierte der Präregistrierungsstatus der Replikation nicht den Zusammenhang zwischen den Originaleffektstärken und den Replikationseffektstärken, $r = 0{,}153$, $SE = 0{,}118$, $t = 1{,}30$, $p = .196$.

Abbildung 3

Streudiagramm des Zusammenhangs zwischen Originaleffektstärken und Replikationseffektstärken, markiert nach statistischer Power und Präregistrierungsstatus der Replikationsstudie.



Hinweis: Dieses Streudiagramm zeigt den Zusammenhang zwischen Effektstärken in originalen Studien (x-Achse) und deren Replikationen (y-Achse). Die Datenpunkte wurden basierend auf dem Präregistrierungsstatus der Replikation eingefärbt (rot für nicht-präregistriert, grün für präregistriert), wobei die Punktgröße das Power-Niveau von 0,25 bis 1,00 anzeigt. Die Linie stellt den Trend mit einem 95%-Konfidenzintervall dar.

Exploratorische Analysen

Hinzunahme von nicht validierten Einträgen

In einer zweiten explorativen Analyse, die nicht validierte Einträge einschloss, wurden 1401 von 2082 Einträgen (67,29 %) analysiert, nachdem diejenigen mit fehlenden Werten für relevante Variablen (Originaleffektstärke, Replikationseffektstärke, Präregistrierungsstatus der Replikation oder Zugehörigkeit zum Replikationsartikel) ausgeschlossen wurden. Es ist zu beachten, dass dies die einzige Analyse in dieser Studie war, die nicht validierte Einträge einbezog. Das lineare gemischte Modell, das zufällige Effekte berücksichtigte, ermittelte einen signifikanten Zusammenhang zwischen den Originaleffektstärken und den Replikationseffektstärken, $r = 0.560$, $SE = 0.034$, $t = 16.40$, $p < .001$. Der Präregistrierungsstatus der Replikation hatte keinen signifikanten Einfluss auf die Replikationseffektstärken, $r = -0.048$, $SE = 0.041$, $t = -1.19$, $p = .235$. Darüber hinaus moderierte der Präregistrierungsstatus der Replikation nicht den Zusammenhang zwischen den Originaleffektstärken und den Replikationseffektstärken, $r = 0.120$, $SE = 0.096$, $t = 1.25$, $p = .210$.

Bewertung des Replikationserfolgs basierend auf übereinstimmenden Signifikanzmustern

In der explorativen Analyse, die den Replikationserfolg als Übereinstimmung der Signifikanzmuster zwischen dem Originaleffekt und dem Replikationseffekt definierte, wurden 391 von 908 Einträgen (43,06 %) ohne fehlende Werte in den relevanten Variablen (Präregistrierungsstatus der Replikation, Ergebnis) einbezogen. Davon wurden 197 als erfolgreiche Replikationen klassifiziert, während 194 als fehlgeschlagen identifiziert wurden. Fünfzehn Einträge wurden als „inconclusive“ und drei Einträge als „practical failure to replicate“ klassifiziert. Diese Einträge wurden ausgeschlossen. Ein generalisiertes lineares gemischtes Modell mit einer binomialen Verteilung und zufälligen Effekten zeigte, dass der Präregistrierungsstatus der Replikation keinen signifikanten Einfluss auf den Replikationserfolg hatte, log-odds Koeffizient = -0,479, $SE = 0,2578$, $z = -1,858$, $p = .063$.

Unterschiedliche Präregistrierungsvorlagen

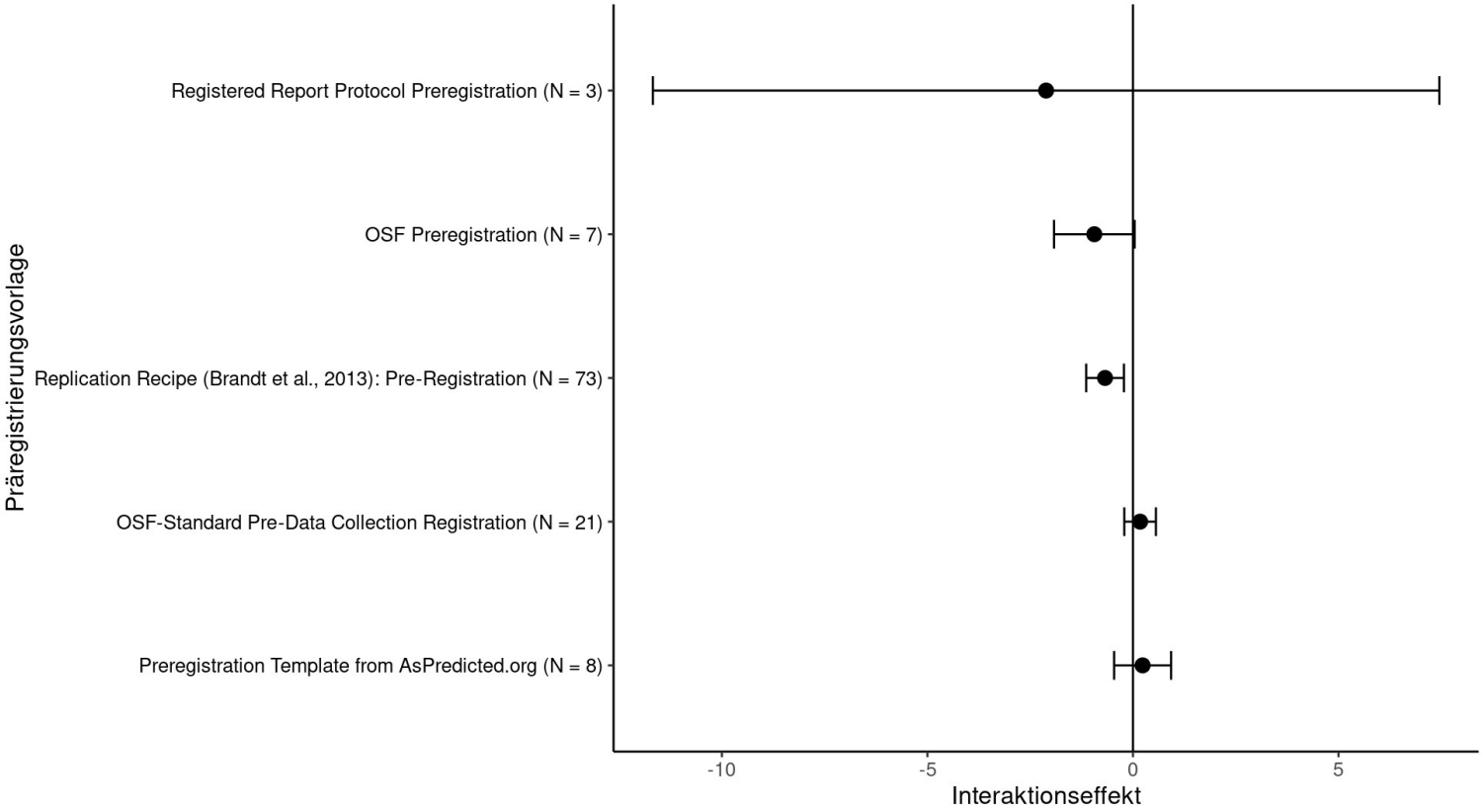
Nach der Extraktion der verschiedenen Präregistrierunsvorlagen aus den API-Links wurden unterschiedliche Vorlagentypen analysiert. Das lineare gemischte Modell mit zufälligen Effekten untersuchte, ob die unterschiedlichen Präregistrierungsvorlagen den Zusammenhang zwischen den Originaleffektstärken und den Replikationseffektstärken moderierten (siehe Abbildung 4). Insgesamt wurden 116 von 908 Einträgen (13 %), für die Werte zur Präregistrierungsvorlage verfügbar waren, in die Analyse einbezogen.

