

Aus der Forschung

Was ist Publikationsbias?

Bei Publikationsbias bestimmt das Ergebnis, ob eine Studie erscheint. In einer bekannten Studienmenge wurden die meisten Nullergebnisse nie ausgearbeitet.

Autor	Veröffentlicht	Lesezeit	Wörter
Sinan Isoglu, MBA (Quantic) Führungskraft für Umsatzwachstum, Dozent und Doktorand	18. Oktober 2026	17 Min.	3.651

Management Summary Publikationsbias ist die Auswahl von Studien zur Veröffentlichung nach Richtung oder Signifikanz ihrer Ergebnisse; die übrigen bleiben in der Schublade. Der Leitfaden definiert ihn (Head et al.; Munafò et al.; Franco et al.), grenzt ihn vom P-Hacking ab, berichtet Sterlings Auszählung von vier psychologischen Zeitschriften, in denen 286 von 294 Artikeln mit Signifikanztests die Nullhypothese verwarfen, und die Nachverfolgung von 221 Umfrageexperimenten (Franco et al.), in der 64,6 Prozent der Nullergebnisse nie ausgearbeitet wurden, erklärt Rosenthals Toleranz für Nullergebnisse, das Fail-Safe-N, und was sie offenlässt, und bietet ein Testregister für ein kommerzielles Analyseteam. Die Quellen sind eine Auszählung veröffentlichter Artikel, die Nachverfolgung eines Forschungsprogramms, eine Methodennotiz, eine Vertriebs-Meta-Analyse, die sie anwendet, und zwei Perspektivbeiträge.

Inhalt	§ 1	Was ist Publikationsbias?	3
	§ 2	Was zählte Sterling in vier psychologischen Fachzeitschriften?	5
	§ 3	Wie groß war die Schublade in einer bekannten Menge von Studien?	6
	Abbildung 1	Was aus 221 Umfrageexperimenten wurde, nach der Stärke ihrer Ergebnisse	7
	§ 4	Wer füllte die Schublade: Redaktionen oder Forschende?	8
	§ 5	Wie viele verborgene Studien würden ein Ergebnis kippen?	10
	§ 6	Was lässt ein großes Fail-Safe-N offen?	11
	§ 7	Wie sieht eine Schublade im Unternehmen aus?	12
	Abbildung 2	Ein Testregister	13
	§ 8	Wie weit tragen diese Befunde bei Testergebnissen im Unternehmen?	14

Referenzen

15

Was ist Publikationsbias?

§ 1 von 8

Gemeint ist, dass Studien danach zur Veröffentlichung ausgewählt werden, was sie gefunden haben. Franco et al. (2014) beginnen mit einer Definition, die sie einem früheren Aufsatz entnehmen (Nr. 1 ihres Literaturverzeichnisses): Publikationsbias liegt vor, wenn die Veröffentlichung von Studienergebnissen von deren Richtung oder Signifikanz abhängt („the direction or significance of the findings“). Eine Form davon besteht nach ihren Worten darin, dass statistisch signifikante Ergebnisse bei gleicher Forschungsqualität eher veröffentlicht werden als nicht signifikante. Eine solche Auswahl macht es, so die Autoren, wahrscheinlicher, dass veröffentlichte Ergebnisse Fehler erster Art sind, also falsch-positive Befunde, statt Schätzungen des wahren Effekts, und treibt die veröffentlichten Effektstärken nach oben.

Für den Namen des Ortes, an dem der Rest bleibt, zitieren Franco et al.¹, Head et al.² und Munafò et al.³ alle drei Rosenthal (1979)⁴: In der extremen Lesart des Schubladenproblems bestehen die Zeitschriften aus den 5 Prozent der Studien, die Fehler erster Art zeigen, während in den Schubladen die 95 Prozent mit nicht signifikanten Ergebnissen liegen („The extreme view of the ‘file drawer problem’ is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results“, S. 638). Er beginnt mit dem Hinweis, dass Forschende und Statistiker seit Langem vermuteten, die in den Verhaltenswissenschaften veröffentlichten Studien seien eine verzerrte Stichprobe der tatsächlich durchgeführten.

¹ Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502–1505.

² Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3), e1002106.

³ Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 0021.

⁴ Rosenthal, R. (1979). The „file drawer problem“ and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641.

