Erschienen in: Forschung und Lehre, Heft 2/2018 (hier erweitert um Literaturangaben)

Fallstricke statistischer Signifikanz

Wissenschaftliche Fachzeitschriften und die Replikationskrise

Edgar Erdfelder

In wissenschaftlichen Disziplinen, die mit statistischen Signifikanztests arbeiten, wird seit einigen Jahren über eine „Replikationskrise" intensiv diskutiert. Welche Rolle spielen Herausgeberentscheidungen von Fachzeitschriften in diesem Zusammenhang? Welche Maßnahmen wurden zur Überwindung der Replikationskrise implementiert?

Die *Open Science Collaboration* – ein Zusammenschluss von 270 psychologischen Wissenschaftlerinnen und Wissenschaftlern von mehr als 120 Universitäten weltweit – hat im Jahr 2015 die Resultate eines vielbeachteten Replikationsprojekts in der Zeitschrift *Science* publiziert (Open Science Collaboration, 2015). Demnach ließen sich nur knapp 40 Prozent der in drei führenden psychologischen Fachzeitschriften im Jahr 2008 veröffentlichten empirischen Befunde als „statistisch signifikant" replizieren. Spätestens seit diesem Zeitpunkt ist vielerorts von einer „Replikationskrise in der Psychologie" die Rede, obwohl diese Bezeichnung in doppelter Hinsicht irreführend ist: Eine genauere Analyse zeigt nämlich, dass nicht alle Teilbereiche der Psychologie in gleicher Weise vom Replizierbarkeitsproblem betroffen sind und dass Replizierbarkeitsprobleme letztlich in allen Wissenschaften zu finden sind, die zur Befundabsicherung statistische Signifikanztests verwenden, insbesondere in den Sozialwissenschaften, den Neurowissenschaften und in anderen biomedizinischen Wissenschaften (z.B. Button et al., 2013; Eklund, Nichoils & Knutsson, 2016; Vul, Harris, Winkielman, & Pashler, 2009). Spezifisch für die Psychologie ist somit nicht das Replizierbarkeitsproblem, sondern eher der offensive Umgang damit. Vor allem die *Open-Science*-Bewegung in der Psychologie zielt darauf ab, das Problem zu thematisieren und Randbedingungen herzustellen, die dabei helfen, Replizierbarkeitsprobleme zukünftig zu überwinden (vgl. Shrout & Rodgers, 2018; Nelson, Simmons, & Simonsohn, 2018)

Herausgeberentscheidungen und ihre Konsequenzen

Wissenschaftlichen Fachzeitschriften kommt hierbei eine zentrale Rolle zu. Bevor Instrumente hierfür vorgestellt werden, muss zunächst eingeräumt werden, dass Herausgeberentscheidungen in der Vergangenheit einen gehörigen Anteil an der Entstehung der Replikationskrise hatten. In dem Bestreben, nur solche Ergebnisse zur veröffentlichen, die vermeintlich statistisch abgesichert sind, wurden nahezu ausschließlich „statistisch signifikante" Resultate publiziert. Hierunter sind anhand einer Stichprobe von Probanden oder Versuchstieren beobachtete Effekte zu verstehen, die unter der Nullhypothese eines faktisch nicht vorhandenen Effekts eine geringe Wahrscheinlichkeit (in der Regel $p < .05$) aufweisen. Man bezeichnet die Rate tolerierter fälschlich positiver Entscheidungen auch als Signifikanzniveau α. Die Praxis, ein maximales α von 5% oder gar 1% zuzulassen, mag auf den ersten Blick plausibel erscheinen, hatte aber zwei gravierende Probleme zur Folge. Erstens waren nichtsignifikante

Resultate damit praktisch unpublizierbar, dies auch dann, wenn sie mit sorgfältiger Methodik und großen Stichproben ermittelt wurden. War also ein signifikanter Befund einmal in einer führenden Fachzeitschrift publiziert worden, dann war es schwer bis unmöglich, ihn als nicht-replizierbar zu entkräften, weil statistisch insignifikante Replikationsstudien in vorauseilendem Gehorsam des Autors entweder gar nicht erst zur Publikation eingereicht wurden oder aber vom Herausgebergremium führender Fachzeitschriften abgelehnt wurden. Robert Rosenthal hat das schon 1979 als *file drawer problem* in der psychologischen Forschung erkannt – das Phänomen, dass statistisch insignifikante Befunde in der Schreib-tischschublade des Labors verschwinden und für die Fachöffentlichkeit unsichtbar bleiben (Rosenthal, 1979).