Das Registered Report Protocol Preregistration (N = 3) zeigte keinen Interaktionseffekt auf den Zusammenhang zwischen den Originaleffektstärken und den Replikationseffektstärken, r = -2.113, SE = 4.880, t = -0.433, p = .667. Die Vorlage OSF Preregistration (N = 7) wies ebenfalls keinen signifikanten Interaktionseffekt auf, r = -0.940, SE = 0.500, t = -1.878, p = .065. Im Gegensatz dazu zeigte das Replication Recipe (Brandt et al., 2013): Pre-Registration (N = 73) einen signifikanten negativen Interaktionseffekt, r = -0.679, SE = 0.234, t = -2.897, p = .005. Das OSF-Standard Pre-Data Collection Registration (N = 21) zeigte keinen signifikanten Interaktionseffekt, r = 0.174, SE = 0.196, t = 0.887, p = .377. Ähnlich wies auch die Präregistrierungsvorlage von AsPredicted.org (N = 8) keinen signifikanten Interaktionseffekt auf, r = 0.235, SE = 0.354, t = 0.664, p = .509. Interaktionseffekte konnten für die Präregistrierungsvorlagen Prereg Challenge und Open-Ended Registration nicht berechnet werden. Drei Einträge wurden herausgefiltert, da sie nur die Vorlage Replication Recipe (Brandt et al., 2013): Post-Completion enthielten aber keine entsprechende Präregistrierungsvorlage Replication Recipe (Brandt et al., 2013): Pre-Registration aufwiesen.

Zusammenfassend wurde nur für die Präregistrierungsvorlage Replication Recipe (Brandt et al., 2013): Pre-Registration ein signifikanter (negativer) Interaktionseffekt gefunden.

Abbildung 4

Forest-Plot der Schätzungen der Interaktionseffekte über verschiedene Präregistrierungsvorlagen für den Zusammenhang zwischen Originaleffektstärke und Replikationseffektstärke.



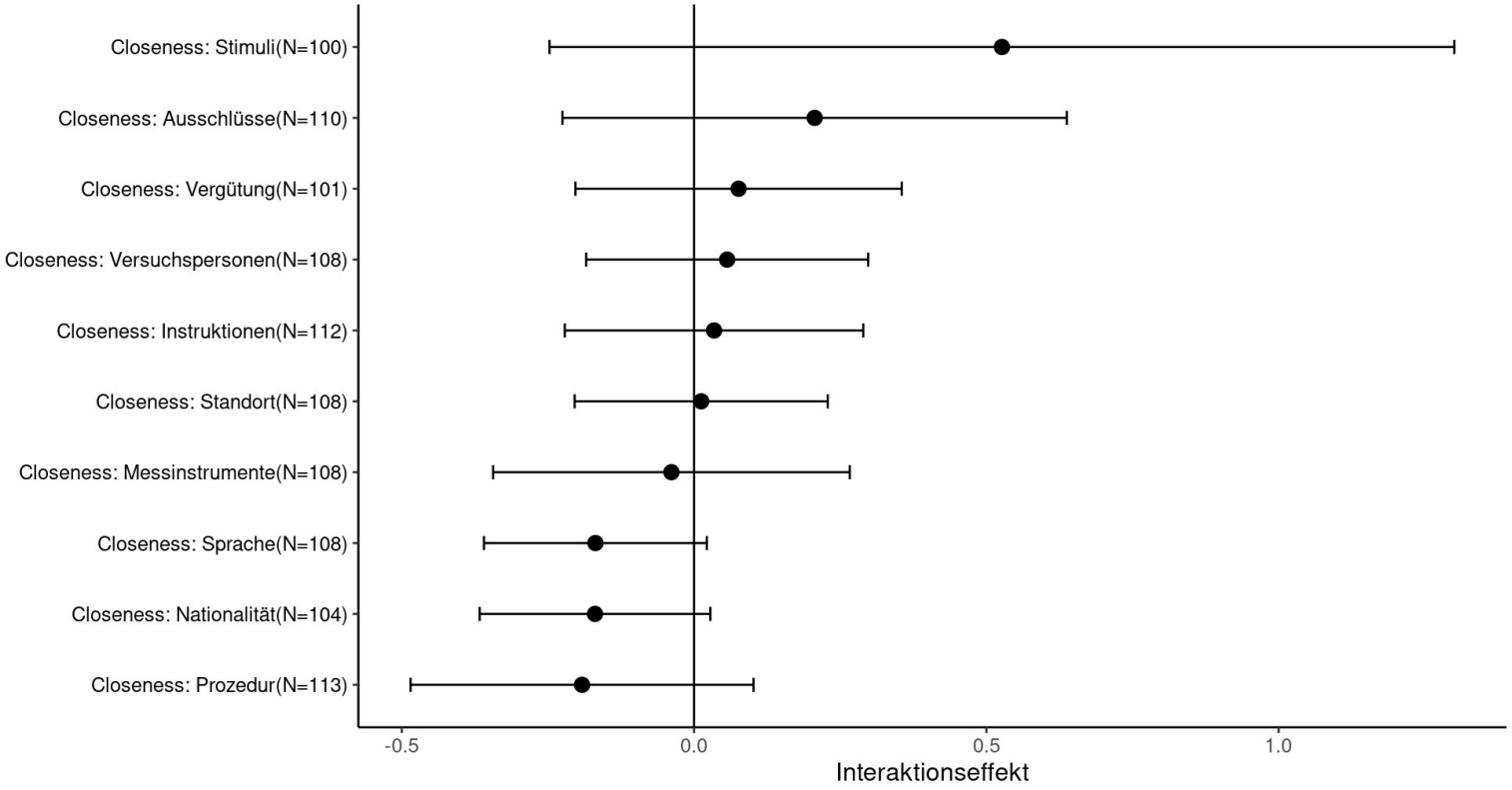
Hinweis: Schätzungen der Interaktionseffekte mit 95% Konfidenzintervallen für jede Art von Präregistrierungsvorlage. Die Anzahl der Replikationseffekte (N) pro Vorlage ist in Klammern angegeben.