Der Begriff wird auf zweierlei Weise verwendet. Head et al. (2015)² unterscheiden zwei anerkannte Formen von Publikationsbias, die von Forschenden ausgehen („two widely recognized types of

researcher-driven publication bias“, S. 1): die Selektion, den Schubladeneffekt („the file drawer effect“), bei dem Studien mit nicht signifikanten Ergebnissen seltener veröffentlicht werden, und die Inflation, also das P-Hacking. Munafò et al. (2017)³ verwenden den Begriff nur für die erste Form: Publikationsbias, auch Schubladenproblem genannt („also known as the file drawer problem“), bezeichne die Tatsache, dass weit mehr Studien durchgeführt als veröffentlicht werden („refers to the fact that many more studies are conducted than published“, S. 3). Studien mit positiven und neuen Ergebnissen würden eher veröffentlicht als Studien mit negativen Ergebnissen oder Replikationen, sodass die veröffentlichte Literatur stärkere Evidenz zeige, als es sie tatsächlich gibt („stronger evidence for findings than exists in reality“, S. 3).

In diesem Leitfaden heißt Publikationsbias die erste Form: Ob eine Studie veröffentlicht wird, hängt von der Richtung oder der statistischen Signifikanz ihres Ergebnisses ab, sodass die veröffentlichten Studien eine ausgewählte Stichprobe der durchgeführten sind und die übrigen in der Schublade bleiben. Die Auswahl können Autorinnen und Autoren treffen, die eine Studie nicht ausarbeiten oder nicht einreichen, oder Redaktionen und Gutachtende, die sie ablehnen, wie Franco et al.¹ festhalten; die Definition hängt weder davon ab, wer auswählt, noch von einer Absicht. P-Hacking, nach der Lesart von Head et al.² die zweite Form, wirkt innerhalb einer Studie, nämlich darauf, welche Analyse berichtet wird, und bleibt einem eigenen Leitfaden überlassen. Publikationsbias ist ein Auswahlmechanismus auf der Ebene ganzer Studien; der Leitfaden zum Selektionsbias behandelt allgemein, wie der Mechanismus, der Einheiten in eine Stichprobe bringt, die Antwort verändern kann.

Was zählte Sterling in vier psychologischen Fachzeitschriften?

§ 2 von 8

Eine fast lückenlose Folge signifikanter Ergebnisse. Sterling (1959)⁵ nahm alle Hefte eines Jahrgangs (1955, bei einer der Zeitschriften 1956) von vier psychologischen Forschungszeitschriften, die er zufällig aus vier großen Teilgebieten der Psychologie ausgewählt hatte. Von 362 Forschungsberichten verwendeten 294 Signifikanztests. Von diesen 294 verwarfen 286 die Nullhypothese, 8 taten es nicht: 97,28 und 2,72 Prozent. Maßstab war das 5-Prozent-Niveau, bis auf zwei Berichte auf dem 10-Prozent-Niveau, die Sterling als Verwerfung zählte, weil ihre Autoren es ausdrücklich so angegeben hatten (S. 32). Keiner der 362 Berichte war die Replikation eines bereits veröffentlichten Experiments. Zur Kontrolle zog er 100 Forschungsartikel zufällig aus den Psychological Abstracts der Jahre 1952 bis 1957: 94 berichteten positive Ergebnisse, 5 negative, und einer war eine Replikation (S. 31 bis 32).

⁵ Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance, or vice versa. *Journal of the American Statistical Association*, 54(285), 30–34.

Er nahm nicht an, dass ausdrückliche Regeln der Redaktionen das Muster erzeugten. Der größte einzelne Faktor, so seine Vermutung, könne eine stillschweigende Übereinkunft unter den Autoren sein, und die Bezeichnung Veröffentlichungspolitik verwendete er weitgehend der Bequemlichkeit halber; was die Tabellen zeigen, ist, was gedruckt wird. Sein Argument gilt den Folgen. Wo Forschung mit nicht signifikanten Ergebnissen nicht veröffentlicht wird, wissen andere nicht, dass sie durchgeführt wurde; sie kann daher unabhängig wiederholt werden, bis zufällig ein signifikantes Ergebnis auftritt, ein Fehler erster Art, und dieses Ergebnis erscheint; signifikante Ergebnisse werden in solchen Gebieten selten durch unabhängige Replikation überprüft; also kann die Literatur eines Gebiets zu einem erheblichen Teil aus falschen Schlüssen bestehen (S. 30). Für die Lesenden lasse sich das vom Autor angegebene Risiko daher nicht beim Wort nehmen („The risk stated by the author cannot be accepted at its face value“), sobald die Schlüsse gedruckt sind: Die angegebenen 5 Prozent gelten für ein Experiment, während die Lesenden womöglich das eine von mehreren ähnlichen Experimenten vor sich haben, das zufällig einen großen Unterschied fand (S. 34).

