AUS DER FORSCHUNG

Evidenz statt Anekdote – was eine Zahl überstanden haben muss.

Was eine Promotion zu grenzüberschreitendem Wachstum über Beweise lehrt – und wie man die Zahl in einer kommerziellen Entscheidung am selben Maßstab hält.

Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand

VERÖFFENTLICHT	AKTUALISIERT	LESEZEIT	WÖRTER
1. August 2026	9. September 2026	14 Min.	2.954

https://isoglu.com/de/insights/journal/evidenz-statt-anekdote/

INHALT

MANAGEMENT SUMMARY

Die Zahl in der Anbieterpräsentation und die Zahl im Fachaufsatz sind beide Überlebende eines Auswahlprozesses – und nur eine der beiden Seiten zeigt, was ihre Zahl überstanden hat. Dieser Beitrag versammelt, was gemessen ist: die eine Zählung mit ehrlichem Nenner – sämtliche 126 Studien zweier US-Nudge-Einheiten, im Schnitt 1,4 Prozentpunkte, gegen eine publizierte Stichprobe mit 8,7 – und was das Nachprüfen mit begutachteten Befunden getan hat, quer durch Psychologie, Ökonomie, Medizin und Krebsforschung – in Beträgen ohne gemeinsame Konstante. Er übergibt drei Fragen, die jede Zahl vor einer kommerziellen Entscheidung einordnen: Aus wie vielen? Wann festgelegt? Verglichen womit? Und er benennt die Grenze: Die Fragen sortieren grob, sie liefern keine korrigierte Zahl, und wo keine von ihnen beantwortbar ist, bleibt als ehrliche Grundlage nur erklärte Urteilskraft.

Schlagwörter: Evidenzqualität · Selektionseffekte · Replikation · Kommerzielle Entscheidungen · Fallstudien

Eine Anbieterpräsentation und ein Fachaufsatz landen in derselben Woche auf Ihrem Tisch, vor derselben Entscheidung. Die Präsentation zeigt einen Kunden, der aussieht wie Sie, und einen zweistelligen Zuwachs. Der Aufsatz zeigt einen Durchschnittseffekt und eine Seite Einschränkungen. Wenn Sie ehrlich sind, bewegt Sie die Präsentation mehr. Die Regel, die mir die Promotion eingebaut hat, sagt: Der Instinkt zielt auf die falsche Achse. Eine Zahl wird nach dem Auswahlprozess beurteilt, den sie überstanden hat – nicht nach dem Dokument, in dem sie ankommt. Der eigentliche Unterschied zwischen den beiden Seiten ist, dass die eine zeigt, was ihre Zahl überstanden hat, und die andere nichts davon.

„Evidenz statt Anekdote“ wird meist gelesen als: Aufsätze schlagen Präsentationen. Das ist hier nicht die These. Auch die Präsentation ist Evidenz – für irgendetwas, auf irgendeine Frage. Die These ist: Zahlen erreichen Entscheidungen durch Filter; der Filter entscheidet, was die Zahl beweisen kann; und das Genre der Präsentation zeigt nichts von diesem Filter. Die Arbeitsfrage lautet deshalb nicht, welchem Dokument man Respekt schuldet, sondern: Was musste jede dieser Zahlen überstehen, um auf die Seite zu gelangen?

Was macht reine Überlebensselektion aus gefeierten kommerziellen Benchmarks?

Die sauberste Messung dieses Filters stammt aus einer unglamourösen Ecke: von staatlichen Nudge-Einheiten. Stefano DellaVigna und Elizabeth Linos haben sämtliche randomisierten Studien zweier der größten US-Einheiten zusammengetragen – 126 Studien mit 23 Millionen Menschen, nichts blieb in der Schublade – und sie gegen eine Stichprobe publizierter Nudge-Studien aus zwei akademischen Metaanalysen gehalten. Der Econometrica-Aufsatz berichtet in der publizierten Stichprobe einen durchschnittlichen Effekt von 8,7 Prozentpunkten, ein Plus von 33,4 Prozent gegenüber der Kontrollgruppe. Über alles, was die Einheiten tatsächlich durchführten: 1,4 Punkte, ein Plus von 8,0 Prozent – in den Worten der Autoren „still sizable and highly statistically significant“, also weiterhin beachtlich und statistisch hochsignifikant.

