Die Geschlossene Gesellschaft und ihre Freunde

Frank Renkewitz

1 Einleitung

Der Versuch, Gesundheitskommunikation evidenzbasiert zu gestalten, ist offenkundig zunächst auf eine methodisch valide Bewertung vorliegender Daten angewiesen. Diese Bewertung von Evidenz muss geeignet sein, empirisch bewährte, theoretische Erkenntnisse von bloßen Zufallsbeobachtungen und falsch positiven Befunden zu trennen. Sie muss zudem eine zumindest näherungsweise angemessene Beschreibung des Gültigkeits- und Anwendungsbereichs bewährter Theorien ermöglichen. Diese Notwendigkeit ist in der Gesundheitskommunikation insoweit schwerer zu erfüllen als in anderen angewandten Disziplinen, als hier eine zweifach gelingende Evidenzbewertung erforderlich ist: Gesundheitsrelevante Botschaften sollten unter einem inhaltlichen Aspekt zunächst eine kritische bio- und medizinwissenschaftliche Prüfung überstanden haben. Eine im Hinblick auf das jeweilige Kommunikationsziel optimierte Gestaltung solcher Botschaften benötigt zudem eine valide Bewertung der entsprechenden kommunikationswissenschaftlichen Evidenz.

Mit einem Schwerpunkt in der Psychologie hat es in den letzten Jahren in zahlreichen Sozial- und Biowissenschaften umfangreiche Bemühungen gegeben, die wissenschaftliche Praxis der Evidenzbewertung zu evaluieren. Diese Bemühungen tragen der Einsicht Rechnung, dass eine umfassende und allgemeingültige Definition von „guter“ Evidenz kaum möglich ist – eine allgemein akzeptierte Antwort auf die Fragen, ob eine vermeintliche neue Erkenntnis wirklich belastbar und wichtig ist oder ob eine neue Theorie tatsächlich die bessere Theorie ist, findet sich oft erst nach Jahren und Jahrzehnten. Allerdings müssen empirische Forschungsergebnisse bestimmten Kriterien genügen, um zumindest potenziell eine Grundlage für Erkenntnisgewinn sein zu können. Ein zentrales und – gerade unter dem Gesichtspunkt der Evidenzbasierung – besonders offensichtliches Kriterium ist dabei Replizierbarkeit: Ein Forschungsresultat kann keine angemessene Grundlage für evidenzbasierte Entscheidungen sein, wenn die fragliche Evidenz in Replikationen der ursprünglichen Studie „verschwin-

det“. Entsprechend stützen sich Evaluationen der Evidenzbewertung in verschiedenen Wissenschaftsdisziplinen zunächst und vor allem auf Untersuchungen der Replizierbarkeit von publizierten Forschungsbefunden. Um das ernüchternde Ergebnis dieser Untersuchungen vorwegzunehmen: Forschungsbefunde in den Sozial- und Biowissenschaften sind weit seltener replizierbar als man erwarten sollte. Dies impliziert auch, dass eine Evidenzbasierung von Entscheidungen und Empfehlungen in der Praxis derzeit zumindest nicht systematisch gelingen kann.

Kapitel 2 des vorliegenden Beitrags liefert eine kurze Übersicht über entsprechende Untersuchungen zur Replizierbarkeit. Die darauf folgenden Teile befassen sich mit den Gründen für mangelnde Replizierbarkeit. Diese Gründe können auf unterschiedlichen Ebenen beschrieben werden: Aus methodischer Perspektive kann aufgezeigt werden, welche (augenscheinlich weit verbreiteten) Praktiken bei der Planung, Durchführung, Analyse und Interpretation von Studien Replizierbarkeit vermindern (s. Kapitel 3). Unter wissenschaftssoziologischen Gesichtspunkten kann beleuchtet werden, welche Anreizstrukturen im Wissenschaftssystem diese Praktiken befördern. Eine Gemeinsamkeit dieser beiden Beschreibungsebenen erklärt den Titel dieses Beitrags, der offensichtlich eine Anspielung auf Poppers (2003) sozialphilosophisches Hauptwerk „Die offene Gesellschaft und ihre Feinde“ ist. Diese Gemeinsamkeit besteht darin, dass sich sowohl in der gängigen Methodenanwendung (also im Verhalten des einzelnen Wissenschaftlers oder der einzelnen Wissenschaftlerin) als auch in den gegenwärtig gültigen „Spielregeln“ des Wissenschaftsbetriebs (also in den entsprechenden Anreizstrukturen) ein fehlender Wille zu Kritik, Irrtumskorrektur und Falsifikation manifestiert. Die Wissenschaftspraxis widerspricht damit in großen Teilen nicht nur Poppers (2005) wissenschaftstheoretischer Konzeption (also vor allem den Ideen von Fallibilismus und Falsifikationismus), sondern auch seiner Vorstellung von einer pluralistischen Gesellschaft, die im Meinungswettbewerb mit den Mitteln der Kritik nach besseren Lösungen sucht. Die empirische Praxis gestaltet sich in ihrem Ziel und Ergebnis tatsächlich überraschend häufig rein ‚affirmativ‘. Diese These soll im Abschnitt 4 untermauert werden.