Sterlings Prämissen lauteten: Experimente, die die Nullhypothese verwerfen, werden mit größerer Wahrscheinlichkeit gedruckt als die

übrigen; ein gedrucktes Experiment wird sehr wahrscheinlich nicht repliziert; und es werden weit mehr Experimente durchgeführt, als in den Zeitschriften erscheinen. Er nannte sie substanziell und schrieb, starke Belege für sie seien über das hier Gezeigte hinaus schwer zu finden („beyond that given here, is hard to come by“, S. 33). Er fügte hinzu, schon ein paar Minuten Blättern in Zeitschriften der Biologie, Chemie, Medizin, Physiologie und Soziologie zeigten dieselben Gepflogenheiten in anderen Wissenschaften („show that the same usages are widespread“, S. 34); das war ein Eindruck, keine Zählung. Eine Auszählung von Zeitschriften zeigt, was gedruckt wurde; sie kann nicht zeigen, was durchgeführt und weggelassen wurde. Der Leitfaden zur Reproduzierbarkeit trennt das Nachrechnen eines Ergebnisses aus seinen eigenen Eingaben von der Replikation mit neuen Daten, also von dem Schritt, den Sterling vermisste.

Wie groß war die Schublade in einer bekannten Menge von Studien?

§ 3 von 8

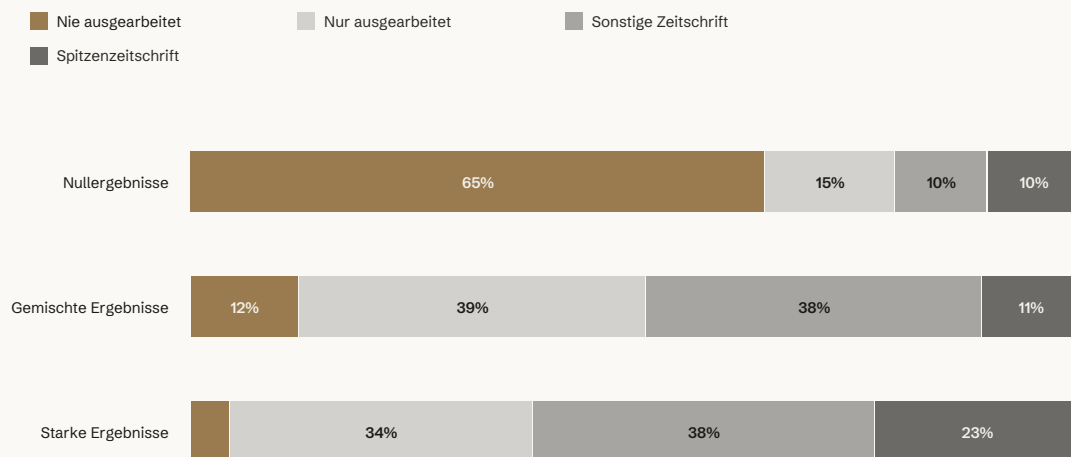
Mehr als die Hälfte der Studien war in keiner Zeitschrift erschienen, und die meisten Nullergebnisse wurden nicht einmal ausgearbeitet. Franco et al. (2014)¹ fanden, was Sterling fehlte: eine Menge von Studien, deren unveröffentlichte Mitglieder sich zählen ließen. Sie nutzten TESS (Time-sharing Experiments in the Social Sciences), ein von der US-amerikanischen National Science Foundation gefördertes, 2002 gegründetes Programm, in dem Forschende Umfrageexperimente vorschlagen, die an landesweit repräsentativen Stichproben durchgeführt werden; TESS lässt die Vorschläge begutachten und vergibt die Mittel im Wettbewerb. Weil TESS eine ausreichende Teststärke verlangt, verfehlte eine Studie die Signifikanz nicht einfach wegen einer zu kleinen Stichprobe (S. 1 der Science-Express-Fassung).