Closeness

Um den potenziellen moderierenden Effekt der zehn verschiedenen Closeness-Variablen auf den Zusammenhang zwischen Originaleffektstärke und Replikationseffektstärke zu untersuchen, wurde ein linear gemischtes Modell mit Zufallseffekten verwendet. Es wurden keine signifikanten Interaktionseffekte für die zehn Closeness-Variablen gefunden.

Abbildung 5

Interaktionseffekt-Schätzungen für verschiedene Closeness-Dimensionen im Zusammenhang zwischen Originaleffektstärke und Replikationseffektstärke.



Hinweis: Schätzungen der Interaktionseffekte mit 95% Konfidenzintervallen für verschiedene Closeness-Dimensionen. Die Anzahl der Replikationseffekte (N) pro Dimension ist in Klammern angegeben.

Diskussion

Um zu untersuchen, ob Replikationsstudien Anzeichen auf Hacking aufzeigen, wurde für diese Studie die bisher umfassendste Replikationsdatenbank FReD analysiert. In die präregistrierte Analyse sind nach allen Exklusionen 361 Replikationseffekte mit zugehörigen Originaleffekten eingegangen, von denen 124 Replikationseffekte präregistriert waren. Das Ergebnis dieser Analyse ergab, dass der Präregistrierungstatus der Replikationsstudie den Zusammenhang zwischen Originaleffekt und Replikationseffekt nicht signifikant moderierte. Die Hypothese, dass es in Replikationsstudien zu Hacking kommt, konnte daher nicht untermauert werden.

Angesichts der 2348 Replikationseffekte in der Datenbank stellt sich die Frage, warum so viele davon ausgeschlossen wurden, wodurch die statistische Power der Analysen eingeschränkt wurde. Beim größten Anteil der ausgeschlossenen Effekte handelte es sich um nicht-validierte Einträge (N=1147). Dies könnte damit erklärt werden, dass die Datenbank noch relativ jung ist und der manuelle Validierungsprozess umfangreiche Ressourcen erfordert. Der zweitgrößte Anteil der ausgeschlossen Effekte (N=547) wurde deshalb automatisch von der präregistrierten Analyse ausgeschlossen, weil es mindestens einen fehlenden Wert in den relevanten Variablen (Originaleffektstärke, Replikationseffektstärke, Präregistrierungsstatus der Replikation oder zugehöriger Replikationsartikel) gab. Die Variablen „Präregistrierungsstatus der Replikation“ und „zugehöriger Replikationsartikel“ stellten dabei nicht das Problem dar; vielmehr traten viele fehlende Werte bei den Effektstärken auf. Dies ist nicht unbedingt ein Problem der Datenbank sondern eher der Studien selbst. So wurden zum Beispiel bei der Originalstudie der Ankereffekte von Jacowitz und Kahnemann (1995) weder Effektstärken noch Teststatistiken berichtet, sondern es wurde ein Ankerindex angegeben, der hohe und niedrige Anker miteinander vergleicht. Zu dieser Originalstudie sind 452 Replikationseffekte in der Datenbank eingetragen. Diese 452 Replikationseffekte konnten in der Analyse nicht berücksichtigt werden, weil keine Originaleffektstärke zu den Replikationseffektstärken vorlag.

Um dem großen Anteil von Exklusionen durch fehlende Validierung Rechnung zu tragen, wurde eine explorative Analyse durchgeführt, die nicht-validerte Einträge in der Berechnung berücksichtigt. Auch bei dieser Analyse, die insgesamt 1401 Replikationseffekte mit zugehörigen Originaleffekten einbezieht, wurde kein signifikanter Interaktionseffekt auf den Zusammenhang zwischen Originaleffektstärke und Replikationseffektstärke festgestellt.

Auch die weiterführende explorative Analyse, die den Replikationserfolg anhand eines übereinstimmenden Signifikanzmusters zwischen Originalstudie und Replikationsstudie misst, zeigte, dass der Präregistrierungsstatus der Replikation keinen signifikanten Einfluss auf den Replikationserfolg hat.

In dieser Studie wurde somit sowohl ein binäres Modell des Replikationserfolgs berücksichtigt, das sich auf die Wahrscheinlichkeit eines Replikationserfolgs als kategorisches Ergebnis konzentriert, als auch ein lineares Modell erstellt, das eine kontinuierliche Beziehung zwischen zwei Variablen untersucht. Es ist zu beachten, dass die gewählte Operationalisierung des Replikationserfolgs im binären Modell eine wesentliche Rolle spielt. Andere Operationalisierungen, wie etwa die Betrachtung des Verhältnisses von Replikationseffektstärke zur Originaleffektstärke (Soto, 2019), könnten zu unterschiedlichen Ergebnissen führen.

Zusammengenommen konnte die Hypothese, dass Replikationen Anzeichen auf Hacking aufzeigen, mit der vorliegenden Studie nicht untermauert werden.

