AUS DER FORSCHUNG

Was das Peer-Review strich, ist im Netz geblieben.

Vier Assistenten lieferten einen Wert, den die veröffentlichte Arbeit nicht enthält; meist verwiesen sie auf Material, in dem er tatsächlich steht.

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

VERÖFFENTLICHT	AKTUALISIERT	LESEZEIT	WÖRTER
27. Juli 2026	28. August 2026	12 Min.	2.526

https://isoglu.com/de/insights/journal/was-peer-review-strich/

INHALT

MANAGEMENT SUMMARY

Nach einer viel zitierten Forschungszahl und ihrer Originalquelle gefragt, lieferten vier KI-Assistenten über dreizehn gewertete API-Durchläufe hinweg einen Wert, den das Peer-Review aus der Arbeit gestrichen hatte. In allen dreizehn wich die Qualitätszahl von der Verlagsfassung ab; in allen bis auf einen enthielt das zitierte Material die gemeldeten 40 Prozent. Der Beitrag trennt die in den Antworten beobachtete Provenienz von der versionsgenauen Evidenz: Das HBS-Working-Paper von 2023 und der veröffentlichte Artikel von 2026 liegen beide vor; die Antwortzählungen bleiben ein eigener Datensatz.

Schlagwörter: Evidenz und Provenienz · Retrieval-augmented Generation · Zitationspraxis · Preprint und Verlagsfassung · Projekt DEAL

Stellen Sie einem Assistenten eine Frage, deren Antwort ein Forschungsbefund ist, und hängen Sie vier Wörter an: Nennen Sie die Originalquelle. Sie bekommen eine Zahl und einen Beleg, und beides sieht richtig aus. Bei einer der beiden hier geprüften Arbeiten war die Zahl in jedem Durchlauf falsch – und der Beleg führte in allen bis auf einen auf ein Dokument, das sie tatsächlich enthält.

Zuerst das Ergebnis. Dreizehn Durchläufe, fünf Modellkonfigurationen, eine bekannte Studie über Beratende im Einsatz von GPT-4: Jede Antwort trug einen Qualitätswert von 40 Prozent, den die veröffentlichte Arbeit nicht enthält. Die Verlagsfassung nennt 33,9 und 29,9 Prozent – im Peer-Review wurde der Wert nach unten korrigiert und die 40-Prozent-Aussage aus dem Abstract entfernt. Keiner der dreizehn Durchläufe zitierte diese Fassung. Alle bis auf eine dieser Antworten passten zu Material in ihrer eigenen Belegmenge, in dem tatsächlich 40 Prozent steht: zu einem Dokument aus der Preprint-Zeit, zur Marketingseite der Beratung oder zu einer Studierendenzeitung. Der Abruf funktionierte. Der Bestand war knapp drei Jahre alt.

Warum war das Replikationsdesign bewusst auf Falsifikation hin angelegt?

Ursprünglich wollte ich zeigen, dass kommerzielle Kennzahlen keine belegbare Herkunft haben – die fünffachen Akquisekosten, die 57 Prozent der Buying Journey, all die Zahlen, hinter denen keine auffindbare Studie steht. Die erste davon ist inzwischen geprüft und ausgewertet: nach der Herkunft gefragt, nannte ein Assistent in jedem Durchlauf Bain und Frederick Reichheld – und die Zuschreibung stammte von der Statistikseite eines Anbieters. Ein Design, das ausschließlich nach nicht rückverfolgbaren Zahlen fragt, erzeugt seinen eigenen Befund. Das Protokoll verlangte deshalb, dass mindestens ein Drittel der Fragen solche sind, für die eine saubere Primärquelle existiert, und hielt vorab fest, welches Ergebnis die These einschränken würde: Zitieren die Assistenten die Primärquelle zuverlässig, wo es eine gibt, dann liegt das Problem in der Literatur – und der Beitrag sagt das.