2 Replizierbarkeit in den Sozial- und Biowissenschaften

Die bislang umfassendste Untersuchung zur Replizierbarkeit veröffentlichter Befunde in einer Wissenschaftsdisziplin ist das *Reproducibility*

Project: Psychology (Open Science Collaboration, 2015). Im Rahmen dieses Projekts wurden 100 Studien aus drei führenden und zitationsstarken psychologischen Fachzeitschriften repliziert. Inhaltlich handelte es sich bei den replizierten Studien überwiegend um experimentelle Untersuchungen aus den Bereichen der Sozialpsychologie und der Kognitiven Psychologie. Methodisch unterlagen die Replikationen einer relativ strikten Kontrolle: Der Stichprobenplan für die Auswahl der Originalstudien ließ kaum Raum für eine Selektion unglaubwürdiger oder methodisch zweifelhafter Befunde, die Replikationen folgten einem standardisierten Protokoll, das (soweit möglich) vor der Durchführung mit den Autorinnen und Autoren der Originalstudien abgestimmt wurde, die Stichproben der Replikationen wurden gegenüber den Originalstudien in den meisten Fällen vergrößert, Protokolle, Studienmaterialien, Daten und Analyseskripte wurden öffentlich zugänglich gemacht (osf.io/ezcuj). Trotz dieser Bemühungen scheiterte die Mehrzahl der Replikationen. Während 97 der Originalstudien ein statistisch signifikantes Ergebnis berichteten, fanden nur 36% der entsprechenden Replikationen abermals ein signifikantes Ergebnis. Die Evidenz, die zur Veröffentlichung der Originalstudien geführt hatte, war also in fast zwei Dritteln der Replikationen nicht länger auffindbar. Entsprechend waren auch die Effektstärken der Replikationen drastisch reduziert: Betrug die mittlere Effektstärke der Originalstudien noch $r = 0{,}40$, so lag der Mittelwert der Replikationen bei $r = 0{,}20$. Schwerwiegende Replikationsprobleme zeigten sich dabei sowohl in sozialpsychologischen als auch in kognitionspsychologischen Untersuchungen. Allerdings demonstriert das Reproducibility Project auch, dass durchaus systematische Unterschiede zwischen eng verwandten Wissenschaftsbereichen bestehen können: Die Replikationsquote im Sinne des Signifikanzkriteriums lag für kognitionspsychologische Studien bei 50 %, für sozialpsychologische hingegen bei lediglich 25 %. Auch die Effektstärken der Replikationen lagen für sozialpsychologische Untersuchungen deutlich näher bei 0 als für kognitionspsychologische Studien.

Die letzte Beobachtung führt natürlich zu der Frage, ob derartige Replikationsprobleme auf die Psychologie beschränkt sein könnten. Dies lässt sich empirisch recht eindeutig mit nein beantworten. Beispiele aus zwei anderen Disziplinen: Camerer et al. (2016) berichten in einer Studie mit Replikationen von 18 ökonomischen Experimenten eine immer noch enttäuschend schwache Replikationsquote von 60 %. Der Unterschied zur Replikationsquote in der Psychologie lässt sich kaum interpretieren, da hier allein Originalstudien mit einfaktoriellen Designs repliziert wurden –

was auch in der Psychologie die Replikationsquote deutlich verbessert. Belege aus der Medizin stammen unter anderem von Pharma-Unternehmen, die regelmäßig solche grundlagenwissenschaftlichen Befunde replizieren, die sie zur Entwicklung von Heilmitteln nutzbar machen wollen. Prinz et al. (2011) berichten für Bayer eine Replikationsquote von 25 %, laut Begley und Ellis (2011) beläuft sich die Replikationsquote bei Amgen sogar auf lediglich 11 %. Aufgrund der anderen Auswahlprozeduren verbietet sich auch hier ein Vergleich mit der Psychologie, dennoch machen diese Ergebnisse deutlich, dass Replikationsprobleme sicher nicht auf eine einzelne Wissenschaftsdisziplin beschränkt sind.