Die Lücke ist kein Betrug, und sie liegt überwiegend nicht einmal an den Maßnahmen. Das Modell der Autoren führt rund 70 Prozent der Differenz auf selektive Publikation, verschärft durch geringe statistische Power, zurück: einzeln ehrliche Studien, gefiltert danach, welche Ergebnisse auffällig genug waren, um aufgeschrieben zu werden. Den größten Teil des Rests führen die Autoren auf die Maßnahmen selbst zurück, die sich in Thema und Kanal zwischen den beiden Stichproben unterscheiden. Die Zahl der Einheiten ist aus genau einem Grund ehrlich: Sie haben den Nenner behalten. Jeder Versuch zählte, nichts konnte still fehlen.

Zwei Vorbehalte reisen mit dem Zahlenpaar, und sie wiegen schwerer als das Paar. Es geht um staatliche US-Nudge-Einheiten und Inanspruchnahme-Quoten, nicht um Kampagnen und Umsatz: Die Größenordnungen übertragen sich nicht. Und 8,7 zu 1,4 ist ein Befund über eine publizierte Stichprobe gegen eine Vollzählung in einem Feld, kein Abschlagsfaktor für Ihren Pilot. Einen solchen Faktor gibt es nicht, wie der nächste Abschnitt zeigt.

Warum schrumpfen selbst begutachtete Management-Zahlen bei empirischer Nachprüfung?

Ein publizierter Befund hat bereits ein feindseliges Gutachterverfahren überstanden – und schrumpft im Durchschnitt trotzdem, wenn jemand ihn nachmisst. Das ist die unbequeme Hälfte der Replikationsbilanz. Die Hälfte, die für kommerzielle Leser zählt, ist, was die Schrumpfung verweigert: eine einzige Zahl zu sein.

Tabelle 1 Was das Nachprüfen mit publizierten Zahlen getan hat, je Feld

Feld	Was getan wurde	Was mit den publizierten Zahlen geschah	Welche Größe das ist
Psychologie: 100 Studien, drei Journale	Nachgemessen mit hoher Power und, wo verfügbar, Originalmaterialien	Replikationseffekte im Schnitt halb so groß wie die Originale; 97 Prozent der Originale signifikant. Von den Replikationen: 36 Prozent signifikant, 47 Prozent der Originaleffekte im 95-Prozent-Konfidenzintervall der Replikation, 39 Prozent als repliziert eingestuft, 68 Prozent signifikant bei kombinierter Evidenz	Effektstärken-Verhältnis plus vier Erfolgsmaße – die Autoren berichten bewusst vier, damit kein Einzelurteil existiert
Laborökonomie: 18 Studien, zwei Spitzenjournale	Nachgemessen mit mindestens 90 Prozent Power, nach vorab festgelegten Plänen	Replizierte Effektstärken im Schnitt 66 Prozent des Originals; 61 Prozent signifikant in der Originalrichtung; vier weitere Replikationsindikatoren: 67 bis 78 Prozent	Effektstärken-Verhältnis und Bestehensquoten – Schwund unter Idealbedingungen, an denen fast nichts verändert wurde
Die meistzitierte klinische Forschung: 49 untersuchte Studien	Gegen spätere, größere oder besser kontrollierte Studien gehalten	Von den 45, die der Aufsatz zählt: 7 widerlegt (16 Prozent), 7 anfangs stärker als das, was folgte (16 Prozent), 20 repliziert (44 Prozent), 11 nie ernsthaft nachgeprüft (24 Prozent)	Urteilszählung über berühmte Befunde – ausgewählt nach Ruhm, keine Basisrate. Die meisten wurden nie umgestoßen
Präklinische Krebsforschung	Unabhängige Wiederholung publizierter Experimente	Medianer Replikationseffekt 85 Prozent kleiner als das Original; 92 Prozent der Replikationseffekte fielen kleiner aus	Das Extrem der Spannweite, im Feld, das kommerziellen Lesern am fernsten liegt
Nudges: publizierte Stichprobe gegen Praxis	Vollzählung der 126 Studien zweier US-Einheiten, gegen eine Stichprobe aus zwei Metaanalysen	8,7 Prozentpunkte in der publizierten Stichprobe; 1,4 über alles Durchgeführte; rund 70 Prozent der Lücke werden selektiver Publikation bei geringer Power zugeschrieben	Eine selektierte Stichprobe gegen einen ehrlichen Nenner – kein Abschlagsfaktor