Also zwei Kontrollfragen, beide zu realen Arbeiten mit begutachteter Verlagsfassung und DOI. Die Antwortaufzeichnungen enthielten außerdem für jede Arbeit eine Preprint- oder Working-Paper-Kennung:

- Brynjolfsson et al. (2025) zu KI und Produktivität im Kundenservice: Quarterly Journal of Economics 140(2), 889–942. Die Antwortaufzeichnungen zitierten außerdem das NBER Working Paper 31161 (Brynjolfsson et al., 2023).
- Dell'Acqua et al. (2026) zu BCG-Beratenden im Einsatz von GPT-4: Organization Science 37(2), 403–423. Die Antwortaufzeichnungen trugen außerdem das Harvard Business School Working Paper 24-013 (Dell'Acqua et al., 2023).

Beide Fassungen von Rang liegen hier vor und sind gelesen. Brynjolfsson et al. (2025) untersuchen „the staggered introduction of a generative AI-based conversational assistant using data from 5,172 customer-support agents" und berichten einen Zuwachs „as measured by issues resolved per hour, by 15% on average". Bei Dell'Acqua et al. (2026) „involved 758 knowledge workers", und „subjects were randomly assigned to one of three conditions: no AI access, GPT-4 AI access, or GPT-4 AI access with a prompt engineering overview"; innerhalb der Grenze berichten sie „completing 12.2% more tasks and completing them 25.1% more quickly on average while also delivering solutions of significantly improved quality". Eine Grenze meiner eigenen Lesart: Die hier vorliegende Fassung von Dell'Acqua ist die Articles-in-Advance-Ausgabe, die Seitenzahlen im Literaturverzeichnis stammen also aus dem Verlagsnachweis und nicht aus der Datei auf der Platte.

Das lokale Belegpaket enthält das Dell'Acqua-Working-Paper-PDF jetzt als separate 58-seitige Quittung. Die Beobachtungen zu seiner früheren Formulierung sind daher gegen diese Arbeitsfassung geprüft; die veröffentlichten Zahlen bleiben ein davon getrenntes Ergebnis der Verlagsfassung.

Der Prompt war jedes Mal derselbe: die Frage, dann „Nennen Sie die Zahl und zitieren Sie die Originalquelle." Die Aufforderung zum Beleg verwandelt ein stilles Auslassen in einen Verstoß gegen eine ausgesprochene Anweisung – und darüber lässt sich belastbarer berichten.

Die Kontrollfragen waren dazu da, die These zu erledigen. Sie haben sie durch eine bessere ersetzt.

Warum ließen redaktionelle Überarbeitungen die zugrunde liegenden Fehler unberührt?

Das Brynjolfsson-Preprint nennt 14 Prozent Produktivitätsgewinn über 5.179 Beschäftigte. Die Verlagsfassung nennt 15 Prozent über 5.172. Eine kleine Korrektur, bei der ein fairer Leser mit den Schultern zuckt.

Bei Dell'Acqua liegt der Fall anders. Die Antwortaufzeichnungen trugen wiederholt einen Satz aus der Working-Paper-Zeit, nach dem Beratende mit GPT-4 Ergebnisse von „more than 40% higher quality" erzielten. Das gehaltene HBS-Working-Paper bestätigt, dass dieser Satz zur früheren Fassung gehört; er ist keine Aussage über die veröffentlichte Arbeit. Die gehaltene Verlagsfassung weist in ihrer Ergebnistabelle 33,9 und 29,9 Prozent aus und lässt die 40-Prozent-Aussage vollständig fallen.

Bei der ersten Arbeit lieferten die meisten Systeme die Zahlen des Preprints; zwei nannten die veröffentlichten und wiesen auf die Korrektur hin. Bei der zweiten lieferte über dreizehn Durchläufe keines die veröffentlichten. Alle dreizehn liefen über API-Endpunkte – und zwar über die AI-Optimization-Endpunkte von DataForSEO, nicht über direkte Aufrufe der einzelnen Anbieter, sodass zwischen Prompt und Modell ein Drittanbieter stand. In den Belegen fanden sich eine Studierendenzeitung, LinkedIn, Facebook, ein YouTube-Video, ein Fachportal für Juristinnen und Juristen, eine überregionale Tageszeitung, eine Karriere-Testseite, ein Trainingsanbieter und die Beratung, deren Beschäftigte die Versuchspersonen waren und die die Daten erhoben hat – und kein einziger Link auf die Zeitschrift.