3. Ursachen der Replikationskrise: Ein methodischer Blick

Die statistische Evidenzbewertung der Ergebnisse aus Einzelstudien erfolgt in den Sozial- und Biowissenschaften nahezu ausschließlich mithilfe des Signifikanztests. Durch die Verwendung dieser Tests soll sichergestellt werden, dass beobachtete Effekte nicht plausibel durch Zufall erklärt werden können. Entsprechend gelten statistisch signifikante Effekte in der Forschungspraxis in aller Regel als ausreichendes Indiz für die Existenz eines „wahren" Effektes und die Annahme der untersuchten Forschungshypothese. Gemäß dieser Logik sollten Replikationen mit einer Stichprobengröße, die eine hinreichende Power sicherstellt, zumeist erneut ein signifikantes Ergebnis finden. Augenscheinlich ist diese Logik also fehlerbehaftet. Dies lässt sich leicht demonstrieren, wenn man eine Annahme über die Basisrate korrekter Forschungshypothesen in einem Forschungsgebiet trifft. Tabelle 1 illustriert eine Situation, in der von 1000 geprüften Forschungshypothesen 100 korrekt sind. Die damit spezifizierte Basisrate von 10 % ist zunächst willkürlich gewählt. Die übrigen Angaben in der Tabelle orientieren sich an der gängigen Forschungspraxis: Von 900 falschen Forschungshypothesen erhalten 45 dennoch ein signifikantes Ergebnis – dies entspricht dem üblichen Signifikanzniveau von $\alpha = 5\,\%$. Die 50 signifikanten Ergebnisse bei 100 korrekten Forschungshypothesen entsprechen einer Power von 50 % – und damit einem Schätzwert für die Power in der Psychologie und den Sozialwissenschaften, der über Jahrzehnte immer wieder ermittelt wurde (z. B. Cohen, 1962; Sedlmeier & Gigerenzer, 1989). Die Tabelle ermöglicht es nun, den positiven Vorhersagewert (PPV) zu bestimmen, also die Wahrscheinlichkeit, mit der ein signifikantes Ergebnis tatsächlich mit einer korrekten Forschungshypothese ver-

bunden ist. Unter den vorliegenden Randbedingungen beläuft sich der positive Vorhersagewert auf PPV = 50 / 95 = 53 %. Nahezu die Hälfte der signifikanten Ergebnisse tritt hier also in Untersuchungen auf, in denen tatsächlich kein wahrer Effekt besteht. Damit ist eine zentrale Ursache von Replikationsproblemen identifiziert: Statistische Signifikanz ist in bestimmten Konstellationen nur sehr schwache Evidenz zugunsten einer Hypothese, wird aber dennoch oftmals als ausreichender Beleg akzeptiert. Die Tabelle macht zudem die Faktoren kenntlich, die den Evidenzwert eines signifikanten Ergebnisses beeinflussen und die daher bei der Bewertung eines Studienergebnisses Berücksichtigung finden müssten: Der positive Vorhersagewert sinkt zunächst mit der Basisrate korrekter Hypothesen, also mit der Selektivität, mit der Hypothesen etwa aufgrund der Güte ihrer theoretischen Begründung untersucht und geprüft werden. Er sinkt zudem mit der Power der Studien (pragmatisch also vor allem mit der Stichprobengröße) und schließlich bei steigendem α, wenn also etwa auch „marginal signifikante" Ergebnisse als ausreichende Evidenz akzeptiert werden; vgl. Sedlmeier und Renkewitz, 2018, für eine ausführlichere Erläuterung).

Tabelle 1: Illustration des positiven Vorhersagewerts bei einer Basisrate korrekter Forschungshypothesen von 10 %, einer Power von 50 % und einem α von 5 %.

