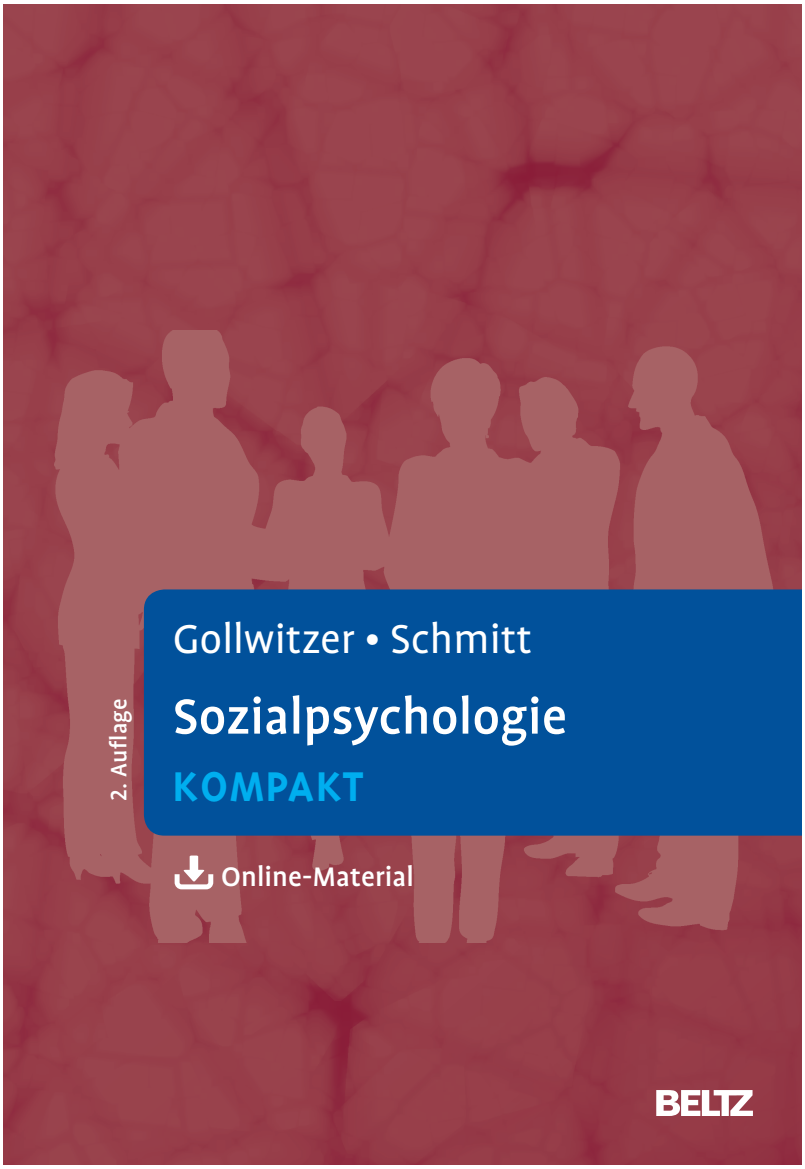
Online-Material

In dieser Datei finden Sie:

- Text: Replizierbarkeit und Generalisierbarkeit sozialpsychologischer Befunde



M. Gollwitzer/M. Schmitt
Sozialpsychologie
kompakt
2. Auflage

Replizierbarkeit und Generalisierbarkeit sozialpsychologischer Befunde

Was Sie in diesem Kapitel erwartet

In Kapitel 19 des Buches haben wir das Gütekriterium der Reliabilität oder Zuverlässigkeit kennen gelernt. Es besagt, dass ein Messinstrument unter gleichen Bedingungen für gleiche Objekte gleiche Messwerte liefern muss. Wenn wir die gleiche Person mit einer Waage mehrmals nacheinander wiegen, erwarten wir, dass die Waage immer das gleiche Ergebnis anzeigt. Von wissenschaftlichen Studien erwarten wir ebenfalls, dass sie unter gleichen Bedingungen gleiche Ergebnisse liefern, Ergebnisse also, die replizierbar sind. In der Psychologie, und insbesondere in der Sozialpsychologie, gibt es seit einigen Jahren eine lebhafte Debatte über die Frage, wie replizierbar empirische Befunde sind und wie es zu erklären ist, dass der Anteil erfolgreicher Replikationsstudien (also Studien, in denen ein statistisch bedeutsamer Befund in der Originalstudie auch in einer Replikationsstudie statistisch bedeutsam war) vergleichsweise gering ist. Auch wenn es bislang noch keine eindeutigen Antworten auf diese Fragen gibt, wollen wir die Entwicklung und den Stand dieser Diskussion hier nachzeichnen.

1 Die Replikationsdebatte

1.1 Anlässe für die Replikationsdebatte

Warum behandeln wir dieses Thema? Im ersten Kapitel haben wir einen kurzen Überblick über die Geschichte der sozialpsychologischen Forschung gegeben und dabei zwei Krisen erwähnt. Ihre erste Krise durchlitt die Sozialpsychologie in den 1960er und 1970er Jahren aufgrund von Zweifeln an der Aussagekraft ihrer Experimente, die als künstlich und lebensfern kritisiert wurden. In ihre zweite Krise geriet die Sozialpsychologie in den Jahren 2011 und 2012, als das Vertrauen in wissenschaftliche Forschung aufgrund von Fälschungen, fragwürdiger Forschungspraktiken und zunehmend deutlichen Hinweisen auf die unzureichende Replizierbarkeit wissenschaftlicher Befunde erschüttert wurde. Diesen Vertrauensverlust gab es nicht nur in der Psychologie. Auch Medizin, Biologie, Pharmazie, Neurowissenschaften und weitere Disziplinen sahen sich etwa gleichzeitig oder bereits früher mit Zweifeln an der Seriosität ihrer wissenschaftlichen Methoden und der Replizierbarkeit ihrer Befunde konfrontiert (Begley & Ellis, 2012; Fanelli, 2009; Ioannidis, 2005, 2008; Prinz et al., 2011). In der Psychologie war die Sozialpsychologie in besonderem Maße von dieser Kritik betroffen. Neu waren die Bedenken allerdings nicht. Auch vor 2011 waren gescheiterte Replikationsversuche kein Geheimnis. Sie wurden allerdings nur anekdotisch berichtet, nur selten publiziert, vor allem aber nicht systematisch aufgearbeitet.

Dies änderte sich, als die Psychologie – und insbesondere die Sozialpsychologie – 2011 durch drei nahezu gleichzeitige Ereignisse aufgerüttelt wurde. Bei dem ersten Ereignis handelte es sich um einen Fall massiven wissenschaftlichen Betrugs durch einen renommierten Sozialpsychologen; bei dem zweiten um eine Serie von Studien, die offenbar den empirischen Nachweis erbrachten, dass Menschen in die Zukunft sehen (»präkognizieren«) können. Das dritte Ereignis war die Veröffentlichung eines Beitrags, in dem gezeigt wurde, dass in typischen psychologischen Experimenten Daten durch ein »geschicktes« Vorgehen so generiert und ausgewertet werden können, dass die inhaltliche Hypothese mit großer Wahrscheinlichkeit bestätigt werden kann, obwohl sie falsch ist – und dies ganz ohne Fälschung oder freie Erfindung von Daten.

Wissenschaftlicher Betrug

Im September 2011 wurde bekannt, dass Diederik Stapel, ein renommierter Sozialpsychologe an der Universität Tilburg (Niederlande), die meisten seiner Daten gefälscht oder frei erfunden hatte. Begonnen hatte sein Fehlverhalten schon während der Dissertation an der Universität von Amsterdam, an der Stapel 1997 promovierte. Mit zunehmendem Karriereerfolg wurden Stapels Betrügereien immer dreister. Er plante seine Untersuchungen gemeinsam mit seinen Doktorandinnen und Doktoranden nach den Regeln der Kunst. Seine theoretischen Ideen waren kreativ und überzeugend. Der Betrug setzte an den Daten an.