Sie verfolgten 221 Studien, die zwischen 2002 und 2012 durchgeführt wurden, 89 Prozent der 249 im Archiv; sieben als Buchkapitel veröffentlichte und 21 mit unklarem Status oder Ergebnis blieben außen vor. Sie suchten nach jeder Veröffentlichung und schrieben die Autorinnen und Autoren von über 100 Studien an, deren Spur sie nicht fanden. Die Ergebnisse wurden danach eingeordnet, wie die Autorinnen und Autoren selbst sie beschrieben: stark, wenn alle oder die meisten Hypothesen gestützt wurden, null, wenn alle oder die meisten nicht gestützt wurden, sonst gemischt. 41 Prozent der Studien berichteten starke Ergebnisse, 37 Prozent gemischte und 22 Prozent Nullergebnisse; 80 Prozent waren zu einem Manuskript ausgearbeitet und 48 Prozent in wissenschaftlichen Zeitschriften veröffentlicht (S. 1

bis 2 der Science-Express-Fassung). Abbildung 1 zeigt, was aus jeder Gruppe wurde.

Abbildung 1 Was aus 221 Umfrageexperimenten wurde, nach der Stärke ihrer Ergebnisse

Von den Studien mit Nullergebnissen wurden 64,6 Prozent nie ausgearbeitet, von denen mit starken Ergebnissen 4,4 Prozent.



Jeder Balken ist eine Ergebnisgruppe und ergibt 100 Prozent. Die zweite Kategorie der Legende umfasst Studien, die ausgearbeitet, aber nicht veröffentlicht wurden; die dritte und vierte unterscheiden Veröffentlichungen in sonstigen Zeitschriften und in Spitzenzeitschriften. Die Balken umfassen 48, 82 und 91 Studien, Summen dieses Leitfadens aus den analysierten Zahlen in Tabelle 2 (S. 6 der Science-Express-Fassung). Die Beschriftungen der Segmente geben den Anteil am Balken an, auf ganze Prozent gerundet; die Tabelle nennt eine Dezimalstelle, ihre Spalte der starken Ergebnisse summiert sich auf 100,1, daher ist das Segment mit 38,5 mit 38 beschriftet, und die 4,4 Prozent der nie ausgearbeiteten starken Ergebnisse sind für eine Beschriftung zu schmal. Die Ergebnisse sind nach der Beschreibung der Autorinnen und Autoren selbst eingeordnet.

Quelle Franco et al. (2014), Tabelle 3, S. 6 der vorliegenden Science-Express-Fassung (S. 1 bis 6, online veröffentlicht am 28. August 2014): Spaltenprozent der 221 TESS-Umfrageexperimente aus den Jahren 2002 bis 2012, nach der von den Autorinnen und Autoren berichteten Stärke der Ergebnisse und nach dem Stand der Veröffentlichung. Buchkapitel und Studien mit unklarem Status oder Ergebnis sind wie in der Analyse der Autoren ausgeschlossen. Werte der Autoren, keine eigene Neuschätzung.

Starke Ergebnisse wurden mit einer um 40 Prozentpunkte höheren Wahrscheinlichkeit veröffentlicht als Nullergebnisse und mit einer um 60 Prozentpunkte höheren Wahrscheinlichkeit ausgearbeitet. Rund 20 Prozent der Studien mit Nullergebnissen erschienen im Druck, gegenüber etwa 60 Prozent der Studien mit starken und 50 Prozent

derer mit gemischten Ergebnissen; Studien mit Nullergebnissen machten mehr als 20 Prozent der Stichprobe aus, aber weniger als 10 Prozent der darauf beruhenden veröffentlichten Artikel. Ein Test auf Unabhängigkeit zwischen Ergebnis und Stand der Veröffentlichung wird klar verworfen ($\chi^2(6) = 80,3, p < 0,001$; S. 1, 2 und 6 der Science-Express-Fassung).

Wer füllte die Schublade: Redaktionen oder Forschende?

§ 4 von 8

In dieser Menge vor allem die Forschenden: Die Auswahl nach dem Ergebnis geschah beim Ausarbeiten, vor der Begutachtung. Nur wenige Nullergebnisse gelangen überhaupt in die Begutachtung, schreiben Franco et al.¹ („Few null findings ever make it to the review process“). 64,6 Prozent der Studien mit Nullergebnissen wurden nie ausgearbeitet, und unter den ausgearbeiteten Studien fanden die Autoren keinen signifikanten Zusammenhang zwischen der Stärke der Ergebnisse und dem Stand der Veröffentlichung (S. 2 der Science-Express-Fassung). Ihr Fazit lautet, dass Autorinnen und Autoren Nullergebnisse nicht ausarbeiten und nicht einreichen („authors do not write up and submit null findings“, S. 1 der Science-Express-Fassung). Das deckt sich mit den früheren Arbeiten, die sie anführen: Auswertungen klinischer Register und der bei medizinischen Kongressen eingereichten Abstracts finden kaum oder keine Benachteiligung von Studien mit Nullergebnissen durch die Redaktionen („little to no editorial bias“), und dass nicht veröffentlicht wird, hängt am stärksten mit der Einschätzung der Autorinnen und Autoren zusammen, Nullergebnisse seien uninteressant (S. 1 der Science-Express-Fassung).