	Forschungshypothese (H_1) korrekt	Forschungshypothese falsch (H_0 korrekt)	*Randsumme*
Testergebnis signifikant ($p \leq 0{,}05$)	50	45	95
Testergebnis nicht signifikant ($p > 0{,}05$)	50	855	905
Randsumme	100	900	**1000**

Das Problem der Überbewertung signifikanter Ergebnisse als schlüssige Evidenz verschärft sich zudem drastisch durch die Verwendung „fragwürdiger Forschungspraktiken". Solche Praktiken können als „Produktionsmittel" zur Herstellung signifikanter Ergebnisse verstanden werden. Das wesentliche Funktionsprinzip besteht in der Wiederholung von

Signifikanztests: Bei einem α von 5% ist nach 20 Tests auch für die unsinnigste Hypothese ein signifikantes Ergebnis zu erwarten. Eine Studie 20-mal zu wiederholen, mag ineffizient und auch unrealistisch erscheinen, aber es gibt andere Mittel, die Zahl der durchgeführten Tests schnell zu erhöhen: In einer Studie können mehrere und potentiell austauschbare Operationalisierungen der abhängigen Variablen erhoben werden, Kovariaten können einbezogen werden oder unberücksichtigt bleiben, in der Regel wird es mehr als ein anwendbares Testverfahren geben, Teilnehmerinnen und Teilnehmer können nacherhoben und der Test bei unterschiedlichen Stichprobengrößen wiederholt werden, Ausreißer können ausgeschlossen werden oder im Datensatz verbleiben und so weiter. Alle diese Vorgehensweisen erhöhen auch dann die Wahrscheinlichkeit eines signifikanten Ergebnisses, wenn die Nullhypothese zutrifft (Simons et al., 2011). Die Auswirkungen solcher fragwürdigen Praktiken lassen sich leicht anhand von Tabelle 1 erläutern: Nehmen wir an, dass aufgrund von fragwürdigen Forschungspraktiken 10 % der ursprünglich nicht signifikanten Ergebnisse in der Literatur als signifikante Resultate erscheinen. Im Beispiel erhöht sich damit die Zahl signifikanter Ergebnisse bei zutreffenden Forschungshypothesen auf 55. Allerdings steigt die Zahl signifikanter Ergebnisse bei falschen Hypothesen auf 130. Dies impliziert einen positiven Vorhersagewert von PPV = 55 / (55 + 130) = 30 %. Nunmehr treten also 70 % der signifikanten Ergebnisse bei falschen Forschungshypothesen auf. Umfragen unter Wissenschaftlerinnen und Wissenschaftlern verschiedener Disziplinen dokumentieren, dass die Verwendung fragwürdiger Forschungspraktiken weit verbreitet ist (z. B. John et al., 2012; Head et al., 2015).

4. Affirmation und Kritiklosigkeit

Die vorschnelle Akzeptanz signifikanter Testergebnisse liefert offensichtlich auch einen ersten Beleg für die Neigung, vermeintlich hypothesenbestätigender Evidenz mit unzureichender Kritik zu begegnen. Die entsprechende affirmative Haltung betrifft allerdings verschiedene Aspekte des Forschungsprozesses und erweist sich auch zeitlich als erstaunlich stabil. Belege für diese These finden sich zunächst bei einem genaueren Blick auf die Faktoren, die den positiven Vorhersagewert des Signifikanztests beeinflussen: Die Bedeutung der Power wurde in den Sozialwissenschaften bereits in den 60er Jahren des vergangenen Jahrhunderts von Cohen (1962) popularisiert. In der methodisch orientierten Literatur ist die Dis-

kussion um die Power und ihre interpretatorische Bedeutung für den Signifikanztest seitdem nie abgerissen. Wie bereits angedeutet, hatte dies für die Forschungspraxis keine nennenswerte Konsequenz. Zumindest für die Psychologie ist durch eine Reihe von Untersuchungen gut dokumentiert, dass durchschnittliche Stichprobengrößen und damit die Power im Verlauf der folgenden Jahrzehnte nicht gestiegen sind. Eine jüngere Untersuchung von Bakker et al. (2012) findet sogar Stichprobengrößen, die die Schlussfolgerung nahelegen, dass die durchschnittliche Power zumindest in manchen Teilgebieten der Experimentalpsychologie bei lediglich 35 % liegen könnte. Kühberger et al. (2014) finden in einer methodischen Übersichtsarbeit über im Jahr 2007 erschienene psychologische Fachartikel, dass ganze 5 % eine explizite Powerkalkulation anstellen. Augenscheinlich hat das Konzept der Power über geraume Zeit also weder bei der Planung von Studiendesigns noch bei der Interpretation von Studienergebnissen an Bedeutung gewonnen. In Anbetracht der intensiven methodischen Diskussion des Konzepts und der offensichtlichen Relevanz der Stichprobengröße für die Belastbarkeit von Befunden wird man dies kaum auf Unwissenheit attribuieren können. Weit eher zeigt sich hier ein erstaunlich geringes Interesse an verlässlicher(er) Evidenz.