Entweder manipulierte Stapel regulär erhobene Daten so, dass sie im Einklang mit seinen Hypothesen standen, oder er erfand Daten völlig frei mit dem gleichen Ziel. Die Studien ließen sich prominent publizieren, da die theoretischen Ideen interessant waren und durch die Daten nahezu perfekt bestätigt wurden. Diese Erfolge waren manchen Kolleginnen und Kollegen zwar suspekt, man konnte Stapel aber kein Fehlverhalten nachweisen. Vielmehr erwarb er sich nach und nach den Ruf eines genialen Forschers, der es besser als andere versteht, gute Theorien zu entwickeln und Studien so durchzuführen, dass ihre Ergebnisse stimmig und überzeugend ausfielen. Erst 2011 gelang es drei Doktoranden, genügend belastbares Material zusammenzutragen und sich damit an den Rektor der Universität zu wenden. Die drei zur Aufklärung des Verdachts eingesetzten Kommissionen gelangten zu dem Schluss, dass Stapel im großen Stil Daten gefälscht oder frei erfunden hatte. Der gemeinsame Bericht der Kommissionen erschien am 28.11.2012. Die Fassung in englischer Sprache ist online verfügbar und sehr lesenswert: https://www.tilburguniversity.edu/upload/3ff904d7-547b-40ae-85fe-bea38e05a34a_Final%20report%20Flawed%20Science.pdf

Der Bericht der Kommissionen war ein Schock für die Sozialpsychologie, da Stapel einer ihrer angesehensten Vertreter war, seine Studien in den besten sozialpsychologischen Zeitschriften publizierte und viele Jahre als Gutachter und Mitherausgeber dieser Zeitschriften tätig war. Allerdings ist Stapel nicht der einzige Wissenschaftler, der Daten erfunden hat. Auch aus anderen Wissenschaften sind ähnliche Fälle wissenschaftlichen Betrugs bekannt (Stroebe et al., 2012). Zudem wurden auch nach der Entdeckung des Fehlverhaltens von Stapel unplausible Datenmuster in der Sozialpsychologie entdeckt, die so unwahrscheinlich sind, dass der Zufall als Erklärung ausscheidet. Stapel ist also kein Einzelfall, allerdings ist dieser Fall besonders gut dokumentiert. Außerdem gilt er als der bislang schwerste Betrugsfall in der Sozialpsychologie und schließlich hat der Zeitpunkt seiner Entdeckung maßgeblich zur Replizierbarkeitsdiskussion beigetragen.

Präkognition

Können Menschen Ereignisse vorhersagen, bevor diese eintreten? Diese tatsächliche oder vermeintliche hellseherische Fähigkeit ist eines von mehreren Phänomenen, mit dem sich die Parapsychologie befasst, das als Präkognition (wörtlich: Vordenken) bezeichnet wird und das seit 1935 Gegenstand vieler Studien war (Honorton & Ferrari, 1989). Die meisten Menschen glauben nicht daran, dass man zukünftige Ereignisse vorhersehen kann, insbesondere dann nicht, wenn diese zufällig sind wie z. B. ein Münzwurf. Daryl Bem, ein höchst angesehener Sozialpsychologe, behauptete das Gegenteil und berichtete in einer Serie von neun Experimenten empirische Befunde, die für seine Auffassung zu sprechen schienen (Bem, 2011). Wir beschreiben nur das erste Experiment; die anderen Experimente waren ähnlich aufgebaut und führten zu ähnlich Ergebnissen.

Experiment

In Experiment 1 bekamen 100 studentische Versuchspersonen (50 Frauen, 50 Männer) folgende Aufgabe: Sie sollten vorhersagen, an welcher Position des Bildschirms (rechts oder links) ein erotisches Bild erscheinen würde. Das Experiment umfasste 100 Durchgänge. Bei 40 Durchgängen wurden je 36 Bilder gezeigt. Davon waren 12 erotisch, 12 negativ und 12 neutral. In 60 weiteren Durchgängen wurden ebenfalls je 36 Bilder gezeigt, davon 18 erotische und 18 positive. Welches Bild an welcher Stelle erschien, wurde durch Zufallsgeneratoren bestimmt, aber erst nachdem (!) die Person ihre Vorhersage getroffen hatte. Da die Position der Bilder völlig zufällig war, erwartet man eine zufällige Trefferquote von 50%. Die tatsächliche Trefferquote entsprach bei negativen, positiven und neutralen Bildern auch fast genau diesem Wert. Bei erotischen Bildern hingegen war sie mit 53.1% signifikant höher.

Bem interpretierte die Befunde dieses Experiments wie folgt: Die Personen »wussten«, wo ein erotisches Bild erscheinen würde. Und sie wussten es, bevor der Zufallsgenerator die Position bestimmt hatte. Die

Versuchspersonen verfügten also entweder über hellseherische Fähigkeiten oder sie konnten den Zufallsgenerator beeinflussen. Beide Erklärungen sind nach den gegenwärtigen wissenschaftlichen Überzeugungen unsinnig.

Wie man sich leicht vorstellen kann, fand der Artikel nicht nur in Fachkreisen, sondern auch den Medien und der Öffentlichkeit viel Beachtung. Kaum überraschen dürfte auch, dass eine heftige und kontroverse Diskussion über den Artikel von Bem losbrach. Seitens der wissenschaftlichen Psychologie wurde insbesondere die Sorge geäußert, dass die Publikation solcher Studien dem Ansehen der Psychologie schade und jenen in die Hände spiele, die der Psychologie ohnehin den Status einer ernstzunehmenden Wissenschaft absprechen. Auch wurden das methodische Vorgehen und die statistischen Datenanalysen heftig kritisiert (Wagenmakers et al., 2011). Versuche, die Befunde von Bem (2011) zu replizieren, scheiterten allesamt (z. B. Galak et al., 2012; Ritchie et al., 2012), sodass man mittlerweile konstatieren muss, dass es keine belastbaren Belege für hellseherische Fähigkeiten gibt.

Falsch Positive Befunde

Unter dem Titel *False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant* veröffentlichten Simmons et al. (2011) einen Artikel, in dem sie zeigten, dass man mit bestimmten Strategien der Erhebung und Auswertung von Daten nahezu jedes beliebige Ergebnis finden kann und somit auch Ergebnisse, die im Einklang mit den Hypothesen einer Untersuchung stehen. Zu den Strategien, auf die wir in Abschnitt 2 genauer eingehen werden, gehört es beispielweise, mehrere abhängige Variablen zu erheben, die Hypothesenprüfung aber nur anhand jener vorzunehmen, bei der sich der theoretisch erwartete Effekt am stärksten zeigt. Eine zweite Strategie besteht darin, von mehreren experimentellen Bedingungen nur jene in die Analyse einzubeziehen, die den erwarteten Effekt generieren. Geht man von der Annahme aus, dass es den Effekt in der Population nicht gibt (dass also – inferenzstatistisch gesprochen – die statistische Nullhypothese gilt), würde man nach der Logik der Inferenzstatistik bei einer vorab festgelegten Fehlerwahrscheinlichkeit von $\alpha = .05$ erwarten, dass von 100 Tests der Nullhypothese 5 auf ein statistisch bedeutsames (»signifikantes«) Ergebnis hindeuten (Eid et al., 2017). Die Wahrscheinlichkeit eines falsch positiven Befunds beträgt somit 5%. Simmons et al. (2011) zeigten nun, dass sich durch eine Kombination der genannten und weiterer Strategien diese Wahrscheinlichkeit auf 61% anheben lässt. Mit anderen Worten: Wenn man einen Effekt, den es nicht gibt, finden möchte, findet man ihn, wenn man nur »geschickt« und lange genug danach sucht. Da in der psychologischen Fachliteratur überwiegend signifikante Effekte berichtet werden und diese Effekte in der Regel mit den Hypothesen übereinstimmen (siehe unten), liegt der Verdacht nahe, dass die von Simmons et al. (2011) beschriebenen Strategien häufig praktiziert werden und dass damit viele der in der Literatur berichteten Effekte falsche Positive sind. Die von Simmons et al. (2011) beschriebenen Strategien werden deshalb auch als fragwürdige Forschungspraktiken bezeichnet. Ergebnisse, die unter Anwendung solcher Praktiken gefunden wurden, sollten sich nicht replizieren lassen, wenn auf diese Praktiken verzichtet wird.

1.2 Reproduzierbarkeit und Replizierbarkeit von Befunden

Aufgrund dieser und weiterer Ereignisse gerieten die Psychologie und insbesondere die Sozialpsychologie sehr stark in die Defensive. Es mehrten sich Stimmen innerhalb und außerhalb der Sozialpsychologie, die eine systematische Aufklärung forderten, welchen empirischen Befunden vertraut werden kann und bei welchen zu befürchten ist, dass sie durch die Anwendung fragwürdiger Forschungspraktiken oder Fälschung zustande gekommen sind. Dieser Forderung nach Aufklärung versucht man durch zwei Maßnahmen Rechnung zu tragen, der Prüfung der Reproduzierbarkeit der statistischen Analyseergebnisse und der unabhängigen Replikation von Befunden. Wir verwenden diese beiden Begriffe hier in der von Asendorpf et al. (2013) vorgeschlagenen Unterscheidung in Reproduzierbarkeit, Replizierbarkeit und Generalisierbarkeit.