Ein zweites Folgeproblem besteht darin, dass Forscher in dem Bestreben, publizierbare Ergebnisse zu erhalten, versucht sein können, Datensätze nach signifikanten Effekten zu durchsuchen, etwa durch Anwendung mehrerer Analysetechniken auf unterschiedliche abhängige Variablen oder auch durch Transformationen dieser Variablen (sog. *p-hacking*). Durch multiple Auswertungen derselben Daten wird ohne entsprechende Adjustierung des Signifikanzniveaus die Falsch-positiv-Rate jedoch drastisch erhöht, von nominell 5% auf deutlich über 60% wie man zeigen kann (Simmons, Nelson & Simonsohn, 2011). Die Kombination von erhöhter Falsch-positiv-Rate infolge *p-hackings* bei gleichzeitiger Verbannung insignifikanter Resultate aus den Fachzeitschriften dürfte neben anderen Einflussfaktoren maßgeblich für das Entstehen der Replikationskrise verantwortlich gewesen sein (Ulrich et al., 2016).

Vier Maßnahmen

Die Diagnose der Determinanten der Replikationskrise legt vor allem vier Maßnahmen zu ihrer Überwindung nahe, die von den Herausgebergremien zahlreicher psychologischer Fachzeitschriften in den letzten Jahren auch implementiert werden: (1) Offenheit für direkte Replikationsstudien, (2) Einführung von präregistrierten Forschungsberichten, (3) Optimierung von Versuchsplanung und statistischer *Power* und (4) Verbesserung metaanalytischer Forschungssynthesemethodik.

(1) Es muss sichergestellt sein, dass sauber geplante direkte Replikationsstudien zu wichtigen Originalbefunden unabhängig von ihrem Ergebnis publizierbar sind. Viele namhafte Fachzeitschriften folgen seit einigen Jahren diesem Prinzip (s. z.B. Lindsay, 2015). Die Sichtbarkeit der Ergebnisse lege artis durchgeführter Replikationsstudien ist wichtig, weil sie hilft, mögliche Replikationsprobleme zu diagnostizieren und zu lokalisieren (Erdfelder & Ulrich, 2018).

(2) Präregistrierte Forschungsberichte sollten als optionales Publikationsformat eingeführt werden. Hierbei erfolgt die Publikation idealerweise in drei Schritten (Greve, Bröder & Erdfelder, 2013): (a) Vor der Datenerhebung wird zunächst eine wissenschaftliche Fragestellung schriftlich dargelegt und begründet, verbunden mit der detaillierten Schilderung eines Experiments, das die Frage beantworten soll. Die Autoren versichern dem Herausgeber durch Unterschrift, dass die entsprechende Untersuchung bislang nicht durchgeführt wurde, auch nicht in Teilen. Dieses Manuskript wird dann einer Begutachtung zugeführt, die in eine Herausgeberentscheidung einmündet. Wird die vorgeschlagene Untersuchung akzeptiert, ist damit zugleich die Publikation unabhängig vom Ergebnis garantiert. (b) Die Autoren erhalten im zweiten Schritt vom Herausgeber die Aufforderung, die Untersuchung genau wie beschrieben durchzuführen und innerhalb einer vorgegebenen Zeitspanne das komplette Manuskript mit

Ergebnisbericht und Diskussion einzureichen. (c) Der letzte Schritt besteht in der Begutachtung des Gesamtmanuskripts, bei der es jedoch nicht mehr um die (bereits erfolgte) Annahmeentscheidung, sondern nur noch um die Korrektur etwaiger Fehler und mögliche Verbesserungen geht. Viele Fachzeitschriften haben dieses Publikationsformat bereits eingeführt, so als eine der ersten Zeitschriften auch *Experimental Psychology* (Stahl, 2014).