Der geringe positive Vorhersagewert in Tabelle 1 ist zudem durch die (angenommene) niedrige Basisrate korrekter Hypothesen verursacht. Dies führt zu der Frage, nach welchen Kriterien Hypothesen in der Forschungspraxis für einen empirischen Test ausgewählt werden. Aus Sicht des kritischen Rationalismus (Popper, 2005) sollten Theorien und die aus ihnen abgeleiteten Hypothesen zunächst einer logischen Prüfung auf Widerspruchsfreiheit und ihren empirischen Gehalt unterzogen werden (Glöckner et al., 2018). Tatsächlich findet eine vorgeordnete, nicht-empirische Prüfung von Hypothesen in den Sozialwissenschaften quasi nicht statt. Im Extremfall sind in der Psychologie auch Studien in führenden Zeitschriften veröffentlicht worden, die Psi-Phänomene (z. B. „Hellseherei“) untersuchen (und bestätigen) und die das Fehlen jeder theoretischen Erklärung für die geprüften Hypothesen explizit betonen (Bem, 2011). Dies ist in dieser Form sicher eine Ausnahme, die mangelnde theoretische Plausibilität untersuchter Hypothesen ist hingegen durchaus kennzeichnend für zahlreiche Arbeiten in den Sozialwissenschaften. Im Kern gilt hier quasi jeder beliebige „Einfall“ als testbar und testenswert. Dies muss eine Verminderung der Basisrate korrekter Hypothesen nach sich ziehen. Hier manifestiert sich also ein fehlender Wille zur Kritik auf der Ebene der Theoriebildung, der allerdings auch auf die Aussagekraft empirischer Prüfun-

gen zurückwirkt: Mit einem wachsenden Anteil falscher Hypothesen muss zwangsläufig auch die Häufigkeit fälschlicherweise bestätigter Hypothesen ansteigen.

In der Verbreitung fragwürdiger Forschungspraktiken zeigt sich schließlich, dass es Wissenschaftlerinnen und Wissenschaftlern oft an jeder Motivation zu einer kritischen Prüfung ihrer Hypothesen mangelt – ihr Interesse gilt in entsprechenden Fällen offensichtlich weniger richtigen Ergebnissen als vielmehr signifikanten Ergebnissen. Diese Motivationslage mag partiell durchaus psychologische Ursachen haben (die etwa aus der verständlichen Auffassung resultieren, dass die eigenen Ideen richtig und wichtig sind), sie kann aber natürlich nicht unabhängig von den Anreizstrukturen des Wissenschaftsbetriebs verstanden werden. Ein wesentlicher Bestandteil dieser Anreizstrukturen lässt sich ebenfalls anhand von Tabelle 1 verdeutlichen. Im Hinblick auf die Basisrate beruht die Tabelle sicher auf einer eher pessimistischen Annahme. Dieser Annahme ist auch der sehr hohe Anteil von 90,5 % nicht signifikanten Testergebnissen geschuldet. Allerdings impliziert eine durchschnittliche Power von 50 % auch, dass selbst dann 50 % aller Tests in nicht signifikanten Ergebnissen enden würden, wenn alle geprüften Hypothesen korrekt wären. Hingegen findet Fanelli (2010) in einer ganzen Reihe von Wissenschaftsdisziplinen (von der Psychologie über die Pharmakologie und die klinische Medizin zur Biologie und Ökonomie), dass der Anteil von signifikanten Hypothesentests in der publizierten Literatur stets im Bereich von 90 % liegt. Publikations-Biases, die die Veröffentlichung von negativen, nicht signifikanten Befunden erschweren oder gar ausschließen, müssen also disziplinübergreifend die Regel sein. Evidenz, die eine empirische Kritik an fehlerhaften Theorien oder nicht erfolgreichen Interventionen begründen und ermöglichen würde, ist in der Literatur systematisch unterrepräsentiert. Damit fehlt dem einzelnen Wissenschaftler und der einzelnen Wissenschaftlerin – die auf Publikationen angewiesen sind – aber auch jeder Anreiz zu einer kritischen Testung von Hypothesen. Seine und ihre Karriereaussichten hängen zumindest bis zu einem gewissen Grade an ihrer Bereitschaft zur Affirmation.