Prüfung der Reproduzierbarkeit von Analyseergebnissen

Eine Maßnahme besteht darin, publizierte Daten erneut zu analysieren, um zu prüfen, ob die Analyseergebnisse reproduzierbar sind. Mit dieser Strategie sollen Auswertungsfehler, Datenmanipulationen und Fälschungen entdeckt werden. Von besonderem Interesse sind dabei Untersuchungen, deren Ergebnisse mit den Hypothesen (annähernd) perfekt übereinstimmen und (so gut wie) keine Unstimmigkeiten aufweisen. Wer über viel Forschungserfahrung verfügt, der weiß, dass Ergebnisse empirischer Untersuchungen so gut wie nie vollkommen stimmig und hypothesenkonform sind. Dies gilt nicht nur in der Psychologie, sondern in grundsätzlich allen empirischen Wissenschaften. Deshalb machen sich Forschende verdächtig, die wiederholt (nahezu) perfekte Ergebnisse publizieren. Man prüft somit den Verdacht, dass das Ergebnis zu gut ist um wahr zu sein. Diese Methode kann auch verwendet werden, um abzuschätzen, wie stark die Ergebnisse von Studien durch die Anwendung von fragwürdigen Forschungspraktiken zugunsten der Hypothese verzerrt sind (Bakker et al., 2012; Simonsohn et al., 2014).

Replizierbarkeit von Befunden

Während man bei der Reproduzierbarkeitsprüfung vorhandene Daten erneut analysiert, werden bei der Prüfung der Replizierbarkeit neue Daten erhoben. Im großen Stil wurde diese Strategie erstmals vom *Reproducibility Project* der Open Science Collaboration (2012) praktiziert, die von Brian Nosek ins Leben gerufen wurde. Zum Zeitpunkt der Vorstellung des Projekts (2012) hatten sich 72 Wissenschaftlerinnen und Wissenschaftler aus 41 Forschungseinrichtungen zusammengetan und bereit erklärt, offen und transparent Studien zu replizieren, die in drei prominenten psychologischen Fachzeitschriften (*Psychological Science, Journal of Personality and Social Psychology, Journal of Experimental Psychology: Learning, Memory, and Cognition*) im Jahre 2008 veröffentlich worden waren. Nach Bekanntwerden der Initiative schlossen sich der Gruppe weitere Wissenschaftlerinnen und Wissenschaftler an. Im August 2015 wurden die Ergebnisse des Projekts in der renommierten Zeitschrift *Science* unter dem Titel *Estimating the reproducibility of psychological science* veröffentlicht. Von den insgesamt 270 Autoren waren 100 Originalstudien repliziert worden. Die Autoren bemühten sich bei der Durchführung ihrer Replikationsstudien darum, die Originaluntersuchung so genau wie möglich nachzustellen, und berieten sich zu diesem Zweck intensiv mit den Autoren der Originaluntersuchungen. Die wichtigsten Ergebnisse des Gemeinschaftswerks waren:

(1) Die Effektstärke in den Replikationsstudien war durchschnittlich nur halb so groß wie in den Originalstudien. In der Maßeinheit des Produkt-Moment-Korrelationskoeffizienten ausgedrückt, betrug die durchschnittliche Effektstärke der Originaluntersuchungen $M(r) = .403$, die durchschnittliche Effektstärke in den Replikationsstudien $M(r) = .197$.
(2) In 97% der Originalstudien war der hypothesenrelevante Effekt signifikant, in den Replikationsstudien waren es nur noch 36%.
(3) Wurden die Studien in kognitionspsychologische und sozialpsychologische unterteilt, war die Replikationsrate bei den sozialpsychologischen Studien nur etwa halb so hoch wie bei den kognitionspsychologischen Studien.
(4) Einige Effekte ließen sich sehr gut und mit annähernd gleicher Stärke replizieren, andere Effekte ließen sich nicht replizieren und in manchen Replikationen wurden sogar gegenteilige Effekte gefunden wie in den Originalstudien.

Zusammengefasst lassen die Ergebnisse dieses äußerst wertvollen Replikationsprojekts mindestens drei Schlussfolgerungen zu. Erstens: Dass in den Originaluntersuchungen 97% der hypothesenrelevanten Effekte signifikant waren, könnte man so interpretieren, dass die Autoren der Studien fast immer mit ihrer Hypothese richtig lagen. Wenn dem tatsächlich so wäre, könnte man auf empirische Studien getrost verzichten und viel Geld sparen. Allerdings ist diese Interpretation nicht sehr plausibel. Zweitens: Der Befund könnte auch bedeuten, dass in der empirischen Forschung keine kühnen Theorien geprüft werden, weil die Gefahr zu groß ist, dass diese Theorien durch empirische Daten nicht gestützt werden. Erkenntnistheoretisch wäre dies geradezu katastrophal, denn wissenschaftlicher Fortschritt lebt von kühnen Ideen. Drittens: Die Ergebnisse bedeuten, dass Zweifel an der Verlässlichkeit psychologischer Forschungsergebnisse begründet sind und dass das Fach diesen Zweifeln systematisch nachgehen und sie so überzeugend

wie möglich ausräumen muss. Diese Schlussfolgerung ist konstruktiv, weil sie Möglichkeiten zur Verbesserung von Forschungsqualität eröffnet. Voraussetzung hierfür sind Analysen der möglichen Gründe für die Nichtreplizierbarkeit von Befunden und die Überschätzung von Effekten.

2 Gründe für die Nichtreplizierbarkeit von Befunden und die Überschätzung von Effekten

Einige Ergebnisse dieser Analyse möchten wir nun vorstellen. Die zu beantwortende Frage lautet: Wie kommt es, dass Befunde, die in der psychologischen Fachliteratur berichtet werden, teilweise falsch oder verzerrt sind? Wir können hier aus Platzgründen nicht auf alle möglichen Erklärungen hierfür eingehen. Wir beschränken uns auf drei Erklärungen, die wir für besonders einschlägig halten und sich ändern lassen, die Anreiz- und Gratifikationsstrukturen in der Wissenschaft, den Wettbewerb unter Fachzeitschriften und fragwürdige Forschungspraktiken.

2.1 Anreize und Gratifikationen in der Wissenschaft

Die Aufgabe von Wissenschaft ist Erkenntnisgewinn. Um diese Aufgabe bestmöglich zu erfüllen, werden Einrichtungen wie Universitäten und Forschungsinstitute geschaffen, die ihre Ziele im Wettbewerb miteinander anstreben. Erfolg in diesem System hängt sehr stark davon ab, wie gut es gelingt, talentierte, leistungsfähige und produktive Wissenschaftlerinnen und Wissenschaftler zu gewinnen. Als Gegenleistung werden diesen Selbstbestimmung ihrer Tätigkeit, Möglichkeiten der Mitgestaltung von Forschung und Lehre sowie gesellschaftlicher Status geboten. Wer sich als einzelner Wissenschaftler in diesem System durchsetzen will, muss sein Talent und seine Leistungsfähigkeit regelmäßig unter Beweis stellen. Das wichtigste Kriterium hierfür ist der Forschungserfolg, und diesen kann man unter anderem durch Veröffentlichungen in renommierten Zeitschriften demonstrieren. Wissenschaftler stehen deshalb unter dem Druck, erfolgreich zu publizieren. Daher rührt der Spruch »Publish or perish« – zu deutsch: Publiziere oder gehe zugrunde.

Der Wettbewerb um wissenschaftlichen Erfolg kann in Konflikt mit dem Gebot geraten, sich ausschließlich von dem Ziel leiten zu lassen, die Wahrheit zu finden. Wenn die Daten einer Studie nicht so sind, wie sie erwartet wurden und nicht so, dass sie sich publizieren lassen, steigt die Versuchung, mittels fragwürdiger Praktiken die Befunde so zurechtzubiegen, dass sie sich doch publizieren lassen. Verstärkt wird diese Versuchung dadurch, dass bei der Beurteilung der wissenschaftlichen Leistungsfähigkeit häufig die Anzahl von Publikationen mehr zählt als ihre Qualität.