Eine Erklärung für das Ergebnis könnte sein, dass die Belohnungen dafür, Replikationen zu manipulieren, von Replikationsforscher*innen als nicht so hoch angesehen werden, wie die damit verbundenen Kosten. Zu den Kosten könnten unter anderem ethische Zweifel und potenzielle Rufschädigungen zählen.

Dies wird bei folgendem Beispiel besonders klar ersichtlich: Replikationsforscher*innen, die absichtlich ihre Replikationen so manipulieren, dass die Forschungsarbeiten von Kolleg*innen nicht repliziert werden können, handeln auf eine Weise, die als äußerst unethisch gegenüber den Kolleg*innen, den wissenschaftlichen Prinzipien und dem Fortschritt des Forschungsfeldes betrachtet werden kann. Null-Hacker müssten auch antizipieren, dass Originalforscher*innen die Replikation ganz genau überprüfen werden und gegebenenfalls noch eine Reanalyse durchführen (Byran et al., 2019). Wenn Null-Hackende Replikationsforscher*innen dann beim Null-Hacken erwischt werden, könnte das potenziell sehr rufschädigend sein. Daher tragen null-hackende Replikationsforscher beim Null-Hacken ein hohes Risiko.

P-Hackende Replikationsforscher*innen könnten ebenfalls ethische Bedenken haben und ein Risiko einer Rufschädigung eingehen. Allerdings müssten sie möglicherweise weniger direkten Druck von den Originalforscher*innen befürchten, da diese wahrscheinlich eher erfreut über eine erfolgreiche Replikation sind oder weil die Replikationsforscher*innen oder Forschende aus dem eigenen Netzwerk selbst die Originalautor*innen sind.

Zudem könnte auch die Erklärung in Betracht gezogen werden, dass Replikationsforscher*innen aus idealistischen Beweggründen forschen. Die Replikationsforschung stellt eine Form der Qualitätskontrolle dar, und Replikationsforscher*innen könnten bestrebt sein, ihre wissenschaftlichen Freiheitsgrade nicht zum Nachteil der Wissenschaft auszunutzen, sondern vielmehr für wissenschaftliche Integrität einzutreten.

Auch Fachzeitschriften könnten Hacking in Replikationen vermindern. Denn Fachzeitschriften, die Replikationen publizieren, stellen an die Replikationen in der Regel hohe Anforderungen. Beispielsweise wird das zusätzliche Veröffentlichen von Analyseskript und Daten häufig erwartet (Center for Open Science, n.d.), was folglich Hacker abschrecken kann. Diese Überlegungen stellen lediglich einige der möglichen Erklärungen für das Ergebnis dar.

Des Weiteren wurde in dieser Studie untersucht, ob unterschiedliche Präregistrierungvorlagen einen moderierenden Effekt auf die Beziehung zwischen Originaleffektstärke und Replikationseffektstärke haben. Dabei konnte festgestellt werden, dass nur die Präregistrierungsvorlage Replication Recipe (Brandt et al., 2013): Pre-Registration einen signifikanten Einfluss auf den Zusammenhang hatte. Der Interaktionseffekt war negativ, was bedeutet, dass der Zusammenhang zwischen Originaleffektstärke und Replikationseffektstärke geringer wird, wenn die Replikation mit dieser Vorlage präregistriert wurde. Die anderen Vorlagen hatten keinen signifikanten Einfluss auf den Zusammenhang. Dabei ist es wichtig zu erwähnen, dass die Stichprobengrößen der unterschiedlichen Präregistrierungsvorlagen sehr gering waren. Bei Metaanalysen gibt es die Daumenregel, dass es mindestens 10 Beobachtungen pro Subgruppe geben muss (Schwarzer et al., 2015). In dieser Analyse hatten nur zwei Präregistrierungsvorlagen mehr als 10 Beobachtungen und von diesen zwei Vorlagen wies eine Vorlage einen signifikant moderierenden Einfluss auf.

Ein Grund, warum die Replikationskrise möglicherweise weniger gravierend ist als angenommen, wird von LeBel (2015, 2019) darin gesehen, dass die durchgeführten Replikationen sich erheblich von den Originalstudien unterscheiden. Diese sogenannte Closeness wurde in dieser Studie auch untersucht und es zeigte sich kein Einfluss auf den Zusammenhang zwischen

Originaleffektstärke und Replikationseffektstärke. Dieses Ergebnis unterstützt den Befund von Röseler et al. (2022), die Closeness in OSF Registries analysiert haben. Dieses Befundmuster überrascht, weil es logisch erscheint, dass die Replikation der Originalstudie möglichst nahe sein sollte. Gründe für den Befund könnten unentdeckte Moderator- oder Mediatorvariablen sein, die den Effekt von Closeness auf den Replikationserfolg beeinflussen. Alternativ könnte auch die Messung der Closeness-Variablen in dieser Studie nicht ausreichend präzise gewesen sein, um ihren Effekt zu erfassen. In der FReD Datenbank sollten Forschende, die ihre Befunde einreichten, subjektiv die Closeness für zehn unterschiedliche Closeness Variablen einschätzen. Dabei stand zur Auswahl, ob sie die Closeness für jede einzelne Variable als exakt, close, different, unknown oder does not apply einschätzen. Da die Variable optional war, könnte es hier zu einem Selektionsbias gekommen sein.

Limitationen

Eine der zentralen Limitationen dieser Untersuchung besteht darin, dass einige Replikationsstudien möglicherweise P-Hacking unterliegen, während andere vom Null-Hacking betroffen sind. Diese Effekte könnten sich gegenseitig aufheben, was zu einem Nullbefund führen würde, wie er auch in der vorliegenden Studie beobachtet wurde. Es käme demnach zu einem Fehler 2. Art, bei dem in der Population tatsächlich ein Effekt besteht, dieser jedoch fälschlicherweise als nicht vorhanden angenommen wird. Daher bräuchte es in Zukunft sensitivere Methoden, um das Ausmaß von P-Hacking und null-Hacking auseinanderdividieren zu können.