Sie fragten nach den Gründen. Von 26 ausführlichen Antworten per E-Mail von Forschenden, deren Studien Nullergebnisse brachten und die kein Manuskript schrieben, gaben 15 an, das Projekt aufgegeben zu haben, weil sie Nullergebnissen keine Aussicht auf Veröffentlichung zutrauten, selbst wenn sie die Ergebnisse persönlich interessant fanden; 9 stellten das Ausarbeiten zurück und wandten sich anderen Projekten zu. Zwei Autoren, deren Studien nicht wie erhofft ausgingen („didn't work out“), veröffentlichten später Artikel, die ihre ursprünglichen Hypothesen stützten, mit Befunden aus kleineren Gelegenheitsstichproben (S. 2 der Science-Express-Fassung). Aus ihrer Korrespondenz mit den Autorinnen und Autoren vieler Experimente mit Nullergebnissen schließen Franco et al.¹, dass manche Forschende mit der Ablehnung solcher Manuskripte rechnen („the rejection of such papers“), während viele schlicht das Interesse an

Projekten verlieren, die sie für gescheitert halten (S. 2 der Science-Express-Fassung). Die Entscheidung, nicht auszuarbeiten, trafen also die Autorinnen und Autoren, doch bei manchen nahm sie vorweg, wie Redaktionen und Gutachtende entscheiden würden.

Die Autoren betonen, dass die TESS-Studien eindeutig keine Zufallsstichprobe der sozialwissenschaftlichen Forschung sind („clearly not a random sample“), und halten es für unwahrscheinlich, dass Publikationsbias anderswo schwächer ist: Die TESS-Daten waren aufwendig zu gewinnen, sodass ihre Autorinnen und Autoren eher stärker motiviert sein sollten, Nullergebnisse zu veröffentlichen. Die Stärke jedes Ergebnisses beruht auf den eigenen Angaben der Autorinnen und Autoren, und nach der Sensitivitätsanalyse der Verfasser halten die Befunde selbst drastischen und unrealistischen Fehlklassifikationsraten stand. Bereits vorgeschlagene Abhilfen wie die zweistufige Begutachtung, Pläne für die Analyse vor der Datenerhebung und die Präregistrierung sollten nach ihren Worten durch Anreize ergänzt werden, nicht signifikante Ergebnisse nicht in Schubladen zu vergraben („bury insignificant results in file drawers“, S. 2 der Science-Express-Fassung). Munafò et al. (2017)³ sagen Ähnliches über die stärkste Form der Präregistrierung, die die Studie mit der Zusage registriert, die Ergebnisse öffentlich zu machen, und Design und Analyse vorab festlegt: Grundsätzlich wirke sie dem Publikationsbias entgegen, weil alle Forschung auffindbar werde, ob sie am Ende veröffentlicht wird oder nicht („In principle, this addresses publication bias by making all research discoverable, whether or not it is ultimately published“, S. 3).

Wie viele verborgene Studien würden ein Ergebnis kippen?

§ 5 von 8

Rosenthal (1979)⁴ stellte die Frage, die die Verfasser einer Übersichtsarbeit beantworten können, ohne die Schublade zu finden: Für kein Forschungsgebiet lasse sich sagen, wie viele Studien durchgeführt, aber nie berichtet wurden („For any given research area, one cannot tell how many studies have been conducted but never reported“, S. 638). Statt sie zu zählen, berechnete er, wie viele Studien mit durchschnittlich null Effekt in den Schubladen liegen müssten, bis das kombinierte Ergebnis der vorliegenden Studien auf gerade noch signifikant fällt, bei $p = 0,05$. Er nannte diese Zahl die Toleranz für künftige Nullergebnisse; brächten schon wenige weitere Nullergebnisse das kombinierte Ergebnis so weit herunter, sei der Befund der Bedrohung durch die Schublade nicht gewachsen („is not resistant to“, S. 639). Verbeke et al. (2011)⁶ berechneten sie in einer Meta-Analyse der Faktoren, die die Leistung im Verkauf bestimmen, als Fail-Safe-N, um zu prüfen, wie plausibel ein Schubladenproblem ist: als die Zahl nicht aufgefundener Studien mit Nullergebnissen, die den kumulierten Effekt über die Studien hinweg bis zur Nichtsignifikanz senken würde („with null results needed to reduce the cumulated effect across studies to the point of non-significance“, S. 413).