2.2 Wettbewerb unter wissenschaftlichen Zeitschriften

Verstärkt wird das Problem auch durch den Wettbewerb unter Fachzeitschriften. Deren Herausgeber und Verleger möchten möglichst erfolgreich sein. Erfolg bemisst sich am Umfang der Leserschaft und einer damit zusammenhängenden Maßzahl, dem sogenannten Impact Factor. Vereinfacht gesagt gibt der Impact Factor einer Zeitschrift an, wie häufig Artikel, die in dieser Zeitschrift publiziert wurden, in anderen Artikeln zitiert werden. Je höher diese Zahl, desto mehr »Einfluss« wird einer Zeitschrift auf die Forschung zugeschrieben. Wie kann eine Zeitschrift nun ihren Impact Factor in die Höhe treiben? Insbesondere dadurch, dass die Artikel, die in ihr erscheinen, besonders interessant, kreativ und innovativ sind und dass die empirischen Ergebnisse klar und eindeutig sind. Replikationsstudien galten bis vor kurzem nicht als besonders innovativ und hatten es somit schwerer, publiziert zu werden. Studien, deren Ergebnisse Ungereimtheiten oder Widersprüche aufweisen, die schwer zu interpretieren sind oder die mehr Fragen aufwerfen als dass sie Antworten liefern, sind vergleichsweise schwerer zu lesen und werden weniger häufig zitiert. Das Gleiche gilt für Studien, in denen ein theoretisch erwarteter und plausibler Effekt empirisch nicht gezeigt werden konnte (weil der entsprechende statistische Effekt nicht signifikant

war; Ferguson & Heene, 2012). Die empirische Bestätigung einer theoretischen Hypothese wird als Erfolg gewertet, während mit einem nicht signifikanten Ergebnis die Sorge verbunden ist, der Forscher habe entweder nicht gründlich genug nachgedacht oder seine Hypothese nicht angemessen geprüft. Diese implizite Unterstellung dürfte in den meisten Fällen falsch sein. Sie ist auch problematisch, denn unerwartete Befunde sind grundsätzlich genauso wertvoll wie erwartete Befunde. Unerwartete Befunde können sogar besonders wertvoll sein, weil sie dazu zwingen, über ein psychologisches Phänomen und seine empirische Untersuchung erneut nachzudenken. Fallen Befunde hingegen wie erwartet aus, ist man dazu nicht gezwungen. Die Präferenz für signifikante Effekte führt dazu, dass in der Psychologie weniger als die Hälfte aller durchgeführten Studien überhaupt publiziert wird (Bakker et al., 2012). Anders gesagt: Studien, deren zentrale Effekte nicht statistisch signifikant sind, landen in der Schreibtischschublade und werden nicht veröffentlicht. Dies ist ein Problem für Sekundäranalysen (z. B. Metaanalysen), denn wenn man zur Schätzung eines Effekts nur diejenigen Primärstudien heranzieht, die auch publiziert wurden, überschätzt man die Größe des Effekts in hohem Maße. Man spricht hier auch vom »Schubladenproblem« (*file drawer problem*).

Gemeinsam mit der Abneigung gegen Replikationen wirkt sich die Präferenz für signifikante Ergebnisse besonders gravierend aus, denn Replikationsstudien, die signifikante Effekte der Originaluntersuchung nicht replizieren können, haben es doppelt schwer, publiziert zu werden. Die beschriebenen Prozesse führen dazu, dass in der Fachliteratur signifikante Effekte und Zusammenhänge über- und insignifikante Befunde unterrepräsentiert sind. Daraus resultiert eine systematische Fehleinschätzung der Gültigkeit von Theorien, den man Publication bias nennt (Fanelli, 2010).

2.3 Fragwürdige Forschungspraktiken

Wissenschaftlerinnen und Wissenschaftler kennen die Spielregeln des Systems. Sie wissen, dass ihr beruflicher Erfolg oder sogar ihre berufliche Existenz davon abhängen, wie viel und wie gut sie publizieren. Sie wissen auch, dass sich signifikante Effekte, schlüssige Ergebnisse und unerwartete Befunde leichter publizieren lassen als insignifikante Effekte, inkonsistente Ergebnisse und Befunde, die kaum überraschen. Dieses Wissen kann dazu verleiten, gegen Prinzipien guter wissenschaftlicher Praxis zu verstoßen und durch Anwendung fragwürdiger Forschungsprinzipien Befunde zu erzeugen, die sich gut publizieren lassen. Welche Forschungspraktiken sind fragwürdig?

Selektive Ergebnisberichte

Zwei der von Simmons et al. (2011) beschriebenen Strategien haben wir bereits beschrieben, die selektive Auswahl unabhängiger und abhängiger Variablen, bei denen sich der erwartete Effekt besonders deutlich zeigt (und das Verschweigen jener Variablen, Methoden oder Testergebnisse, bei denen sich der erwartete Effekt nicht zeigt).

Unzureichende Teststärke (Power)

Die Wahrscheinlichkeit, mit der man einen Effekt, der in der Population existiert, anhand einer Stichprobe statistisch nachweisen kann, heißt Teststärke oder Power. Sie hängt von der Stärke des Effekts, von dem gewählten Risiko einer fälschlichen Ablehnung der Nullhypothese (α-Niveau, meistens 5%) und von der Stichprobengröße ab. Diese Zusammenhänge kann man nutzen, um die Größe einer Stichprobe so zu wählen, dass eine bestimmte Power erreicht wird. Erwünscht ist eine Power von mindestens 80%. Je schwächer der Effekt in der Population ist, desto größer muss die Stichprobe sein, um die gewünschte Power zu erreichen. Diese Zusammenhänge sind lange bekannt und können in jedem Statistikbuch nachgelesen werden (Eid et al., 2017). Dennoch werden sie in der Forschungspraxis häufig missachtet. So fanden Marszalek et al. (2011) mittels einer Analyse der Artikel, die über einen Zeitraum von 30 Jahren in führenden Zeitschriften der Psychologie (*Journal of Abnormal Psychology; Journal of Applied Psychology; Journal of Experimental Psychology: Human Perception and Performance; Developmental Psychology*) publizierten wurden, eine mittlere (Median) Power von lediglich 40%.

Man könnte nun annehmen, dass dadurch das Risiko falsch positiver Befunde sinkt und das Risiko falsch negativer Befunde steigt. Dies wäre richtig, wenn die Hypothese an einer einmaligen Zufallsstichprobe aus der Population geprüft würde. Tatsächlich aber gibt es Hinweise, dass Forscher häufig mehrere kleine Stichproben statt einer ausreichend großen ziehen und nur die Befunde aus jener Stichprobe berichten, in der sich der Effekt als signifikant erwies (Bakker et al., 2012). Man kann sich das Vorgehen an einem Extrembeispiel veranschaulichen. Nehmen wir an, es behauptet jemand, Frauen seien größer als Männer. Dass diese Annahme falsch ist, ließe sich mit einer Zufallsstichprobe von 1.000 Frauen und 1.000 Männern belegen. Wenn stattdessen 200 kleine Stichproben mit je 5 Frauen und 5 Männern gezogen werden, ist darunter mit hoher Wahrscheinlichkeit eine Stichprobe, in der sich 5 untypisch kleine Männer und 5 untypisch große Frauen befinden. Der Unterschied könnte so groß sein, dass er signifikant wird und für die falsche Behauptung spricht. Unzulässig wäre es, nur dieses Ergebnis zu berichten und die Ergebnisse der 199 anderen Stichproben zu verschweigen.

P-hacking

Das zuletzt beschriebene unzulässige Vorgehen ist eine von mehreren Strategien, die unter dem Begriff des p-hacking zusammengefasst werden. Ganz allgemein bedeutet p-hacking, dass Grundsätze der Inferenzstatistik unterlaufen werden, um eine Hypothese zu bestätigen. Wir nennen einige weitere Beispiele für diese Praxis.

Schrittweise Vergrößerung der Stichprobe. Diese Strategie funktioniert so: Man beginnt eine Studie mit einer kleinen Stichprobe und prüft die Nullhypothese. Kann sie verworfen werden, bleibt es bei dieser Stichprobe und der Befund wird berichtet. Führt der Test nicht zu einem signifikanten Ergebnis, wird die Stichprobe etwas vergrößert und der Test erneut durchgeführt. Dieses Vorgehen wird so lange fortgesetzt, bis es erstmals gelingt, die Nullhypothese zu verwerfen.

Gerichtet statt ungerichtet testen. Diese Strategie besteht darin, eine ursprünglich ungerichtete Hypothese (Frauen und Männer können unterschiedlich gut erklären) gerichtet zu testen, nachdem die Daten erhoben wurden und man die Richtung des Unterschieds kennt. Die Wahrscheinlichkeit eines falsch positiven Befunds verdoppelt sich dadurch.

Mehrere statistische Tests verwenden. Die meisten Hypothesen lassen sich mit mehreren statistischen Tests prüfen, z. B. Tests mit und ohne Verteilungsannahmen (Eid et al., 2017). Die problematische Strategie besteht darin, alle geeigneten Tests anzuwenden und nur das Ergebnis desjenigen Tests mitzuteilen, der am deutlichsten für die Hypothese spricht.

Multiples Testen ohne Adjustierung der Fehlerwahrscheinlichkeit. Wenn Zusammenhänge zwischen mehreren Variablen oder Effekte mehrerer unabhängiger Variablen auf mehrere abhängige Variablen untersucht werden, müssen multivariate statistische Tests verwendet werden. Werden stattdessen viele Einzeltests verwendet, ist die effektive Fehlerwahrscheinlichkeit höher als die nominelle. Die nominelle Fehlerwahrscheinlichkeit muss deshalb adjustiert werden. Wird dieses Gebot unterlaufen und nur das Ergebnis eines Einzeltests berichtet, liegt ebenfalls ein Fall von p-hacking vor.