Eine weitere wichtige Limitation besteht darin, dass es für diese Studie essenziell ist, dass die Präregistrierung wirksam Hacking unterbindet. Wenn Hacking durch die Präregistrierung nicht unterbunden wird, wäre es fehlerhaft einen moderierenden Effekt des Präregistrierungsstatus oder einen Unterschied im Replikationserfolg als Anzeichen von Hacking zu anzusehen. Die Wirksamkeit von Präregistrierungen als Gegenmaßnahme zu P-Hacking ist empirisch jedoch noch nicht abschließend geklärt. Besonders die Wirksamkeit von Präregistrierungen bei Replikationen, insbesondere in Bezug auf Null-Hacking, hat bisher keine ausreichende Aufmerksamkeit erhalten. Wie van den Acker et al. (2024) feststellen, sind entscheidende Faktoren für die Wirksamkeit

Präregistrierung die producibility und consistency sowie die Abweichung von der Präregistrierung. Zudem ist auch ein festgelegter Analyseplan für die Wirksamkeit der Präregistrierung von hoher Bedeutung (Brodeur et al., 2024). Diese Qualitätskriterien von Präregistrierungen wurden in der vorliegenden Stichprobe jedoch nicht untersucht, da dies den Rahmen der Studie überschritten hätte. Daher kann wenig Aussage über die Wirksamkeit der Präregistrierung in dieser Stichprobe getroffen werden.

Bryan et. al. (2019) argumentieren zudem, dass besonders für Null-Hacking die Präregistrierung der Replikation nicht ausreichen könnte, weil es zu viele Möglichkeiten gibt Null zu hacken, die nicht von der Präregistrierung erfasst werden.

Das Kriterium der Vollständigkeit von Präregistrierungen kann in der Theorie durch detailliertere Präregistrierungsvorlagen gefördert werden. In der Praxis ist es jedoch so, dass bei offenen Formaten spezifischer als nötig und bei spezifischen Präregistrierungsvorlagen relevante Punkte weggelassen werden können. Daher ist der Vorlagentyp möglicherweise kein guter Schätzer für das Kriterium Vollständigkeit. Außerdem eignen sich einige Vorlagen für bestimmte Studiendesigns besser als andere Vorlagen, wodurch eine Konfundierung mit dem Studiendesign möglich wird. So fragt beispielsweise die Vorlage „Preregistration Template from AsPredicted.org“ nach den Versuchsbedingungen, nach denen die Versuchspersonen zugeteilt werden - was ein experimentelles Design impliziert. Einige Vorlagen sind zudem sehr offen gestaltet und stellen nur wenige Fragen, wobei die Hauptinformationen in angehängten Dokumenten enthalten sind. Zum Beispiel stellt die OSF-Standard Pre-Data Collection Registration Vorlage nur zwei Fragen, während die restlichen Informationen in Form eines freien Dokuments hinzugefügt werden. Damit wird der Vorlagentyp weniger entscheidend; wichtiger ist, wie die Forschenden mit den Vorlagen arbeiten.

Zu den allgemeinen Limitationen gehört, dass die verwendete Datenbank nicht repräsentativ ist. Die Replikationsstudien wurden nach Verfügbarkeit hinzugefügt, was zur Folge hat, dass die Datenbank Studien aus unterschiedlichen Forschungsfeldern enthält, wobei bestimmte Felder über- oder unterrepräsentiert sind. Daher ist die Generalisierbarkeit der Ergebnisse eingeschränkt.

Zudem handelt sich bei dieser Studie nicht um ein experimentelles Design sondern um eine Post-hoc-Analyse. Die Variablen Präregistrierung der Replikation, Auswahl der Präregistrierungsvorlage und Closeness wurden nicht systematisch manipuliert und es wurde keine Randomisierung durchgeführt. Würden diese Variablen Unterschiede im Zusammenhang zwischen Originaleffektstärke und Replikationseffektstärke aufweisen, wären diese mit Selektionsmechanismen konfundiert.

Darüber hinaus stammen 48 der 124 in die präregistrierte Analyse einbezogenen präregistrierten Replikationseffekte (38,71%) aus einer einzigen Replikationsstudie (Liu et al., 2023). Die verbleibenden 76 Effekte sind auf 62 Präregistrierungslinks verteilt, was einem Durchschnitt von 1,23 Effekten pro Präregistrierungslink entspricht. Diese Verteilung deutet darauf hin, dass insbesondere die eine Replikationsstudie (Liu et al., 2023) das Potenzial hat, die Ergebnisse zu verzerren.

Abweichung von der Präregistrierung in dieser Untersuchung

Im R-Skript der Präregistrierung wird im Code zur Erstellung des Graphen die Moderatorvariable „testmod“ anstelle der Variable „preregistration_replication“ verwendet. Die Variable „testmod“ diente als Testmoderator und erzeugte zufällige Werte, um zu überprüfen, ob der Code funktioniert, bevor er präregistriert wurde. Beim Graphen wurde versäumt, den Testmoderator durch die korrekte Variable zu ersetzen. Im finalen R-Skript ist im geänderten präregistrierten Code der Kommentar „CHANGED“ vermerkt.

Offene Forschungsfragen

Die ursprüngliche Idee dieser Studie war es, die Wirksamkeit der Methode Präregistrierung anhand der bis dato größten Replikationsdatenbank zu überprüfen, in dem untersucht werden sollte, ob Originalliteratur die präregistriert ist, eine höhere Replizierbarkeit vorweist. Nach der manuellen Überprüfung der 30 aktuellsten Originalstudien (nicht Originaleffekte) der Datenbank, wurde festgestellt, dass von diesen 30 Studien nur eine Studie präregistriert war. Es wurde daraufhin vermutet, dass die Prävalenz von präregistrierter Originalliteratur eher rückläufig ist, je weiter man in der Zeit zurückgeht, da die Methode relativ jung ist. Aufgrund der zu geringen Anzahl von

präregistrierten Originalliteratur in der Datenbank, wurde diese Forschungsidee verworfen. Dennoch wäre diese Fragestellung für zukünftige Forschung äußerst interessant, da die Wirksamkeit der Methode Präregistrierung als Gegenmaßnahme zu P-Hacking noch wenig empirische Evidenz besitzt.

In Anbetracht der Tatsache, dass nur eine der aktuellsten 30 Originalstudien in der Datenbank präregistriert ist und in der Replikationsliteratur lediglich etwa ein Fünftel der Replikationen präregistriert wurden, sollten Forscher*innen erneut dazu ermutigt werden, dass Präregistrierungen zügig durchgeführt werden können und dass präregistrierte Studien weder weniger signifikante Ergebnisse noch kleinere Effektstärken aufweisen (van den Acker, 2023).