Wenn Theorien und Interventionen oftmals weder empirisch noch logisch einer ernsthaften Prüfung unterzogen werden, könnten zumindest die sie bestätigenden Befunde einer kritischen Betrachtung unterliegen. Dazu würden zum einen Maßnahmen gehören, die die Verwendung fragwürdiger Forschungspraktiken als Ursache positiver Befunde unwahrscheinlicher machen. Zum anderen könnten solche Befunde durch Replikationen

überprüft werden. Leider setzt sich auch hier die Neigung des Wissenschaftsbetriebs zu mangelnder Überprüfung und fehlendem Kritikwillen fort. Fragwürdige Forschungspraktiken könnten im Vorfeld einer Studie vor allem dadurch vermieden werden, dass Wissenschaftlerinnen und Wissenschaftler ihre Hypothesen und den angezielten Hypothesentest bereits vor der Datenerhebung vollständig spezifizieren. Entsprechende Präregistrierungen von Studien waren in den Sozialwissenschaften allerdings bis vor wenigen Jahren nahezu unbekannt und wurden kaum jemals genutzt. Im Nachgang einer Studie könnten fragwürdige Forschungspraktiken und schlichte Fehler (die sich im Forschungsprozess natürlich oftmals völlig unbeabsichtigt und nahezu unvermeidbar ereignen) vor allem anhand der Studienmaterialien, der Daten und der entsprechenden Analyseskripte erkannt werden. Allerdings sind neben der Publikation einer Studie in aller Regel weder die entsprechenden Materialien noch die Daten verfügbar. Zumindest im Hinblick auf Daten verlangen sowohl viele Fachzeitschriften als auch Ethikrichtlinien von wissenschaftlichen Fachgesellschaften (z. B. American Psychological Association, 2009) oftmals seit Jahrzehnten, dass diese mit anderen Wissenschaftlerinnen und Wissenschaftlern geteilt werden. Mehrere Untersuchungen in der Psychologie konnten jedoch zeigen, dass der Anteil der Autorinnen und Autoren von Originalstudien, die die zugehörigen Daten auf Anfrage tatsächlich übersenden, trotz dieser Verpflichtung lediglich im Bereich von 30 % liegt (Wicherts et al., 2006, 2011; Vanpaemel et al., 2015). Der Wissenschaftsbetrieb konstituiert also einerseits einen starken Anreiz, signifikante Ergebnisse zu finden und zu berichten, er fragt aber andererseits kaum danach, wie genau solche Ergebnisse zustande gekommen sind – tatsächlich gewährleistet er in aller Regel noch nicht mal eine Prüfung der Frage, ob überhaupt Daten vorliegen.

Als letzte mögliche Instanz der Prüfung von Befunden verbleiben damit Replikationen. Da falsch positive Befunde im Forschungsprozess unvermeidbar auftreten müssen, sollte man erwarten, dass unabhängige Replikationen zumindest zur Prüfung von Ergebnissen, die sich als einflussreich erweisen, die Regel sind. Auch hier bestätigt sich aber das bisher geschilderte Muster: In der Biologie (Palmer, 2000) und der Psychologie (Neuliep & Crandall, 1990, 1993) wurde schon vor geraumer Zeit dokumentiert, dass Herausgeberinnen und Herausgeber sowie Gutachterinnen und Gutachter von Zeitschriften eine starke Tendenz haben, Replikationen abzulehnen (oftmals mit dem offenkundig falschen Argument, dass der fragliche Befund bereits bekannt und gesichert sei). Entsprechend finden

Makel et al. (2012), dass deutlich weniger als 1 % der psychologischen Forschungsartikel unabhängige, direkte Replikationen enthalten.