Open Science Collaboration (2015), Science 349(6251), aac4716, Abstract; Camerer et al. (2016), Science 351(6280), 1433–1436, Abstract; Ioannidis (2005), JAMA 294(2), 218–228; Errington et al. (2021), eLife 10:e71601; DellaVigna & Linos (2020), NBER Working Paper No. 27594, Abstract. Jede Zeile berichtet eine andere Größe; eine gemeinsame Skala existiert nicht – genau das ist der Befund.

Lesen Sie die Richtung der Tabelle, nicht ihre Zellen. Ein Detail verdient einen eigenen Satz. In der klinischen Reihe – Ioannidis' JAMA-Erhebung – schnitten die sechs hochzitierten nichtrandomisierten Studien deutlich schlechter ab als die Studien mit Zufallszuteilung: Fünf wurden später widerlegt oder hatten stärkere Effekte berichtet als

das, was folgte, gegenüber neun von neununddreißig randomisierten. Je schwächer das Design, desto tiefer der Fall. Gemessen an sechs und neununddreißig Fällen, also ein Wegweiser, kein Gesetz.

Beachten Sie, was die stärksten dieser Prüfungen genossen. Die ökonomischen Replikationen liefen mit über 90 Prozent Power nach vorab festgelegten Plänen; das Psychologie-Projekt arbeitete, wo verfügbar, mit den Originalmaterialien; in beiden wurden die Fehlschläge veröffentlicht. Die klinische Zeile unterscheidet sich in der Methode wie im Ruhm: keine Wiederholung, sondern ein Vergleich mit den späteren, größeren Studien – näher an der Art, wie eine kommerzielle Zahl altert. Wo eine publizierte Zahl echter Prüfmaschinerie begegnete und trotzdem schrumpfte, wurde der Schwund unter Bedingungen gemessen, die gebaut waren, ihn zu fangen. Die Zahlen in einer kommerziellen Präsentation haben in aller Regel nichts dergleichen durchlaufen. Das ist meine Folgerung, kein Messergebnis; der nächste Abschnitt sagt, warum ich sie für tragfähig halte.

Welche drei epistemischen Filter trennen Management-Kennzahlen von der Realität?

Die erste Tür ist die Selektion – der oben gemessene Mechanismus. Die Fallstudie einer Präsentation ist aus der Menge der Versuche gewählt, und gewählt danach, wie sie sich liest. In der Forschung ließ sich die Größe dieses Filters messen, weil zwei Einheiten ihren Nenner behalten hatten. Kommerzielles Berichtswesen führt selten einen: In den Pipelines, die ich kenne, hat nie jemand verlangt, ihn zu führen, und nichts im Genre der Präsentation zwingt die gescheiterten Versuche auf die Seite. Die Richtung des Filters ist also bekannt. Seine Größe, für die Zahl vor Ihnen, nicht.

Die zweite Tür ist die Verzerrung, und ihr Vorzeichen ist unbekannt. Selbst die Zahlen des eigenen Dashboards – ganz ohne Anbieter – weichen regelmäßig von dem ab, was ein Experiment findet. Brett Gordon und Kollegen nahmen fünfzehn US-Werbeexperimente bei Facebook, mit den Nutzerdaten der Plattform selbst, und verglichen die experimentelle Antwort mit dem, was branchenübliche Beobachtungsmethoden berichten: „in half of our studies, the estimated percentage increase in purchase outcomes is off by a factor of three across all methods“ – in der Hälfte der Studien lag die Schätzung also über alle Methoden hinweg um den Faktor drei daneben, so die Übersetzung des Verfassers. Und zwar in beide Richtungen, was für Entscheider schlimmer ist als ein konstanter Fehler. Die Daten stammen aus der Zeit vor den Datenschutzumbrüchen von 2019, die Beobachtungsmessung seither schwerer gemacht haben, nicht leichter.