(3) Neben Replikationsstudien und präregistrierten Forschungsberichten müssen Originalarbeiten selbstverständlich weiterhin die zentrale Rolle in Fachzeitschriften spielen. Originalarbeiten bedürfen jedoch einer sorgfältigeren Versuchsplanung als das in der Vergangenheit oftmals der Fall war. Bei Verwendung statistischer Tests gilt dies etwa für den Stichprobenumfang bzw. die Anzahl der Beobachtungen, welche dem Hypothesentest zugrunde liegen. Diese Kennwerte beeinflussen (neben anderen Einflussfaktoren) die *Power* des Tests, d.h. die Wahrscheinlichkeit $1-\beta$, bei Vorliegen eines wahren Effekts bestimmter Mindestgröße auch ein signifikantes Analyseergebnis zu erhalten. Nur wenn gezeigt werden kann, dass ein Test bei einem Signifikanzniveau von z.B. $\alpha = 5\%$ zugleich eine hohe *Power* von z.B. $1-\beta = 95\%$ aufweist, kann die betreffende Untersuchung zwischen Nulleffekten und wahren Effekten zuverlässig trennen, was eine Voraussetzung für die Aussagekraft der Untersuchung ist. Auch scheinbar beeindruckende Serien von beispielsweise fünf erfolgreichen konzeptuellen Replikationen einer Hypothese verlieren ja schnell ihre Überzeugungskraft, wenn man erfährt, dass jede einzelne dieser Studien eine *Power* von nur $1-\beta = .30$ hatte (Schimmack, 2012). Selbst wenn in allen Fällen der erwartete Effekt wirklich vorlag, hätte das Ergebnis eine Wahrscheinlichkeit von lediglich $.30^5 = 0.00243$. Das Ergebnis von fünf Signifikanzen in angeblich nur fünf durchgeführten Tests scheint unter diesen Randbedingungen also „zu gut um wahr zu sein“, um es mit den Worten von Gregory Francis auszudrücken (Francis, 2012a, 2012b). Als Konsequenz aus derartigen Einsichten wird von vielen Fachzeitschriften inzwischen eine Fallzahlbegründung mittels A-priori-*Power*-Berechnungen verlangt.

(4) Forschungssynthese – die Integration von Befunden zu eine bestimmten Forschungsfrage – gewinnt im Kontext einer mit statistischen Tests arbeitenden Wissenschaft besondere Bedeutung. Dies liegt daran, dass perfekte Replizierbarkeit statistischer Testergebnisse nicht erwarten werden kann. Aufgrund der tolerierten Rate fälschlich positiver Befunde – sei sie nun 5% oder 1% – und der nicht perfekten *Power* wird es immer Einzelbefunde geben, die sich als nicht-replizierbar erweisen. Klarheit über den empirischen Status einer Forschungshypothese erhält man daher nur durch metaanalytische Integration von verschiedenen Primärstudien. Standardverfahren der Metaanalyse sind seit Dekaden bekannt, gehen aber von der problematischen Prämisse aus, dass die vorliegenden (publizierten) Primärstudien eine repräsentative Auswahl aller durchgeführten Studien sind, was zu bezweifeln ist. Benötigt werden daher neue metaanalytische Techniken, die Verzerrungen durch *Publication Bias* (und auch *p-hacking*, vgl. Simonsohn, Nelson, & Simmons, 2014; Ulrich & Miller, 2017) diagnostizieren und möglichst auch statistisch korrigieren können (für einen Überblick vgl. Ulrich, Miller, & Erdfelder, 2018). Die Förderung dieser Forschungsrichtung haben sich ebenfalls einige Fachzeitschriften zum Ziel gesetzt, darunter auch die *Zeitschrift für Psychologie*, die jährlich ein Themenheft „*Hotspots in Psychology*“ zu Metanalysen und ihren methodischen Verbesserungen herausgibt (Erdfelder & Bosnjak, 2016).