Die Replikationsdatenbank FReD sollte außerdem mehr Beachtung finden. Forschende, Lehrende, Studierende und Praktizierende sollten darüber informiert werden, dass es eine Plattform gibt, auf der sie einfach und übersichtlich überprüfen können, ob ihre verwendete Literatur repliziert wurde. Zudem wäre es wünschenswert, wenn diese Personen ihre eigene Forschung in die Datenbank eintragen, um deren Umfang zu erweitern und zukünftig weitere wertvolle Erkenntnisse zu ermöglichen. Zusätzlich sollten im Rahmen von Forschungsförderungen mehr Mittel bereitgestellt werden, um zum Beispiel das Problem der geringen Validierungen zu adressieren, da der Aufbau und die Pflege der Datenbank erhebliche Ressourcen erfordern.

Fazit

Diese Studie untersuchte als eine der ersten Studien, Hacking in Replikationsstudien anhand der aktuell umfangreichsten Replikationsdatenbank. Dabei konnte die Hypothese, dass es in Replikationsstudien zu Hacking kommt, in dieser Studie nicht untermauert werden.

Nach Veranschaulichung der zentralen Limitationen dieser Studie (v.a. Ausgleich von Null- und P-Hacking, Wirksamkeit der Präregistrierungen) braucht es jedoch weitere Forschung, die Hacking in Replikationen untersucht, da dieses Thema von außerordentlicher Relevanz ist. Wenn Replikationen, die den Anschein nach Qualitätskontrolle und Fortschritt der Wissenschaft machen, in Wirklichkeit gehacked sind, kann das gefährliche Konsequenzen haben.

Referenzen

Aldhous, P. (2011). Journal rejects studies contradicting precognition. *New Scientist*. https://www.newscientist.com/article/dn20447-journal-rejects-studies-contradicting-precognition/

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *The Leadership Quarterly*, *28*(1), 5–21. https://doi.org/10.1016/j.leaqua.2017.01.006

Attali, D., & Baker, C. (2023). ggExtra: Add Marginal Histograms to 'ggplot2', and More 'ggplot2' Enhancements (Version 0.10.0) [Computer software]. https://CRAN.R-project.org/package=ggExtra

Bakan, D. (1970). *Toward a Reconstruction of Psychological Investigation* (4thEdition Aufl.). Jossey-Bass Inc, Publishers.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

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. https://doi.org/10.1177/1745691612459060

Bakker, M., Veldkamp, C. L. S., Assen, M. A. L. M. van, Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, *18*(12), e3000937. https://doi.org/10.1371/journal.pbio.3000937

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear Mixed-Effects Models Using Eigen and S4* (Version 1.1-23) [Computer software]. https://CRAN.R-project.org/package=lme4

Berinsky, A. J., Druckman, J. N., & Yamamoto, T. (2021). Publication Biases in Replication Studies. *Political Analysis*, *29*(3), 370–384. https://doi.org/10.1017/pan.2020.34

Boyce, V., Mathur, M., & Frank, M. C. (2023). Eleven years of student replication projects provide evidence on the correlates of replicability in psychology. *Royal Society Open Science*, *10*(11), 231240. https://doi.org/10.1098/rsos.231240

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F., Geller, J., & Giner-Sorolla, R. (2013). The replication recipe: What makes for a convincing replication? J. *Exp. Soc. Psychol.*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

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

Brodeur, A., Cook, N., Hartley, J., & Heyes, A. (2024). Do Pre-Registration and Pre-Analysis Plans Reduce p-Hacking and Publication Bias?: Evidence from 15,992 Test Statistics and Suggestions for Improvement. *Journal of Political Economy Microeconomics*. https://doi.org/10.1086/730455

Bryan, C. J., Walton, G. M., Rogers, T., & Dweck, C. S. (2011). Motivating voter turnout by invoking the self. *Proceedings of the National Academy of Sciences*, *108*(31), 12653–12656. https://doi.org/10.1073/pnas.1103343108

Bryan, C. J., Yeager, D. S., & O'Brien, J. M. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences*, *116*(51), 25535–25545. https://doi.org/10.1073/pnas.1910951116

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science (New York, N.Y.)*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Carey, B. (2015, August 27). Many Psychology Findings Not as Strong as Claimed, Study Says. *The New York Times*. https://www.nytimes.com/2015/08/28/science/many-social-science-findings-not-as-strong-as-claimed-study-says.html

Carey, B. (2016, März 3). New Critique Sees Flaws in Landmark Analysis of Psychology Studies. *The New York Times*. https://www.nytimes.com/2016/03/04/science/psychology-replication-reproducibility-project.html

Carter, E. C., Schachtman, T. R., & Lin, Y. (2020). *broom.mixed: Convert Statistical Analysis Objects into Tidy Data Frames* (Version 0.2.8) [Computer software]. https://CRAN.R-project.org/package=broom.mixed

Center for Open Science. (n.d.). *TOP Factor: Journals* (Transparency and Openness Promotion). Retrieved October 25, 2024, from https://topfactor.org/journals

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., & others. (2018). *pwr: Basic Functions for Power Analysis* (Version 1.3-0) [Computer software]. https://CRAN.R-project.org/package=pwr

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10), 211037. https://doi.org/10.1098/rsos.211037

Clarke, B., Lee, P. Y., Schiavone, S. R., Rhemtulla, M., & Vazire, S. (2023). *The Prevalence of Direct Replication Articles in Top-Ranking Psychology Journals*. OSF. https://doi.org/10.31234/osf.io/sa6rc

Clarke, B., Lee, P. Y. (K.), Schiavone, S. R., Rhemtulla, M., & Vazire, S. (2024). The prevalence of direct replication articles in top-ranking psychology journals. *American Psychologist.* Advance online publication. https://doi.org/10.1037/amp0001385

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods*, *2*(4), 447–452. https://doi.org/10.1037/1082-989X.2.4.447

Dickersin, K., Chan, S., Chalmersx, T. C., Sacks, H. S., & Smith, H. (1987). Publication bias and clinical trials. *Controlled Clinical Trials*, *8*(4), 343–353. https://doi.org/10.1016/0197-2456(87)90155-3