Bei eBay fanden Blake, Nosko und Tadelis experimentelle Renditen der Suchwortwerbung, die nur einen Bruchteil der nichtexperimentellen Schätzungen betrugen; Anzeigen auf den eigenen Markennamen zeigten „no measurable short-term benefits“, keinen messbaren kurzfristigen Nutzen. Ein einzelner Anbieter mit dominanter Marke: Nehmen Sie das Setting mit, nicht ein Gesetz der Suchwortwerbung.

Die dritte Tür ist die im Nachhinein gefundene Metrik. Wer genug Versuche fährt und genug Kennzahlen beobachtet, findet immer etwas, das sich bewegt hat; ein Erfolg, der nach dem Ergebnis ausgerufen wird, ist eine andere Art von Behauptung als ein Erfolg, der vorher definiert war. Die Forschung begegnet dieser Tür im Industriemaßstab: Microsofts Bing, mit „over 200 concurrent experiments“ an einem gewöhnlichen Tag, behandelt „a high occurrence of false positives“ – ein hohes Aufkommen falscher Treffer – als operative Tatsache, die bewirtschaftet werden muss. Und es ist die Tür, die John Ioannidis’ Aufsatz in PLoS Medicine modelliert: Flexibilität „in designs, definitions, outcomes, and analytical modes“, also in Studiendesigns, Definitionen, Endpunkten und Auswertungswegen, zersetzt die Verlässlichkeit des Behaupteten. Seine Schlussfolgerung im Titel ist durchaus umstritten – Goodman und Greenland stehen im Quellenverzeichnis –, doch die Richtung des Flexibilitätsproblems ist nicht der strittige Teil.

Warum überhaupt Maschinerie gegen die eigenen Zahlen bauen? Weil Urteilskraft allein im Maßstab immer wieder versagte. Das Bing-Team berichtet, viele Ideen bewegten Kernmetriken um rund ein Prozent und seien „not well estimated a-priori“, vorab also kaum zu schätzen – und das System habe negative Neuerungen gestoppt, „despite key stakeholders’ early excitement“, trotz früher Begeisterung zentraler Stakeholder. Ron Kohavi und Kollegen haben die Lektion in eine Formel gefasst, die ihren Ruhm verdient: Ertrag entsteht, wenn Teams auf ihre Kunden hören, „not to the Highest Paid Person’s Opinion (HiPPO)“ – nicht auf die Meinung der bestbezahlten Person.

Beachten Sie, was das Beispiel einräumt. Die Korrekturmaschinerie existiert auch im kommerziellen Leben, und Bing betrieb sie industriell. Die Asymmetrie ist nicht Forschung gegen Wirtschaft; sie ist das Artefakt. Ein publizierter Befund ist an die Maschinerie gebunden, die ihn geprüft hat: benannte Methoden, ein Gutachterverfahren, eine Spur, die das nächste Team angreifen kann – so, wie die Replikationsprojekte im Schwesterbeitrag dieses Clusters ihre Vorlagen angegriffen haben. Eine Präsentation kann mit oder ohne all das entstanden sein und sieht in beiden Fällen gleich aus. Genau diese Gleichheit bepreist der nächste Abschnitt.

Welche drei Prüffragen zeigen, ob eine kommerzielle Behauptung übertragbar ist?

Eine Promotion befähigt nicht dazu, einen Markteintritt zu randomisieren, und die Experimentatoren selbst benennen die Grenzen ihres Instruments – „both technical and organizational“, technischer wie organisatorischer Art, so Kohavis Überblicksarbeit. Was die Ausbildung tatsächlich verändert, ist billiger. Sie installiert drei Fragen, eine je Tür, stellbar in jeder Besprechung, ganz ohne Statistik:

1. Aus wie vielen?
2. Wann festgelegt?
3. Verglichen womit?

Aus wie vielen bepreist die Selektionstür. Die Vollzählung zeigt, wie eine echte Antwort aussieht: jeder Versuch gezählt, nichts fehlt. Die kommerzielle Fassung ist kleiner. Auf wie vielen Piloten, Kunden, Tests steht diese Zahl – und wie viele liefen, die ich nicht sehe? Oft lautet die ehrliche Antwort, dass niemand einen Nenner geführt hat. Eine Weigerung an dieser Stelle sagt Ihnen, dass der Nenner die Entscheidung nicht erreichen wird – nicht, dass es ihn nicht gibt, und nicht, dass die Zahl falsch ist. Der Unterschied zählt.