Umgang mit Ausreißern

In den meisten psychologischen Untersuchungen finden sich so genannte Ausreißer. Dabei handelt es sich um Personen mit untypisch hohen oder niedrigen Messwerten. Solche Ausreißer können Zusammenhänge oder Effekte stark beeinflussen, im Extremfall einen Zusammenhang oder Unterschied sogar umkehren. Beispiel: Wenn in einer Stichprobe von 4 Frauen und 4 Männern alle Frauen 170 cm und alle Männer 180 cm groß sind, sind die Frauen durchschnittlich kleiner (M = 170 cm) als die Männer (M = 180 cm). Befinden sich nun in die Strichprobe eine 5. Frau, die 200 cm groß ist und ein 5. Mann, der 150 cm groß ist, kehrt sich die durchschnittliche Größe um. Die Frauen sind nun durchschnittlich 176 cm groß, die Männer 174. In einer solchen Situation stellt sich die Frage, ob man die Ausreißer in der Stichprobe belassen oder ausschließen sollte. Zum Umgang mit Ausreißern gibt es Konventionen (Eid et al., 2017). Wie man mit Ausreißern in einem konkreten Fall umgeht, muss im Einzelfall entschieden und begründet werden. Die Regel, nach der Ausreißerwerte beibehalten vs. eliminiert werden, muss jedoch unabhängig von ihrer Auswirkung auf das zentrale statistische Analyseergebnis festgelegt und angewendet

werden. Inakzeptabel ist es also, aus mehreren Möglichkeiten des Umgangs mit Ausreißern diejenige auszuwählen, die zur Folge hat, dass die ursprüngliche Hypothese am ehesten bestätigt wird.

Umgang mit Kovariaten

In vielen psychologischen Untersuchungen werden demographische Variablen wie Alter, Geschlecht und Bildung routinemäßig miterhoben, selbst wenn mit diesen Variablen keine Hypothesen verbunden sind. Dennoch korrelieren diese Variablen meistens mit den relevanten Untersuchungsvariablen. Deshalb können sich Zusammenhänge und Effekte zwischen diesen ändern, wenn irrelevante Variablen als Moderatoren oder Kontrollvariablen in die Analysen einbezogen werden. So zeigt sich unter Umständen, dass der Effekt einer UV auf eine gemessene AV erst signifikant wird, wenn man eine Drittvariable (»Kovariate«) aus der AV herauspartialisiert (denn die Auspartialisierung verringert den Anteil der unerklärten Varianz in der AV). Unproblematisch ist es, im Vorhinein aufgrund plausibler Überlegungen festzulegen, welche Kovariaten kontrolliert werden. Inakzeptabel ist es hingegen, die Entscheidung über die Kontrolle von Kovariaten davon abhängig zu machen, welche Konsequenzen dies für die statistische Hypothesenprüfung hat.

Hypothesen nach Kenntnis der Ergebnisse formulieren (HARKing)

Noch fragwürdiger wäre es, zusätzlich zu einer strategischen Entscheidung über den Umgang mit Kovariaten eine theoretische Begründung hierfür zu erfinden, die es ursprünglich nicht gab. Post-hoc Hypothesen sind generell problematisch. Im Extremfall würde das fragwürdige Vorgehen so aussehen, dass man eine Untersuchung ohne Hypothesen durchführt, die Daten auswertet, dann Hypothesen formuliert, die zu den Ergebnissen passen und in der Veröffentlichung behauptet, die Studie zur Prüfung genau dieser Hypothesen durchgeführt zu haben. Dieses Vorgehen wird als HARKing (»Hypothesizing After the Results are Known«; Kerr, 1998) bezeichnet.

Mit der Kritik an diesem Vorgehen sagen wir nicht, dass exploratorische (erkundende) und induktive Forschung per se schlecht oder wertlos ist. Ganz im Gegenteil. In der Geschichte der Naturwissenschaften gibt es viele Beispiele dafür, dass wichtige Entdeckungen ungeplant gemacht und Theorien zu ihrer Erklärung erst anschließend entwickelt wurden. Fragwürdig ist es jedoch, die Ergebnisse einer exploratorischen Untersuchung im Nachhinein als gezielte Prüfung einer Hypothese darzustellen.

Wie häufig sind fragwürdige Forschungspraktiken?

Die beschriebenen Praktiken sind so offensichtlich fragwürdig, dass selbst Laien mühelos erkennen, zu welchen Verzerrungen und Fehlschlüssen sie führen. Man sollte deshalb erwarten, dass Wissenschaftlerinnen und Wissenschaftler aufgrund ihrer Ausbildung in Wissenschaftstheorie und Forschungsmethoden auf solche Praktiken verzichten. Eine anonyme Befragung von 2.155 Wissenschaftlerinnen und Wissenschaftlern der Psychologie durch John et al. (2012) ergab jedoch, dass die meisten mindestens eine dieser Praktiken schon einmal angewendet hatten, häufig sogar mehrere. Drei Beispiele: Die Frage nach Unterlassung der Nennung aller gemessenen abhängigen Variablen wurde von über 60% der Befragten bejaht. Fast 60% der Befragten gaben zu, die Erhebung weiterer Daten davon abhängig gemacht zu haben, ob das Ergebnis signifikant war. Fast 50% gaben an, in einem Artikel nur diejenigen Studien berichtet zu haben, die »funktionierten«. Aus diesen Prozentwerten kann man nicht schließen, dass fragwürdige Praktiken ständig angewendet werden. Aber bereits die einmalige Anwendung ist problematisch. Der Sozialpsychologie muss zu denken geben, dass in der Studie von John et al. (2012) aus ihrem Kreis am häufigsten fragwürdige Forschungspraktiken zugegeben wurden.

2.4 Systematische Unterschiede zwischen Original- und Replikationsstudien

Alle bisher genannten Gründe für die Nichtreplizierbarkeit oder Überschätzung von Effekten sind für die Wissenschaft und ihre Glaubwürdigkeit schädlich. Dies gilt nicht für einen Grund, den wir jetzt behandeln. Dieser birgt im Gegenteil Chancen für die Wissenschaft, indem er zur Differenzierung und Weiterentwicklung von Theorien und zur Gewinnung neuer Erkenntnisse beitragen kann.

Eine Replikationsstudie sollte nur unter gleichen Bedingungen zu den gleichen Ergebnissen wie die Originalstudie führen. Wenn sich Ergebnisse einer Originalstudie nicht replizieren lassen, kann dies somit bedeuten, dass die Replikationsstudie unter anderen Bedingungen durchgeführt wurde als die Originalstudie. Zu erkennen, in welcher Hinsicht die Durchführung der Replikationsstudie sich von der Durchführung der Originalstudie unterschieden hat, ist wissenschaftlich fruchtbar, trägt zu einem besseren Verständnis des untersuchten Phänomens bei und hilft, die Generalisierbarkeit des Effekts (»externe Validität«; s. Abschn. 19.2.1 im Buch) besser abzuschätzen. Um welche Unterschiede könnte es sich dabei handeln?

- Die Stichproben können sich in relevanter Weise unterscheiden (etwa hinsichtlich ihrer Größe, ihrer Eigenschaften etc.),
- die Untersuchungssituationen können sich unterscheiden (etwa hinsichtlich des Untersuchungslabors, des Versuchsleiters, des Ortes, der Tageszeit der Durchführung etc.),
- die Operationalisierungen der experimentellen Bedingungen können sich unterscheiden,
- die Messeigenschaften der Instrumente für die abhängigen Variablen können sich unterscheiden,
- die historischen Umstände können sich gewandelt haben und
- die kulturellen oder subkulturellen Kontexte können verschieden sein.

All diese und weitere Unterschiede in den Untersuchungsbedingungen kommen prinzipiell zur Erklärung unterschiedlicher Befunde in Frage. Wir möchten das Prinzip an den beiden letztgenannten Bedingungen, den historischen Umständen und den kulturellen Kontexten, erläutern.