Ellemers, N. (2013). Connecting the dots: Mobilizing theory to reveal the big picture in social psychology (and why we should do this). *European Journal of Social Psychology*, *43*(1), 1–8. https://doi.org/10.1002/ejsp.1932

Elms, A. C. (1975). The crisis of confidence in social psychology. *American Psychologist*, *30*(10), 967–976. https://doi.org/10.1037/0003-066X.30.10.967

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist*, *35*(9), 790–806. https://doi.org/10.1037/0003-066X.35.9.790

Fanelli, D. (2010). Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data. *PLoS ONE*, *5*(4), e10271. https://doi.org/10.1371/journal.pone.0010271

Fishman, D. B., & Neigher, W. D. (1982). American psychology in the eighties: Who will buy? *American Psychologist*, *37*(5), 533–546. https://doi.org/10.1037/0003-066X.37.5.533

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Garnier, S. (2018). *viridis: Color Scales for R* (Version 0.5.1) [Computer software]. https://CRAN.R-project.org/package=viridis

Gerber, A., Huber, G., & Fang, A. (2018). Do Subtle Linguistic Interventions Priming a Social Identity as a Voter Have Outsized Effects on Voter Turnout? Evidence From a New Replication Experiment. *Political Psychology*, *39*(4), 925–938. https://doi.org/10.1111/pops.12446

Gerber, A. S., Huber, G. A., Biggers, D. R., & Hendry, D. J. (2016). A field experiment shows that subtle linguistic cues might not affect voter behavior. *Proceedings of the National Academy of Sciences*, *113*(26), 7112–7117. https://doi.org/10.1073/pnas.1513727113

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, *26*(2), 309–320. https://doi.org/10.1037/h0034436

Giner-Sorolla, R. (2012). Science or Art? How Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science. *Perspectives on Psychological Science*, *7*(6), 562–571. https://doi.org/10.1177/1745691612457576

Good, I. J. (1972). Statistics and Today's Problems. *The American Statistician*, *26*(3), 11–19. https://doi.org/10.2307/2682859

Greenwald, A. G. (Hrsg.). (1976). An editorial. *Journal of Personality and Social Psychology*, *33*(1), 1–7. https://doi.org/10.1037/h0078635

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., … Zwienenberg, M. (2016). A Multilab Preregistered Replication of the Ego-

Depletion Effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Heirene, R., LaPlante, D., Louderback, E. R., Keen, B., Bakker, M., Serafimovska, A., & Gainsbury, S. M. (2021). *Preregistration specificity & adherence: A review of preregistered gambling studies & cross-disciplinary comparison*. OSF. https://doi.org/10.31234/osf.io/nj4es

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology*, *66*, 81–92. https://doi.org/10.1016/j.jesp.2015.09.009

In praise of replication studies and null results. (2020). *Nature*, *578*(7796), 489–490. https://doi.org/10.1038/d41586-020-00530-6

Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, *58*(6), 543–549. https://doi.org/10.1016/j.jclinepi.2004.10.019

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A „many labs“ replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kravitz, D., & Mitroff, S. (2020). *Quantifying, and correcting for, the impact of questionable research practices on false discovery rates in psychological science*. OSF. https://doi.org/10.31234/osf.io/fu9gy

Lakens, D. (2019). *The Value of Preregistration for Psychological Science: A Conceptual Analysis*. https://doi.org/10.31234/osf.io/jbh4w

Lakens, D. (2023). *Concerns about Replicability, Theorizing, Applicability, Generalizability, and Methodology across Two Crises in Social Psychology*. OSF. https://doi.org/10.31234/osf.io/dtvs7

Larivière, V., & Costas, R. (2016). How Many Is Too Many? On the Relationship between Research Productivity and Impact. *PLOS ONE*, *11*(9), e0162709. https://doi.org/10.1371/journal.pone.0162709

LeBel, E. P. (2015). A New Replication Norm for Psychology. *Collabra*, *1*(1), 4. https://doi.org/10.1525/collabra.23

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A Brief Guide to Evaluate Replications. *Meta-Psychology*, *3*. https://doi.org/10.15626/MP.2018.843

Lubin, A. (1957). Replicability as a publication criterion. *American Psychologist*, *12*(8), 519–520. https://doi.org/10.1037/h0039746

Lüdecke, D. (2018). *esc: Effect Size Computation for Meta Analysis* (Version 0.5.1) [Computer software]. https://CRAN.R-project.org/package=esc

Mahoney, M. J. (1979). Review Paper: Psychology of the Scientist: An Evaluative Review. *Social Studies of Science*, *9*(3), 349–375. https://doi.org/10.1177/030631277900900304

Martin, B. (1992). SCIENTIFIC FRAUD AND THE POWER STRUCTURE OF SCIENCE. *Prometheus*, *10*, 83–98. https://doi.org/10.1080/08109029208629515

Michaels, D., & Monforton, C. (2005). Manufacturing Uncertainty: Contested Science and the Protection of the Public's Health and Environment. *American Journal of Public Health*, *95*(S1), S39–S48. https://doi.org/10.2105/AJPH.2004.043059

Miller, C., & Bamberger, P. (2016). Exploring Emergent and Poorly Understood Phenomena in the Strangest of Places: The Footprint of Discovery in Replications, Meta-Analyses, and Null Findings. *Academy of Management Discoveries*, *2*. https://doi.org/10.5465/amd.2016.0115

Moreau, D., & Wiebels, K. (2023). Ten simple rules for designing and conducting undergraduate replication projects. *PLOS Computational Biology*, *19*(3), e1010957. https://doi.org/10.1371/journal.pcbi.1010957

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration Is Hard, And Worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

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. https://doi.org/10.1177/1745691612459058

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Ooms, J. (2014). *jsonlite: A Json Parser/Generator* (Version 1.6) [Computer software]. https://CRAN.R-project.org/package=jsonlite

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

Pereboom, A. C. (1971). Some Fundamental Problems in Experimental Psychology: An Overview. *Psychological Reports*, *28*(2), 439–455. https://doi.org/10.2466/pr0.1971.28.2.439

Pfeiffer, T., Bertram, L., & Ioannidis, J. P. A. (2011). Quantifying Selective Reporting and the Proteus Phenomenon for Multiple Datasets with Similar Bias. *PLoS ONE*, *6*(3), e18362. https://doi.org/10.1371/journal.pone.0018362