An dieser Stelle schließt sich der Kreis: Mit dem Signifikanztest wird ein Verfahren zur Evidenzbewertung eingesetzt, das nicht ausschließt, dass die Mehrzahl hypothesenbestätigender Befunde tatsächlich falsch positiv ist, und das zudem – willentlich oder unwillentlich – relativ leicht „manipulierbar" ist. Die Publikationspraxis schafft für die einzelne Wissenschaftlerin und den einzelnen Wissenschaftler hohe Anreize, hypothesenbestätigende Resultate zu berichten, da negative Ergebnisse in der Literatur kaum auftreten. Der Forschungsprozess und die Daten, die einem berichteten signifikanten Ergebnis zugrunde liegen, unterliegen aber keiner adäquaten Prüfung und werden zumeist auch nicht in einer Weise dokumentiert, die sie zumindest potenziell kritisierbar machen würde. Schließlich werden die falsch positiven Befunde, die als Resultat dieses Zustands zu erwarten sind, auch nach ihrer Publikation bestenfalls selten und mit großer zeitlicher Verzögerung entdeckt, da auch Replikationen zur Befundsicherung von der Veröffentlichung weitgehend ausgeschlossen sind. In der Literatur entsteht so letztlich das Bild einer Wissenschaft, die Erfolge und Innovationen beinahe unterbrechungslos aneinanderreiht. Tatsächlich zeigt dieses Bild eine Wissenschaft, die kaum noch in der Lage ist, ihre Irrtümer von Erkenntnissen zu unterscheiden.

5. Die offene Wissenschaft

Dieser pessimistischen Perspektive steht allerdings die gute Nachricht entgegen, dass die Wissenschaftspraxis seit dem Aufkommen der Diskussion um eine Replikationskrise bereits einem raschen Wandel unterliegt. Zahlreiche Anzeichen sprechen dafür, dass sich dieser Wandel weiterhin fortsetzen und intensivieren wird. Inhaltlich lassen sich die entsprechenden Veränderungen dadurch charakterisieren, dass sie auf mehr Transparenz in allen Aspekten des Wissenschaftsbetriebs zielen. Sie betreffen damit sowohl das Handeln der einzelnen Wissenschaftlerinnen und Wissenschaftler als auch die institutionellen Rahmenbedingungen. In der Verantwortung des Einzelnen liegt es zunächst, in konfirmatorischen Untersuchungen theoriegeleitet vorzugehen, also zu prüfende Hypothesen aus relevanten Theorien in nachvollziehbarer Weise zu deduzieren. Entsprechende hypothesenprüfende Studien sollten zudem präregistriert werden, die geprüfte Hypothese und das Studiendesign sollten also bereits vor der

Durchführung dokumentiert werden. Dazu haben sich auch die technischen Voraussetzungen in jüngster Zeit deutlich verbessert: Mit dem Open Science Framework (osf.io) steht eine internetbasierte, disziplinübergreifende und leicht handhabbare Registratur zur Verfügung. Dort finden sich auch diverse Templates als Hilfestellung für die Abfassung einer Präregistrierung. Der OSF lässt sich zudem als Repositorium nutzen, er ermöglicht es also, Studienmaterialien, Daten und Analyseskripte öffentlich zu teilen. Ein entsprechender institutioneller Wandel zeigt sich zum Beispiel darin, dass inzwischen über 100 Zeitschriften „Registered Reports" als Publikationsformat anbieten oder verlangen. Bei dieser Publikationsform wird anhand der Präregistrierung über eine spätere Veröffentlichung entschieden. Das Format gewährleistet damit zum einen eine Qualitätssicherung auf Ebene der Präregistrierung (also bereits bei der Studienplanung), zum anderen aber auch die dringend benötigte ergebnisunabhängige Publikation von Studien. Somit wirkt es auch einem Publikations-Bias entgegen. Eine deutliche Änderung der Publikationspraxis zeigt sich zudem darin, dass inzwischen hunderte von Zeitschriften aus unterschiedlichen Disziplinen die „Transparency and Openess Promotion guidelines" (Nosek et al., 2015) unterzeichnet haben. Sie werden also künftig in unterschiedlichen, standardisierten Kategorien dokumentieren, bis zu welchem Grade sie Transparenz in ihren Artikeln fördern oder verlangen (also z. B. öffentlich zugängliche Daten, unabhängig geprüfte statistische Analysen oder Präregistrierungen). Auch die Richtlinien der Forschungsförderung ändern sich in entsprechender Weise. Die Deutsche Forschungsgemeinschaft fordert seit 2009, dass Daten aus öffentlich-rechtlich geförderten Projekten nach Projektabschluss öffentlich zugänglich gemacht werden. Die Deutsche Gesellschaft für Psychologie (Schönbrodt et al., 2017) hat in der Folge eine Empfehlung zum Umgang mit Forschungsdaten herausgegeben, die es – unabhängig von der Förderung – für jede Forschungsarbeit zur Regel erhebt, dass die Daten nach einer Embargofrist veröffentlicht werden. Schließlich ist zumindest in der Psychologie das Bewusstsein dafür, dass Replikationen das Fundament einer empirischen Wissenschaft bilden, rasch gewachsen. Entsprechend werden Replikationen in jüngster Zeit drastisch häufiger veröffentlicht, diverse Zeitschriften haben eigene Rubriken und Publikationsformate für Replikationen eingerichtet.