Historischer Wandel

Viele sozialpsychologisch relevante Variablen unterliegen einem historischen Wandel. Dies gilt für Einstellungen zu Sexualität und Partnerschaft, Rollenbilder, religiöse Überzeugungen, Erziehungsziele, Kommunikationsgewohnheiten, Möglichkeiten der Freizeitgestaltung, Mobilität, Unterrichtsformen in Bildungseinrichtungen, Jugendkulturen sowie den Umgang mit ethnischer, kultureller und sprachlicher Heterogenität. Schauen wir uns ein Beispiel etwas genauer an, die Äußerung von Vorurteilen. In den 1960er Jahren war es in den USA üblich und in weiten Teilen der Bevölkerung akzeptabel, Schwarze als »Neger« zu bezeichnen und sich geringschätzig über sie zu äußeren. Durch die Bürgerrechtsbewegung und daran anschließende gesellschaftliche Entwicklungen hat sich das geändert. Heute ist es sozial unerwünscht, einen Schwarzen als Neger zu bezeichnen. Selbst die Bezeichnungen »Schwarze« und »Weiße« sind verpönt. Politisch korrekt ist es, von »African Americans« und »European Americans« zu sprechen. Als Folge dieser Entwicklungen haben sich die Messeigenschaften von Fragebögen für Vorurteile und Rassismus geändert. Dies ist der Grund für die Konstruktion neuer Instrumente, die zwischen subtilen und unverhohlenen Vorurteilen unterscheiden, eine Unterscheidung, die vor 50 Jahren weder nötig noch sinnvoll gewesen wäre (Pettigrew & Meertens, 1995).

Eine ähnliche historische Entwicklung hat es bei Geschlechtsstereotypen gegeben. Folglich haben sich auch die Messeigenschaften von Instrumenten für sexistische Einstellungen geändert. In den 1960er Jahren konnte man sexistische Einstellungen mit Items wie »Frauen sollten sich ihrem Ehemann unterordnen« messen. Dieses Item wäre heute psychometrisch extrem schwer und deshalb wenig trennscharf.

Kulturunterschiede

Kulturunterschiede können ebenfalls für unterschiedliche Ergebnisse in Studien verantwortlich sein, wenn diese in verschiedenen kulturellen Kontexten durchgeführt wurden. Die kulturvergleichende Psychologie hat eine Vielzahl sozialpsychologisch relevanter Kulturunterschiede gefunden. Ein bekanntes Beispiel sind independente (unabhängige) und interdependente (abhängige) Selbstkonzepte, die in individualistischen Kulturen (Nordamerika, Europa) und kollektivistischen Kulturen (Asien) unterschiedlich häufig vorkommen bzw. unterschiedlich stark ausgeprägt sind (Markus & Kitayama, 1991). Personen mit einem independenten Selbstkonzept definieren sich über ihre Individualität und somit ihre Unterschiedlichkeit zu anderen Menschen. Personen mit einem interdependenten Selbstkonzept definieren sich über ihre Zugehörigkeit zu Gruppen und somit ihre Ähnlichkeit mit anderen Menschen. Kulturen unterscheiden sich neben der typischen Ausprägung von Selbstkonzepten auch in deren Zusammenhängen mit anderen

Variablen. So hängen das Selbstwertgefühl und das emotionale Wohlbefinden einer Person davon ab, wie kompatibel ihr Selbstkonzept mit den vorherrschenden Normen der Kultur ist, in der sie leben. Je höher die Passung zwischen Selbstkonzept und soziokulturellen Erwartungen, desto positiver sollten Selbstwertgefühl und Wohlbefinden sein. Empirischen Studien bestätigen diese Erwartung (Gebauer et al., 2015). Ein bestimmtes Selbstkonzept wie »dominant sein« kann somit in einer Kultur positiv mit Wohlbefinden zusammenhängen, in einer anderen Kultur negativ. Für die Frage nach der Replizierbarkeit psychologischer Befunde bedeutet das: Eine Originalstudie kann sich – je nach kulturellem Kontext – in Bezug auf ihre Befunde von einer Replikationsstudie unterscheiden. Dieser Unterschied bedeutet nicht, dass in einer der beiden Studien oder in beiden schlechte Forschung betrieben wurde. Ganz im Gegenteil kann das Verständnis eines solchen Unterschieds zur Weiterentwicklung von Theorien und erweitertem Verständnis sozialpsychologischer Phänomene beitragen.

2.5 Generalisierbarkeit sozialpsychologischer Befunde

Die beiden zuletzt behandelten Beispiele zeigen, dass sich historische Veränderungen und Kulturunterschiede auf die Ergebnisse von Studien auswirken können und somit auch auf deren Replizierbarkeit. Unterschiedliche Ergebnisse sind in diesem Fall aber nicht das Ergebnis schlechter Forschung. Vielmehr haben sie systematische Ursachen, deren sorgfältige Analyse zur Weiterentwicklung von Theorien und Forschungsmethoden führen kann. Dies ist eine erste wichtige Feststellung.

Eine zweite lautet, dass Forschungsbefunde nicht ungeprüft über die historische Zeit und über Kulturen generalisiert werden dürfen. Vielmehr müssen regelmäßig theoretische Überlegungen angestellt werden, welche historischen Veränderungen und kulturellen Unterschiede welche Einschränkungen der Generalisierbarkeit mit sich bringen könnten. Diese Überlegungen sollten in empirischen Untersuchungen explizit berücksichtigt und gezielt überprüft werden. Gleiches gilt für die historische und kulturelle Äquivalenz von Messinstrumenten, die wir am Beispiel von Maßen für ethnische Vorurteile und sexistische Einstellungen erläutert haben. Äquivalenz darf nicht stillschweigend vorausgesetzt werden, sondern muss regelmäßig untersucht werden.

Historische Veränderungen und kulturelle Unterschiede sind Beispiele für Moderatoren, die den Zusammenhang zwischen Untersuchungsvariablen beeinflussen können. Weitere Moderatoren, die theoretisch in Betracht kommen und die Generalisierbarkeit von Effekten einschränken, sind das Geschlecht, das Alter, Persönlichkeitseigenschaften und kognitive Fähigkeiten.

In Metaanalysen versucht man, Moderatoren von Effekten und Zusammenhängen zu identifizieren (Rosenthal & Rosnow, 1984). Dabei wird zunächst statistisch geprüft, ob die Variabilität von Zusammenhängen und Effekten über Studien hinweg unsystematisch (d. h. durch Stichprobenfehler verursacht) oder systematisch ist. Wird die Variabilität als systematisch beurteilt, versucht man, sie durch Moderatoren aufzuklären. Moderatoren sind in diesem Fall Merkmale und Bedingungen der Studien wie die Art der verwendeten Messinstrumente, die Reihenfolge, in der die Variablen erhoben wurden, die Gestaltung experimenteller Bedingungen oder der Inhaltsbereich, in dem die Theorie geprüft wurde. Das größte Problem bei diesem Vorgehen besteht darin, dass nur solche Moderatoren geprüft werden können, zu denen Informationen aus allen oder den meisten der einbezogenen Studien vorliegen. Häufig sind diese Informationen theoretisch wenig ergiebig, da zum Zeitpunkt der Planung einer Originalstudie noch nicht bekannt ist, welche Moderatoren bei einer späteren Metaanalyse theoretisch von Interesse sein werden. Dennoch ist bereits der Nachweis einer systematischen Variabilität der Effekte über Studien hinweg wertvoll, weil er die Aufgabe stellt, diese in künftigen Untersuchungen theoriegeleitet aufzuklären.

Ein Beispiel zur Erläuterung: In einer Metaanalyse des Zusammenhangs zwischen impliziten und expliziten Einstellungen wird eine systematische Variabilität der Korrelationskoeffizienten zwischen den untersuchten Primärstudien gefunden. Die Studien waren u. a. danach kodiert worden, wie spontan explizite Einstellungen geäußert wurden. Es zeigte sich, dass Spontaneität den Zusammenhang zwischen impliziten und expliziten Einstellungen verstärkte (Hofmann et al., 2005). Auf der Grundlage dieses Befundes und seiner Interpretation anhand von Zweiprozesstheorien wurde ein Modell der Konvergenz impliziter und expliziter Dispositionen entwickelt und systematisch empirisch untersucht

(s. Abschn. 16.6.3 im Buch). Gemeinsam mit den Überlegungen zum historischen Wandel und Kulturunterschieden zeigt dieses Beispiel, dass Replikationsstudien, Metaanalysen und Generalisierbarkeitsstudien wertvolle Beiträge zur Weiterentwicklung von Theorien und Methoden der Sozialpsychologie leisten.

3 Lehren aus der Replikationsdebatte

Wie die meisten Debatten und Krisen, so ist auch die Replikationsdebatte lehrreich. Die Psychologie hat ihr Problem erkannt und nimmt es ernst. Welche Konsequenzen können aus der Debatte gezogen und welche Empfehlungen zur Verbesserung psychologischer Forschung gegeben werden?