Protzko, J. (2018). *Null-hacking, a lurking problem*. OSF. https://doi.org/10.31234/osf.io/9y3mp

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.1) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org

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. https://doi.org/10.1177/0956797614553946

Registered replication reports. (n.d.). Association for Psychological Science. Retrieved October 7, 2024, from https://www.psychologicalscience.org/publications/replication

Replication crisis. (2024, October 6). In Wikipedia. https://en.wikipedia.org/wiki/Replication_crisis

Rohrer, D., Pashler, H., & Harris, C. R. (2015). Do subtle reminders of money change people's political views? *Journal of Experimental Psychology: General*, *144*(4), e73–e85. https://doi.org/10.1037/xge0000058

Röseler, L., Gendlina, T., Krapp, J., Labusch, N., & Schütz, A. (2022). *Successes and Failures of Replications: A Meta-Analysis of Independent Replication Studies Based on the OSF Registries*. OSF. https://doi.org/10.31222/osf.io/8psw2

Röseler, L., Kaiser, L., Doetsch, C. A., Klett, N., Seida, C., Schütz, A., Aczel, B., Adelina, N., Agostini, V., Alarie, S., Albayarak-Aydemir, N., Aldoh, A., Al-Hoorie, A. H., Azevedo, F., Baker, B. J., Barth, C. L., Beitner, J., Brick, C., Brohmer, H., … Zhang, Y. (2024). *The Replication Database: Documenting the Replicability of Psychological Science*. OSF. https://doi.org/10.31222/osf.io/me2ub

Röseler, Lukas. (o. J.). *Open Science: Eine Einführung.* osf.io/2qxwv

Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research*. Appleton-Century-Crofts.

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

Rothstein, H., Sutton, A. J., & Borenstein, M. (2005). *Publication Bias in Meta-analysis: Prevention, Assessment And Adjustments* (Annotated Edition). John Wiley & Sons Inc.

Schauberger, G., & Walker, H. (2021). *openxlsx: Read, Write and Edit xlsx Files* (Version 4.2.5) [Computer software]. https://CRAN.R-project.org/package=openxlsx

Schauer, J. M., & Hedges, L. V. (2021). Reconsidering statistical methods for assessing replication. *Psychological Methods*, *26*(1), 127–139. https://doi.org/10.1037/met0000302

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007467. https://doi.org/10.1177/25152459211007467

Schimmack, U. (2020). A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology / Psychologie canadienne*, *61*(4), 364–376. https://doi.org/10.1037/cap0000246

Schöch, C. (2023). Repetitive research: A conceptual space and terminology of replication, reproduction, revision, reanalysis, reinvestigation and reuse in digital humanities. *International Journal of Digital Humanities*, *5*(2), 373–403. https://doi.org/10.1007/s42803-023-00073-y

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. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). *Life after P-Hacking* (SSRN Scholarly Paper 2205186). https://doi.org/10.2139/ssrn.2205186

Simonsohn, U. (2023, November 13). *[115] Preregistration Prevalence*. Data Colada. https://datacolada.org/115

Singh Chawla, D. (2021). 8% of researchers in Dutch survey have falsified or fabricated data. *Nature*. https://doi.org/10.1038/d41586-021-02035-2

Soto, C. J. (2019). How Replicable Are Links Between Personality Traits and Consequential Life Outcomes? The Life Outcomes of Personality Replication Project. *Psychological Science*, *30*(5), 711–727. https://doi.org/10.1177/0956797619831612

Sovacool, B. K. (2008). Exploring Scientific Misconduct: Isolated Individuals, Impure Institutions, or an Inevitable Idiom of Modern Science? *Journal of Bioethical Inquiry*, *5*(4), 271–282. https://doi.org/10.1007/s11673-008-9113-6

Sterling, T. D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.2307/2282137

Syed, M. (2023). *Some Data Indicating that Editors and Reviewers Do Not Check Preregistrations during the Review Process*. OSF. https://doi.org/10.31234/osf.io/nh7qw

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113

van den Akker, O. R., Bakker, M., van Assen, M. A. L. M., Pennington, C. R., Verweij, L., Elsherif, M. M., Claesen, A., Gaillard, S. D. M., Yeung, S. K., Frankenberger, J.-L., Krautter, K., Cockcroft, J. P., Kreuer, K. S., Evans, T. R., Heppel, F. M., Schoch, S. F., Korbmacher, M., Yamada, Y., Albayrak-Aydemir, N., … Wicherts, J. M. (2024). *The potential of preregistration in psychology: Assessing preregistration producibility and*

preregistration-study consistency. https://chesterrep.openrepository.com/handle/10034/628700

van den Akker, O. R., van Assen, M. A. L. M., Bakker, M., Elsherif, M., Wong, T. K., & Wicherts, J. M. (2023a). Preregistration in practice: A comparison of preregistered and non-preregistered studies in psychology. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02277-0

van den Akker, O. R., van Assen, M. A. L. M., Enting, M., de Jonge, M., Ong, H. H., Rüffer, F., Schoenmakers, M., Stoevenbelt, A. H., Wicherts, J. M., & Bakker, M. (2023b). Selective Hypothesis Reporting in Psychology: Comparing Preregistrations and Corresponding Publications. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231187988. https://doi.org/10.1177/25152459231187988

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* [Computer software]. Springer.

Wickham, H., & Bryan, J. (2019). *httr: Tools for Working with URLs and HTTP* (Version 1.4.2) [Computer software]. https://CRAN.R-project.org/package=httr

Wickham, H., François, R., Henry, L., & Müller, K. (2018). *dplyr: A Grammar of Data Manipulation* (Version 0.8.5) [Computer software]. https://CRAN.R-project.org/package=dplyr

Xie, Y. (2015). *knitr: A Package for Dynamic Report Generation in R* (Version 1.33) [Computer software]. https://CRAN.R-project.org/package=knitr

Zhu, H. (2021). *kableExtra: Construct Complex Table with Kable and Pipe* (Version 1.3.4) [Computer software]. https://CRAN.R-project.org/package=kableExtra

Schauberger, G., & Walker, H. (2021). *openxlsx: Read, Write and Edit xlsx Files* (Version 4.2.5) [Computer software]. https://CRAN.R-project.org/package=openxlsx