Wann festgelegt bepreist die Flexibilitätstür. Stand die Erfolgsmetrik fest, bevor das Ergebnis existierte – oder wurde sie hinterher in dem gefunden, was sich gerade bewegt hat? Die Norm ist aus der Forschungspraxis importiert, wo es das vorab registrierte Design gibt, weil sein Fehlen teuer wurde; ihr Praktiker-Echo sind Kohavis vorab fixierte Bewertungskriterien. Deutschland vertraut übrigens seit Jahrzehnten einer Institution, die zwei dieser drei Türen institutionell verriegelt: Die Stiftung Warentest kauft ihre Prüfmuster anonym im Handel – der Anbieter stellt die Stichprobe nicht zusammen – und erstellt das Prüfprogramm in der Testplanung, bevor geprüft wird. Eine Analogie für die Praxis, kein Beleg für einen Effekt; aber der Instinkt, dem Testurteil mehr zu trauen als dem Prospekt, ist genau der Instinkt, den dieser Beitrag in den Besprechungsraum tragen will.

Was KI in Revenue Operations wirklich verändert hat die Frage nach der vorab festgelegten Metrik an einer einzelnen Literatur durchexerziert; hier wird sie zur stehenden Ausrüstung.

Verglichen womit bepreist die Verzerrungstür. Was hätte diese Metrik gezeigt, wenn die Initiative nichts bewirkt hätte – eine Kontrollgruppe, ein zeitlicher Versatz, eine Region, oder ein ehrliches „wir können es nicht wissen“?

Das ist die Frage, die die Facebook-Studie mit dem Faktor drei beantwortet, auf Daten in Plattformqualität. Eine Zahl ohne Vergleich ist nicht falsch; sie ist unbepreist.

Drei Fragen sortieren grob. Sie liefern keine korrigierte Zahl; es gibt keinen Schrumpfungsfaktor zum Weiterreichen, und Tabelle 1 ist die Verweigerung. Und sie stufen eine Zahl nur als Beleg über ihr eigenes Umfeld ein: Ob die Welt von Bing, eBay oder einer Nudge-Einheit bis zu Ihrer reicht, ist ein Urteil, das keine Frage automatisiert. Was die Fragen kaufen, ist ökonomisch. Eine Pipeline kann die Antwort verweigern oder ins Ungefähre gehen – und Ungefähres auf eine direkte Frage ist ebenfalls eine Antwort. Stilles Weglassen ist billig und bestreitbar; eine falsche Antwort auf eine direkte Frage ist beides nicht. Das – nicht statistische Raffinesse – ist der Ertrag des Maßstabs.

Wo endet wissenschaftliche Evidenz und wo beginnt unternehmerisches Handeln?

Die größten kommerziellen Entscheidungen sind die, denen das Instrument am wenigsten dient. Ein Markteintritt, eine Übernahme, eine Preisreform: ein Versuch, kein ehrlicher Nenner, oft kein Vergleichsmaßstab. Dort geben die drei Fragen „nicht beantwortbar“ zurück – und dieser Ausgang ist der nützliche. Jacob Cohen hat ein Forscherleben lang zugesehen, wie aus dem Signifikanztest, dem hauseigenen Ritual der publizierten Forschung, Gewissheit gewrungen wurde, und den Satz geschrieben, der sich verallgemeinert: Der Test „does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!“ – er sagt uns nicht, was wir wissen wollen; und wir wollen es so dringend wissen, dass wir aus Verzweiflung glauben, er täte es doch. So die Übersetzung des Verfassers.

Wenn die Fragen leer zurückkommen, ist die ehrliche Grundlage der Entscheidung Urteilskraft: Erfahrung, Struktur, Appetit auf das Verlustszenario – erklärt als Urteilskraft, nicht verkleidet als geliehene Zahl. Und die Präsentation behält ihre eigentliche Aufgabe. Eine nach Relevanz gewählte Fallstudie kann ehrliche Evidenz für Existenz sein – jemand wie Sie hat das zum Laufen gebracht –, sofern ihre eigene Zahl die beiden anderen Türen überstanden hat. Was sie nicht tragen kann, ist ein Schätzwert für die Größenordnung: dafür, was der durchschnittliche Versuch liefern würde. Zu wissen, welche Frage eine Zahl beantworten kann, war immer schon der Kern von „Evidenz statt Anekdote“.