3.1 Anreize für theoretisch informierte Replikationen

Eine erste Schlussfolgerung lautet, dass der wissenschaftliche Wert von Replikationen vor Beginn der Debatte unterschätzt wurde. Daraus folgt die Empfehlung, mehr Anreize für Replikationsstudien zu schaffen. Zeitschriften müssen sie als gleichwertig mit Originalstudien anerkennen und ihnen Raum geben. Diese Veränderung ist bereits im Gange. Auch bei der Bewertung wissenschaftlicher Leistungen müssen Replikationsstudien stärker gewichtet werden als bisher. Honoriert werden sollten vor allem Replikationen, die Erklärungen für unterschiedliche Ergebnisse in den Blick nehmen und Vorkehrungen treffen, diese gezielt zu überprüfen. Replikationen, wie sie von der Open Science Collaboration (2015) durchgeführt wurden, sind unschätzbar wertvoll. Sie sind aber nur ein erster Schritt. Sie müssen ergänzt werden um Studien, in denen mögliche Ursachen für Ergebnisunterschiede zwischen Originalstudie und Replikationsstudie gezielt überprüft werden. Dazu bedarf es gründlicher Theoriearbeit.

3.2 Regeln guter wissenschaftlicher Praxis

Die Vermittlung von Regeln guter wissenschaftlicher Praxis muss bereits im Studium beginnen. Ihrer Einhaltung muss von Betreuerinnen und Betreuern wissenschaftlicher Qualifikationsarbeiten, Gutachtern und Herausgebern von Fachzeitschriften und Einrichtungen der Wissenschaftsförderung hohe Priorität eingeräumt werden.

Flankiert werden müssen diese Maßnahmen durch eine solide Ausbildung von Studierenden in Wissenschaftstheorie und Forschungsmethoden. Hochwertige Forschung setzt beispielsweise ein Verständnis für den Zusammenhang zwischen Effektstärke, Power, Stichprobengröße und den beiden Fehlerwahrscheinlichkeiten (fälschliche Verwerfung der Nullhypothese, fälschliche Beibehaltung der Nullhypothese), Kenntnisse der Bayesschen Statistik, Verständnis für den Unterschied zwischen induktiver und deduktiver Forschung sowie für die Problematik multipler statistischer Tests voraus. Unzureichende Kenntnis oder Beachtung dieser Prinzipien weist auf Defizite in der Methodenlehre hin, denen konsequent entgegenwirkt werden muss. Diese Forderung gilt für alle psychologischen Studiengänge weltweit. Gute Forschung ist ohne fortgeschrittene Methodenkenntnisse unmöglich.

3.3 Reform wissenschaftliche Anreize

Anreize für Replikationsstudien haben wir bereits empfohlen. Die vielen Projekte der Open Science Collaboration (wie etwa die Replizierbarkeitsstudie aus dem Jahr 2015 oder die so genannten »Many Labs«-Studien) können hierbei als Vorbild dienen. Solche Gemeinschaftsprojekte müssen honoriert und dauerhaft etabliert werden. Künftige Replikationen sollten sich zunächst auf besonders bekannte und häufig zitierte Effekte konzentrieren, da diese in Lehrbücher und damit in das sozialpsychologische Wissen sehr vieler Psychologiestudierender einfließen.

Um die Einhaltung von Regeln guter wissenschaftlicher Praxis zu fördern, sollte bei der Bewertung wissenschaftlicher Leistungen die Qualität der Forschung höher gewichtet werden als die Menge von Publikationen. Der Grundsatz »publish or perish« muss seine Gültigkeit verlieren. Honoriert werden müssen sorgfältige Theoriearbeit, gründliche Aufarbeitung des Forschungsstands, begründete Auswahl und sorgsame Konstruktion von Messinstrumenten, konstruktvalide experimentelle Bedingungen, messgenau erhobene abhängige Variablen, angemessene Power zur Entdeckung von Effekten, die Verwendung von statistischen Modellen, die für die Daten optimal geeignet sind und jeglicher Verzicht auf fragwürdige Forschungspraktiken.

3.4 Offenheit und Transparenz

Für besonders wirksam halten wir schließlich alle Maßnahmen, die eine lückenlose Transparenz von Studien, der verwendeten Methoden, der erhobenen Daten und ihrer Auswertung gewährleisten. Dazu gehören unter anderem

- die Präregistrierung von Versuchs- und Auswertungsplänen,
- die angemessene Bereitstellung von erhobenen Daten zum Zwecke der Reanalyse und
- die Veröffentlichung der verwendeten Untersuchungsmaterialien.

Präregistrierung

Mit dem Instrument der Präregistrierung soll verhindert werden, (1) dass Wissenschaftlerinnen und Wissenschaftler dazu verführt werden, fragwürdige Analysepraktiken zu verfolgen (insbesondere, analytische Entscheidungen davon abhängig zu machen, welche Konsequenzen diese auf die Ergebnisse haben; s. Abschn. 2), und (2) dass die Ergebnisse von Untersuchungen Einfluss auf ihre Publikationschancen haben. Eine Präregistrierung funktioniert wie folgt: Autoren beschreiben – bevor sie ihre Untersuchung durchführen – ihre Fragestellungen, die theoretischen Grundlagen, die Hypothesen, die Methoden, den Versuchsplan, die Stichprobe und die geplanten statistischen Auswertungsschritte so genau wie möglich. Diese Beschreibung machen sie öffentlich zugänglich (z. B. auf einer Internetplattform) oder geben sie sogar – bei einer Zeitschrift, die sich für solche »registered reports« eignet – zur Begutachtung frei. Mit der Veröffentlichung (entweder ohne Begutachtung oder nach Einarbeitung von Rückmeldungen durch unabhängige Gutachterinnen und Gutachter) erklären die Autoren, die Daten nicht bereits erhoben zu haben und versichern, die Studie genau wie beschrieben durchzuführen. Die Beschreibung kann nach dieser Veröffentlichung nicht mehr verändert werden. Im Falle eines begutachteten »registered reports« wird die Studie dann durchgeführt, der Ergebnis- und Diskussionsteil wird verfasst und das fertige Manuskript wird dann veröffentlicht – unabhängig davon, ob die Ergebnisse hypothesenkonform oder -konträr sind.

Damit entfällt sowohl der Anreiz als auch die Möglichkeit, durch Anwendung fragwürdiger Praktiken, wie wir sie in Abschnitt 2.3 beschrieben haben, die Ergebnisse im Nachhinein zu schönen. Zunehmend mehr Zeitschriften bieten die Möglichkeit der Präregistrierung an. In anderen Disziplinen wird dieses Instrument schon länger verwendet. Es hat nachweislich dazu geführt, dass der teilweise absurd hohe Anteil signifikanter Effekte und somit Umfang falsch positiver Befunde deutlich zurückgegangen ist.

Offen zugängliche Daten

Eine zweite Säule offener Wissenschaft sind offen zugängliche Daten. Damit werden mehrere Ziele verfolgt. Erstens kann die Reproduzierbarkeit von Analyseergebnissen leicht überprüft werden, wenn die Daten einer Studie zugänglich sind. In Kombination mit der Präregistrierung einer Studie kann Auswertungsfehlern oder fragwürdigen Praktiken leichter auf die Spur gekommen werden. Bereits das Wissen von Autoren um die Entdeckung solcher Fehler und Praktiken in der eigenen Forschung dürfte dazu führen, dass sie seltener werden. Das zweite Ziel besteht in der besseren Nutzbarkeit von Daten für Metaanalysen, denn indem alle Daten, die zur Prüfung einer bestimmten Hypothese erhoben wurden, verfügbar sind (unabhängig davon, ob die Ergebnisse für oder gegen die Hypothese sprechen), verringert sich nicht nur das »Schubladenproblem«; die metaanalytische Schätzung eines mittleren Effekts wird

durch die verfügbaren Informationen auch umfangreicher und genauer. Das dritte Ziel ist die Möglichkeit der Nachnutzung von Daten. Daten werden zur Untersuchung einer bestimmten Fragestellung erhoben, eignen sich aber oft auch für die Untersuchung weiterer Fragestellungen, an denen die Autoren ursprünglich kein Interesse hatten oder die ihnen nicht in den Sinn kamen. Wenn Daten offen zugänglich sind, können sie später für neue Fragestellungen genutzt werden. Die offene Zugänglichkeit von Daten bringt allerdings Probleme mit sich. Bestimmungen des Datenschutzes müssen penibel beachtet werden, damit die Daten nicht den Versuchspersonen zugeordnet werden können. Außerdem müssen die Interessen der ursprünglichen Autoren und die Interessen von Nachnutzern ausgeglichen werden. Mit der Erhebung, Aufbereitung und Dokumentation von Daten ist in der Regel viel Aufwand verbunden. Ein Nachnutzer hat diesen Aufwand nicht und kann somit schneller zu Ergebnissen und Publikationen kommen. Die ursprünglichen Autoren müssen deshalb bevorzugten Zugriff auf die Daten oder einen Teil der Daten haben. Um Interessenskonflikten und anderen Problemen vorzubeugen, hat die Deutsche Gesellschaft für Psychologie eine Kommission zur Erarbeitung von Regeln für den Umgang mit Forschungsdaten beauftragt. Die Vorschläge der Kommission wurden veröffentlicht und werden künftig auf der Basis von Erfahrungen regelmäßig evaluiert und angepasst (Schönbrodt et al., 2017).