⁶ Verbeke, W., Dietz, B., & Verwaal, E. (2011). Drivers of sales performance: A contemporary meta-analysis. Have salespeople become knowledge brokers? *Journal of the Academy of Marketing Science*, 39(3), 407–428.

Die Rechnung beruht auf der Addition von Z-Werten. Der p-Wert jeder Studie wird in einen Wert Z der Standardnormalverteilung umgerechnet; die Z-Werte der k Studien werden addiert und durch die Wurzel aus k geteilt, was ein kombiniertes Z ergibt. Setzt man dieses kombinierte Z, nach Hinzunahme von X weiteren Studien mit $Z = 0$, gleich 1,645 und löst auf, erhält man Rosenthals alternative Formel $X = (\text{sum } Z)^2 / 2,706 - k$, wobei 2,706 das Quadrat von 1,645 ist; seine Gleichung 3 sagt dasselbe mit dem mittleren Z (S. 639). In seinem ersten Beispiel hatten 94 Experimente zu zwischenmenschlichen sich selbst erfüllenden Prophezeiungen ein mittleres Z von 1,014, und 3.263 Studien mit durchschnittlich null Effekt hätten in Schubladen liegen müssen, um sie zu kippen; eine spätere Zusammenfassung von 311 Studien mit einem mittleren Z von 1,180 ergab 49.457. In der Gegenrichtung haben 15 Studien mit durchschnittlich $Z = 0,50$ ein kombiniertes p von 0,026, und 6 Studien mit Nullergebnis in einer Schublade würden es nicht signifikant machen; 2 Studien mit

durchschnittlich $Z = 2,00$ ergeben ein p von etwa 0,002, und 4 würden genügen (S. 639 bis 640).

Ab welcher Zahl verborgene Studien unplausibel werden, ließ er offen: Feste Richtlinien dafür, welche Zahl nicht aufgefundener oder unveröffentlichter Studien unwahrscheinlich ist, ließen sich noch nicht angeben, schrieb er. In manchen Gebieten könnten 100 oder sogar 500 unveröffentlichte Studien plausibel sein, in anderen erscheinen schon 10 oder 20 unwahrscheinlich. Vorsichtig schlug er vor, ein Ergebnis als widerstandsfähig zu betrachten, wenn die Toleranz $5k + 10$ erreicht, ein Niveau, das er für konservativ, aber vernünftig hielt, und dass Übersichtsarbeiten die Toleranz neben jedem kombinierten p angeben (S. 640).

Was lässt ein großes Fail-Safe-N offen?

§ 6 von 8

Drei Dinge, nach der Lesart dieses Leitfadens. Erstens betrifft es die Signifikanz, nicht die Größe. Rosenthals Frage ist, ob der Effekt „real“ ist, also kein Fehler erster Art („not a Type I error“, S. 639), und Verbeke et al.⁶ bestimmen die Zahl über den Punkt der Nichtsignifikanz. Darüber, wie stark die veröffentlichten Effektstärken die wahren überzeichnen, sagt es nichts, und genau diese Folge nennen Franco et al.¹, wenn sie schreiben, eine Auswahl nach der Signifikanz verzerre die Effektstärken nach oben („biasing effect sizes upwards“, S. 1 der Science-Express-Fassung): Eine Literatur kann Tausende verborgener Studien brauchen, um ihre Signifikanz zu verlieren, und dennoch zu große Effekte berichten.

Zweitens beruht es auf Annahmen, die Rosenthal nennt. Für die Studien in der Schublade wird ein durchschnittliches Nullergebnis angenommen, $Z = 0$, und seine Gleichungen setzen voraus, dass jede der k Studien von den übrigen unabhängig ist (S. 639). Sprächen die unveröffentlichten Studien stattdessen gegen den Effekt, würden weniger von ihnen genügen, um ihn zu kippen; das folgt aus der Rechnung und wurde in dem Artikel nicht geprüft. Drittens ist die Schwelle eine Ermessensfrage: Rosenthal gibt keine feste Richtlinie und bietet $5k + 10$ als ein vernünftiges Niveau an. Der Leitfaden zur Quellenprüfung führt eine Zitation auf Identität, Fassung und vorliegenden Text des Werks zurück; Publikationsbias betrifft dagegen, welche Werke überhaupt in den Druck gelangten, sodass eine Quelle jeden Schritt dort bestehen und trotzdem zu den Studien gehören kann, die für den Druck ausgewählt wurden.