Abbildung 1 Drei Fragen, bevor eine Zahl in die Entscheidung einzieht

DIE ZAHL, WIE SIE ANKAM	AUS WIE VIELEN?	WANN FESTGELEGT?	VERGLICHEN WOMIT?
Wörtlich, mit Quelle und Datum.	Versuche dahinter – sichtbare und unsichtbare.	Metrik vor oder nach dem Ergebnis fixiert.	Was sie zeigen würde, wenn nichts gewirkt hätte.

Eine leere Zelle ist ein Befund. Drei leere Zellen: bestenfalls eine Existenzantwort – entscheiden Sie auf erklärter Urteilskraft.

Eigenes Arbeitsblatt des Verfassers.

Legen Sie das Blatt an die nächste Zahl an, die Budget verlangt. Nicht, um jemanden zu überführen: Die meisten Pipelines wurden nie gebeten, Nenner zu führen, und das festzustellen ist Diagnose, keine Anklage. Der Punkt ist kleiner und härter. Der Maßstab lag nie darin, wo eine Zahl gedruckt wurde – sondern darin, was sie überstanden hat und ob das jemand sagen kann.

Wo liegen die methodischen Grenzen empirischer Evidenzbewertungen?

Grenze. Die Evidenzschwelle ist ein Entscheidungsinstrument, kein Ersatz für Fachurteil. Wende sie auf die aktuelle Kennzahl an und prüfe sie erneut, wenn Quelle oder Basis wechseln.

Evidenzbasis. Der analytische Rahmen stützt sich außerdem auf diese zusätzlichen Quellen: Blake et al. 2015; Cohen 1994; Goodman and Greenland 2007; Gordon et al. 2019; Ioannidis 2005a; Ioannidis 2005b; Kohavi et al. 2013; Kohavi et al. 2009. Die Links identifizieren die konkreten Werke; sie stützen die hier diskutierten Mechanismen und Geltungsgrenzen, nicht jede einzelne Aussage isoliert.

Blake, T., Nosko, C., & Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. Econometrica, 83(1), 155–174. https://doi.org/10.3982/ECTA12423

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

Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. Econometrica, 90(1), 81–116. https://doi.org/10.3982/ECTA18709

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. eLife, 10, e71601. https://doi.org/10.7554/eLife.71601

Goodman, S., & Greenland, S. (2007). Why most published research findings are false: Problems in the analysis. PLoS Medicine, 4(4), e168. https://doi.org/10.1371/journal.pmed.0040168

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. Marketing Science, 38(2), 193–225. https://doi.org/10.1287/mksc.2018.1135

Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. JAMA, 294(2), 218–228. https://doi.org/10.1001/jama.294.2.218

Ioannidis, J. P. A. (2005b). Why most published research findings are false. PLoS Medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1168–1176. https://doi.org/10.1145/2487575.2488217

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. Data Mining and Knowledge Discovery, 18(1), 140–181. https://doi.org/10.1007/s10618-008-0114-1

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

Stiftung Warentest. (o. J.). So testen wir: Wie die Stiftung Warentest arbeitet. https://www.test.de/unternehmen/testablauf-5017344-0/

ENDE DES BEITRAGS

ZITIERWEISE

Isoglu, S. (2026, 1. August). Evidenz statt Anekdote – was eine Zahl überstanden haben muss. Sinan Isoglu. https://isoglu.com/de/insights/journal/evidenz-statt-anekdote/

WEITERLESEN

Aus der Forschung
Was eine zehnwöchige Roblox-Studie sagen kann und was nicht
https://isoglu.com/de/insights/journal/was-eine-zehnwoechige-roblox-studie-sagen-kann-und-was-nicht/

Aus der Forschung
Ein Uplift braucht zuerst eine Spezifikation
https://isoglu.com/de/insights/journal/ein-uplift-braucht-eine-spezifikation/

Aus der Forschung
Survey-Bias ist zuerst ein Entscheidungsfehler
https://isoglu.com/de/insights/journal/survey-bias-ist-ein-entscheidungsfehler/

Sinan Isoglu, MBA (Quantic)
Führungskraft für Umsatzwachstum, Dozent und Doktorand