Diese institutionellen Änderungen schaffen natürlich auch veränderte Anreize für das individuelle Verhalten des einzelnen Wissenschaftlers und der einzelnen Wissenschaftlerin. Allerdings handelt es sich hier um ein Wechselspiel: Anreize im Wissenschaftsbetrieb resultieren zu einem gro-

ßen Teil eher aus dem „kollektiven“ Handeln der *scientific community* als aus den Entscheidungen übergeordneter Institutionen. Der sich abzeichnende Wandel wird sich daher nur dann ausweiten und verstetigen lassen, wenn wir Kritik an unseren Arbeiten suchen und als Chance auf Verbesserung und Weiterentwicklung begreifen (statt als beängstigende Fehleraufdeckung), wenn wir die Arbeiten anderer mit angemessen kritischer Haltung rezipieren und zitieren, wenn wir als Gutachterinnen und Gutachter sowie Herausgeberinnen und Herausgeber Transparenz fördern und verlangen und wenn wir die durch Transparenz zu erwartende Qualitätsverbesserung auch in Berufungsentscheidungen honorieren. Die Aussichten auf einen solchen Wandel sind aktuell zweifellos deutlich besser als noch vor wenigen Jahren. Ohne diesen Wandel wird Evidenzbasierung allerdings auch in der Gesundheitskommunikation oft eher ein Marketingschlagwort bleiben als eine realistische Option auf besser begründete Entscheidungen und Empfehlungen.

Literaturverzeichnis

American Psychological Association. (2009). *Publication manual of the American Psychological Association*. Washington, D.C.: American Psychological Association.

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

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.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531-533.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433-36.

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

Fanelli, D. (2010). ‘‘Positive’’ Results Increase Down the Hierarchy of the Sciences. *PLoS ONE 5*(4): e10068.

Glöckner, A., Fiedler, S., & Renkewitz, F. (2018). Belastbare und effiziente Wissenschaft. Strategische Ausrichtung von Forschungsprozessen als Weg aus der Replikationskrise. *Psychologische Rundschau, 69*(1), 1-15.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biololgy*, *13*, e1002106.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices with Incentives for Truth-telling. *Psychological Science, 23*(5), 524-532.

Kühberger, A., Scherndl, T., & Fritz, A. (2014). Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLoS ONE*, *9*(9), e105825.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research. *Perspectives on Psychological Science,* 7, 537-542.

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality, 5,* 85-90.

Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior and Personality, 8,* 21-29.

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

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Palmer, A. R. (2000). Quasi-replication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. *Annual Review of Ecology and Systematics*, *31*, 441-480.

Popper, K. R. (2005). *Logik der Forschung* (11. Aufl.). Tübingen: Mohr Siebeck.

Popper, K. R. (2003). *Die offene Gesellschaft und ihre Feinde* (8. Aufl.), Tübingen: Mohr.

Prinz F., Schlange T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews: Drug Discovery, 10*, 712-712.

Schönbrodt, F., Gollwitzer, M., & Abele-Brehm, A. (2017). Der Umgang mit Forschungsdaten im Fach Psychologie: Konkretisierung der DFG-Leitlinien. *Psychologische Rundschau, 68*, 20-35.

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

Sedlmeier, P., & Renkewitz, F. (2018). Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler. München: Pearson.

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

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are We Wasting a Good Crisis? The Availability of Psychological Research Data after the Storm. *Collabra*, *1*(1), Art. 3.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE, 6*, e26828.

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