Wie sieht eine Schublade im Unternehmen aus?

§ 7 von 8

Etwa wie ein Quartalsbericht, der nur die Tests aufführt, die gewonnen haben. Die Quellen handeln von Fachzeitschriften; dieser Abschnitt ist daher die Überlegung des Autors. Ein Analyseteam in Pricing oder Vertrieb kann in einem Jahr viele Tests durchführen: Preispunkte nach Region, Rabattstaffeln, Verlängerungsbedingungen, ein neues Vertriebs-Playbook in einigen Gebieten. Jemand entscheidet, welche davon in den Bericht kommen. Werden Tests mit Nullergebnissen weder ausgearbeitet noch gezeigt, wird der Bericht zu einem Dashboard der Gewinner, und die eigene Bilanz des Teams trägt dieselbe Auswahl, die Franco et al.¹ bei Forschenden fanden, an derselben Stelle: beim Ausarbeiten, über das die Analystin oder der Analyst entscheidet und das, wie bei TESS, vorwegnehmen kann, was der Bericht zeigen soll. Auch Sterlings Kette kann in einem Team ablaufen: Ein Test mit Nullergebnis ohne Notiz wird vielleicht später erneut durchgeführt, und berichtet wird die Wiederholung, die die Schwelle zufällig überschreitet.

Ein Unternehmen hat Lesenden einer Fachzeitschrift eines voraus: Es kann seine eigene Schublade zählen. TESS gab Franco et al.¹ eine bekannte Gesamtheit von Studien; ein Register aller Tests, angelegt vor dem Start jedes Tests, gibt einem Team dasselbe. Der Bericht kann dann zeigen, wie viele Tests liefen, wie viele stark, gemischt oder null ausfielen und wie viele von jeder Gruppe in den Bericht kamen, und Rosenthals Frage, wie viele verborgene Studien es bräuchte, um ein Ergebnis zu kippen, erübrigt sich, weil die Zahl der weggelassenen Tests bekannt ist. Der Beitrag zur Spezifikation eines Uplifts zeigt für Nudge-Studien, wie weit veröffentlichte Effekte von einer Vollerhebung der Studien von Nudge-Einheiten entfernt liegen können und warum eine Aussage die Evidenz benennen sollte, auf der sie beruht.

Abbildung 2 Ein Testregister

Registrieren Sie jeden Test vor dem Start und halten Sie dann Ergebnis, Notiz und Weg in den Bericht fest.

Test	Geplante Größe	Ergebnis	Notiz	Berichtet	Falls nicht
Frage, und wann registriert?	Stichprobe oder Laufzeit und gesuchter Effekt vorab fest?	Stark, gemischt, null laut Plan?	Geschrieben, und wo abgelegt?	Welcher Bericht zeigt es?	Warum blieb es draußen?

Zeigen Sie neben jeder Liste erfolgreicher Tests, wie viele Tests liefen und wie viele null ergaben.

Eine Zeile je Test, angelegt vor dem Start. Die geplante Größe zeigt zusammen mit dem gesuchten Effekt, ob ein Test mit Nullergebnis den Effekt hätte erkennen können, auf den es dem Team ankam; die letzten drei Spalten zeigen, wo die Schublade ist.

Quelle Entscheidungsblatt des Autors. Das Register als bekannte Gesamtheit folgt Franco et al. (2014), S. 1 und 2 der Science-Express-Fassung; das Feld für die geplante Größe folgt ihrem Hinweis, dass TESS ausreichende Teststärke verlangt (S. 1), den gesuchten Effekt neben der Größe ergänzt dieser Leitfaden; die Felder für Notiz und Bericht folgen ihrem Befund, dass die meisten Nullergebnisse nie ausgearbeitet wurden (S. 1 und 2, Tabelle 3 auf S. 6), und Munafò et al. (2017), S. 3; das Feld für den Grund folgt den Erklärungen der Autorinnen und Autoren, die Franco et al. berichten (S. 2). Die Felder sind Eingaben der Lesenden; es werden keine Unternehmensdaten geliefert.