Offene Materialien

Um die Ergebnisse einer Studie nachvollziehen zu können, genügt es nicht zu wissen, wie der Untersuchungsplan aussah (Präregistrierung), und die Daten zu kennen. Man muss auch nachvollziehen können, wie die Daten und Ergebnisse entstanden sind, also welche Materialien verwendet wurden und wie genau die Studie abgelaufen ist. Angaben hierzu sind oft so detailliert oder umfangreich, dass sie nicht in den Text einer Publikation aufgenommen werden können. Dennoch sollten sie verfügbar sein, beispielsweise um die Studie so exakt wie möglich replizieren zu können. Deshalb bieten viele Online-Repositorien sowie Zeitschriften inzwischen die Möglichkeit, alle Materialen zu hinterlegen und zugänglich zu machen. Zu Materialien gehörten z. B. die Messinstrumente, Skripte von Versuchssteuerungsprogrammen, Stimuli, Skripte der durchgeführten Datenanalysen sowie Angaben zu den Daten und zur Transformation von Rohdaten in Primärdaten (vgl. Schönbrodt et al., 2017). Auch hier müssen die Interessen aller Beteiligten sorgfältig gegeneinander abgewogen und Rechte wie das Urheberrecht beachtet werden.

3.5 Ein optimistischer Ausblick

Die Replikationsdebatte hat gezeigt, dass es in der Sozialpsychologie bei der Sicherung der Forschungsqualität Nachholbedarf gibt. Die Sozialpsychologie scheint besonders stark von der Nichtreplizierbarkeit von Befunden, der Überschätzung von Effekten und von der Anwendung fragwürdiger Forschungspraktiken betroffen oder betroffen gewesen zu sein (Open Science Collaboration, 2015; John et al., 2012). Dies ist eine schlechte Nachricht.

Die gute Nachricht lautet, dass sich Sozialpsychologen besonders engagiert für eine Korrektur der bisherigen Fehler eingesetzt haben und aktiv Strategien verfolgen, die Lehren aus der gegenwärtigen Krise für die Zukunft konstruktiv zu nutzen. Zum Ausdruck kommen diese Bemühungen in zahlreichen Artikeln und Sonderheften, die unter maßgeblicher Beteiligung von Sozialpsychologen verfasst oder herausgegeben wurden, so z. B. das Themenheft »Replizierbarkeit« der Psychologischen Rundschau (Jahrgang 69, Heft 1, 2018). Und nicht zuletzt ist dieses Kapitel zur Replizierbarkeitsproblematik Bestandteil eines Lehrbuchs der Sozialpsychologie. All dies sind ermutigende Anzeichen, dass sich die Sozialpsychologie auf einem guten Weg zur Steigerung der Qualität ihrer Forschung befindet.

4 Zusammenfassung

In den Jahren 2011 und 2012 ist die Sozialpsychologie durch drei nahezu gleichzeitige Ereignisse in eine Vertrauenskrise geraten: einen schweren Fall von Datenfälschung, die Publikation einer obskuren Studie zu hellseherischen Fähigkeiten von Menschen und einen Artikel mit dem Nachweis, dass durch eine

Kombination von Entscheidungen über die Erhebung und Auswertung von Daten nahezu jedes beliebige Ergebnis gefunden werden kann. Dabei handelt es sich nicht um Fälschungen, Manipulationen oder die Erfindung von Daten, sondern um fragwürdige Forschungspraktiken, die Regeln der Inferenzstatistik und der unverfälschten Schätzung von Effekten unterlaufen. Diese Ereignisse haben zur Forderung nach Überprüfung von Daten (Reproduzierbarkeit von Ergebnissen anhand der Originaldaten einer Studie) und nach mehr Replikationen geführt.

Ein groß angelegtes Gemeinschaftsprojekt der Open Science Collaboration (2015) hat 100 Studien aus namhaften Zeitschriften repliziert und kam zu einem beunruhigenden Ergebnis: Während in den Originalstudien 97% der statistisch getesteten Effekte signifikant waren, waren es in den Replikationen nur 36%. Außerdem waren die geschätzten Effekte in den Replikationsstudien nur halb so stark wie in den Originaluntersuchungen.

Dass in der Fachliteratur viele falsch positive Befunde und überschätzte Effekte publiziert werden, hat eine Reihe von Gründen. Die wichtigsten sind:

(1) In der Wissenschaft besteht ein starker Druck zu publizieren (Publish or perish), da Publikationen das wichtigste Erfolgskriterium bei der Bewertung wissenschaftlicher Leistungen sind und Wissenschaftlerinnen und Wissenschaftler ebenso wie wissenschaftliche Einrichtungen in einem starken Wettbewerb untereinander stehen.
(2) Auch wissenschaftliche Fachzeitschriften stehen im Wettbewerb untereinander. Sie streben einen möglichst hohen Impact Factor an. Sie möchten deshalb Untersuchungen publizieren, die neue, überraschende, in sich stimmige und signifikante Ergebnisse berichten, die eine Theorie entweder konsistent untermauern oder widerlegen.
(3) Diese beiden Gründe können Wissenschaftler in Versuchung bringen, durch die Anwendung fragwürdiger Forschungspraktiken ihre Ergebnisse so zu schönen, dass sie sich gut publizieren lassen.
(4) Unterschiede in Ergebnissen einer Replikationsstudie zu denen der Originalstudie können auch daran liegen, dass die Untersuchungsbedingungen nicht gleich waren. Solche Unterschiede in den Bedingungen nennt man Moderatoren. Zu ihnen gehören historischer Wandel und Kulturunterschiede. Diese und weitere Moderatoren sind für die Wissenschaft produktiv, indem sie zur Differenzierung und Weiterentwicklung von Theorien beitragen.
(5) Die Prüfung der Generalisierbarkeit ist eine wichtige Aufgabe von Forschung. Sie sollte theoriegeleitet vorgenommen werden. Ein geeignetes Instrument sind Metaanalysen.

Folgende Konsequenzen müssen aus der Replikationsdebatte gezogen werden:

(1) Replikationen müssen stärker als bisher in die Bewertung wissenschaftlicher Leistungen einfließen. Zeitschriften müssen mehr Platz für Replikationen bereitstellen. Besonders honoriert werden sollten umfangreiche Replikationsprojekte wie das der Open Science Collaboration (2015) und Replikationsstudien zu häufig zitierten Effekten, weil diese in Hand- und Lehrbücher eingehen und das Gesamtbild sozialpsychologischer Phänomene besonders stark prägen.
(2) Regeln guter wissenschaftlicher Praxis müssen bereits im Studium vermittelt, ihre Beachtung regelmäßig überprüft werde. Die Ausbildung in Wissenschaftstheorie und Methoden muss weltweit in allen Psychologiestudiengängen auf hohem Niveau stattfinden.
(3) Bei der Bewertung wissenschaftlicher Leistungen anhand von Publikationen muss der Grundsatz gelten: Qualität geht vor Quantität.
(4) Forschung muss offener und transparenter werden. Die wichtigsten Instrumente zur Erreichung dieses Ziels sind Präregistrierung, offen zugängliche Daten und offen zugängliche Untersuchungsmaterialien.

Die Sozialpsychologie scheint die Notwendigkeit erkannt zu haben, die Qualität ihrer Forschung zu steigern, und hat sich durch engagierte Beteiligung an Maßnahmen hierzu bereits hervorgetan.

Weiterführende Literatur

Eine ausführliche Darstellung der Replizierbarkeitsproblematik gibt dieser Artikel:

Asendorpf, J.B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J.A., Fiedler, K., Fiedler, S., Funder, D.C., Kliegl, R., Nosek, B.A., Perugini, M., Brent, W.R., Schmitt, M., van Aken, M.A.G., Weber, H. &

Wicherts, J.M. (2013). Recommendations for increasing replicability in psychology. European Journal of Personality, 27, 108–119.

Der Artikel wurde von 15 Experten/Expertengruppen kommentiert:
Open Peer Commentary. European Journal of Personality, 27, 120–138.

Diese Kommentare wurden wiederum von den Autoren des Artikels kommentiert:
Asendorpf, J.B., Conner, M., De Fuyt, F., De Houwer, J., Denissen, J.A., Fiedler, K., Fiedler, S., Funder, D.C., Kliegl, R., Nosek, B.A., Perugini, M., Brent, W.R., Schmitt, M., van Aken, M.A.G., Weber, H. & Wicherts, J.M. (2013). Replication is more than hitting the lottery twice. European Journal of Personality, 27, 138–144.