Exploratorische Forschung

Fachzeitschriften müssen natürlich auch in Zukunft ein Forum für innovative, ggf. exploratorisch gewonnen Ideen bieten, deren empirisches Fundament den o.g. strengen Anforderungen (noch) nicht standhält. Möglicherweise entstehen aus derartigen Ideen ja bedeutsame theoretische, methodische oder technologische Weiterentwicklungen (Fiedler, Kutzner & Krüger, 2012). Wichtig ist allerdings, dass exploratorisch gewonnene Ideen als solche kenntlich gemacht und nicht als durch gezielte Hypothesentests abgesichert vorgetäuscht werden. Ein wichtiges Instrument hierfür ist die Offenlegung aller Datenanalysen und ggf. Simulationen, die im Kontext einer Forschungsarbeit durchgeführt wurden. Dies kann beispielsweise in Form von Präregistrierungsdiensten, elektronischen Supplements zu Publikationen oder auch durch Dokumentation des Forschungsprozesses in digitalen Repositorien geschehen, wie sie durch den *Open Science Framework* (www.OSF.io) oder auch das *Leibniz-Zentrum für Psychologische Information und Dokumentation* (ZPID, www.leibniz-psychology.org) bereitgestellt werden.

Autor:

Edgar Erdfelder ist Professor für Psychologie an der Universität Mannheim. Von 2007 bis 2010 hatte er das Amt des Editor-in-Chief von *Experimental Psychology* inne. Seit 2017 ist er Editor-in-Chief der *Zeitschrift für Psychologie*.

Weitere Informationen zum Autor:

http://psycho3.uni-mannheim.de/Personen/Prof.%20Dr.%20Edgar%20Erdfelder/

Literatur:

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365-376. doi: 10.1038/nrn3502.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Science*s, 113 (28), 7900-7905, doi: 10.1073/pnas.1602413113.

Erdfelder, E. & Bosnjak, M. (2016). Hotspots in psychology: A new format for special issues of the Zeitschrift für Psychologie. *Zeitschrift für Psychologie*, *224*, 141-144. doi: 10.1027/2151-2604/a000249.

Erdfelder, E. & Ulrich, R. (2018). Zur Methodologie von Replikationsstudien. *Psychologische Rundschau*, *69*, 3-21. doi: 10.1026/0033-3042/a000387.

Fiedler, K., Kutzner, F., & Krueger, J. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*, 661–669. DOI: 10.1177/17456916/2462587.

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

Francis, G. (2012b). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151–156. doi: 10.3758/s13423-012-0227-9.

Greve, W., Bröder, A., & Erdfelder, E. (2013). Result-blind peer-reviews and editorial decisions: A missing pillar of scientific culture. *European Psychologist*, *18*, 286–294, doi: 10.1027/1016-9040/a000144.

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science*, *26*, 1827-1832. doi: 10.1177/0956797615616374.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511-534. doi: 10.1146/annurev-psych-122216-011836.

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

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

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

Shrout, P. E. & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication sciences. *Annual Review of Psychology*, *69*, 487-510. doi: 10.1146/annurev-psych-122216-011845.

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

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

Stahl, C. (2014). Experimental Psychology: Toward reproducible research. *Experimental Psychology*, *61*, 1-2. doi: 10.1027/1618-3169/a000257.

Ulrich, R., Erdfelder, E., Deutsch, R., Strauß, B., Brüggemann, A., Hannover, B., Tuschen-Caffier, B., Kirschbaum, C., Blickle, G., Möller, J., & Rief, W. (2016). Inflation von falsch-positiven Befunden in der psychologischen Forschung: Mögliche Ursachen und Gegenmaßnahmen. *Psychologische Rundschau*, *67*, 163-174. doi: 10.1026/0033-3042/a000296.

Ulrich, R. & Miller, J. (2017). Some properties of *p*-curves, with an application to gradual publication bias. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000125.

Ulrich, R., Miller, J., & Erdfelder, E. (2018). Effect size estimation from *t*-statistics in the presence of publication bias: A brief review of existing approaches with some extensions. *Zeitschrift für Psychologie*, *226*, 56-80. doi:10 .1027/2151-2604/a000319.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzling high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274-290. DOI: 10.1111/j.1745-6924.2009.01125.x.

16131 Zeichen (mit Leerzeichen)