Ein Beispiel ohne Unternehmensdaten: Ein Pricing-Team registriert in einem Jahr 24 Preistests, jeweils mit Frage, Stichprobe, Laufzeit und dem kleinsten Effekt, auf den es ankommt, alle vor dem Start festgelegt. Sieben zeigen einen signifikanten Anstieg des Umsatzes je Kunde. Ohne das Register würde der Quartalsbericht die sieben aufführen. Mit ihm zeigt er alle 24 nach Ergebnis, zu jedem Test mit Nullergebnis liegt eine kurze Notiz dort, wo die nächste Analystin oder der nächste Analyst nachsieht, bevor sie oder er denselben Test vorschlägt, und jede Planzahl, die auf den sieben aufbaut, wird neben der Zahl der Tests gelesen, die null ergaben. Auch die geplante Größe zählt: Ein Nullergebnis aus einem Test, der zu klein war, um den Effekt zu erkennen, auf den es dem Team ankam, sagt wenig in die eine oder andere Richtung; Franco et al.¹ konnten das für TESS ausschließen,

weil das Programm eine ausreichende Teststärke verlangte. Das sind Beispiele des Autors und ungeprüft.

Wie weit tragen diese Befunde bei Testergebnissen im Unternehmen?

§ 8 von 8

Als gemessene Zahlen nicht weit. Sterling (1959)⁵ zählte vier psychologische Fachzeitschriften der Jahre 1955 und 1956 aus: Seine Zählung zeigt, was gedruckt wurde, und seine Erklärung dafür beruht auf Annahmen, die er selbst substanziell nannte. Franco et al. (2014)¹ verfolgten ein Programm begutachteter Umfrageexperimente mit ausreichender Teststärke in den Sozialwissenschaften, mit Ergebnissen, die nach den Angaben der Autorinnen und Autoren eingeordnet wurden. Rosenthal (1979)⁴ liefert eine Rechnung unter genannten Annahmen, keine Messung einer Schublade, und Verbeke et al. (2011)⁶ wenden sie in einer Meta-Analyse an, deren Gegenstand Verkäuferinnen und Verkäufer sind, nicht Publikationsbias. Head et al. (2015)² und Munafò et al. (2017) sind Perspektivbeiträge; herangezogen werden sie hier für ihren Gebrauch des Begriffs und, im Fall von Munafò et al.³, für das, was sie von der Präregistrierung erwarten. Keine dieser Arbeiten hat das Testprogramm eines Unternehmens beobachtet.

Nach der Lesart dieses Leitfadens lassen sich zwei Punkte mit einiger Sicherheit übertragen, weil keiner von der Stichprobe abhängt: Eine Menge von Ergebnissen, die danach ausgewählt wurde, dass sie signifikant sind, überzeichnet die Evidenz, und eine Schublade kann sich füllen, bevor jemand die Arbeit begutachtet, nämlich bei der Entscheidung, was ausgearbeitet wird; so war es zum größten Teil in der einen Menge von Studien hier, deren unveröffentlichte Mitglieder sich zählen ließen. Ob die Schublade eines Analyseteams der von TESS gleicht und ob ein Testregister ändert, was ein Team berichtet oder entscheidet, sagen die Quellen nicht.

 Ende des Beitrags

Referenzen

Zitate stehen in der Sprache des Originals; die deutschen Wiedergaben im Fließtext sind sinngemäß und stammen vom Autor.

- 1 Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502–1505. <https://doi.org/10.1126/science.1255484>
- 2 Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3), e1002106. <https://doi.org/10.1371/journal.pbio.1002106>
- 3 Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 0021. <https://doi.org/10.1038/s41562-016-0021>
- 4 Rosenthal, R. (1979). The „file drawer problem“ and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- 5 Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance, or vice versa. *Journal of the American Statistical Association*, 54(285), 30–34. <https://doi.org/10.2307/2282137>
- 6 Verbeke, W., Dietz, B., & Verwaal, E. (2011). Drivers of sales performance: A contemporary meta-analysis. Have salespeople become knowledge brokers? *Journal of the Academy of Marketing Science*, 39(3), 407–428. <https://doi.org/10.1007/s11747-010-0211-8>

Zitierweise

Isoglu, S. (2026, 18. Oktober). Was ist Publikationsbias? Sinan Isoglu. <https://isoglu.com/de/insights/journal/was-ist-publikationsbias/>
Online-Fassung: <https://isoglu.com/de/insights/journal/was-ist-publikationsbias/>

Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand
© 2026 Sinan Isoglu · isoglu.com · Alle Rechte vorbehalten