Ein System nannte die Marketingseite dieser Beratung wörtlich „the original source". Diese Seite gibt das Ergebnis jenseits der Leistungsgrenze mit „23 Prozent schlechter" an. Die begutachtete Arbeit nennt 19 Prozentpunkte. Die Eigendarstellung der beteiligten Beratung rangierte vor dem Experiment selbst.

Wie begünstigen institutionelle Anreize im Peer-Review das Fortleben falscher Zitate?

Der Fall, der entscheidet, wie das alles zu lesen ist, ist der kleinste.

Ein Assistent berichtete, „40 Prozent der Versuchsgruppe erzielten Ergebnisse höherer Qualität". Das ist keine abgeschwächte Fassung des echten Befundes, sondern eine andere Aussage über eine andere Größe – aus einer Effektstärke ist eine Kopfzahl geworden. Belegt war sie mit dem Harvard Crimson. Dort steht wörtlich:

> Additionally, 40 percent of the trial group produced higher quality results.

Das Modell hat seine Quelle getreu wiedergegeben und korrekt zitiert. Der Fehler gehört der Zeitung.

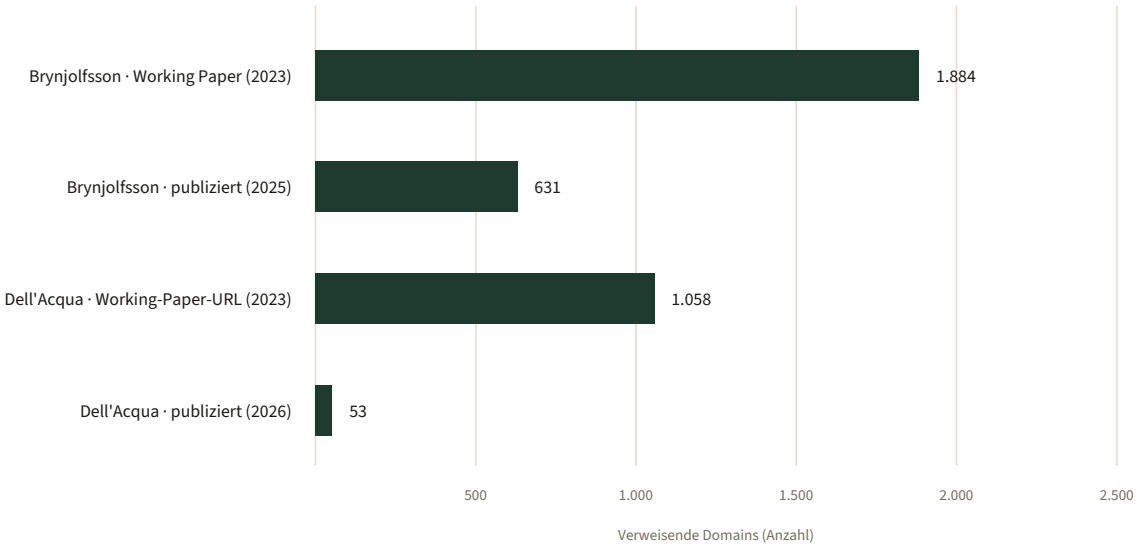
Daneben die andere Richtung der Drift. Ein anderes System nannte den Qualitätsgewinn mit 40,2 Prozent – eine Nachkommastelle, die in keiner der beiden Fassungen vorkommt. Belegt war sie mit einer Karriere-Testseite, die eine Statistikseite zur Working-Paper-Nummer betreibt. In einem Wiederholungslauf erschienen dieselben 40,2 Prozent, diesmal belegt mit dem PDF der Beratung, in dem 40 Prozent steht. Eine Zahl, die als gerundete Untergrenze über zwei Versuchsbedingungen begann, hatte unterwegs eine Nachkommastelle angesetzt, sich von jedem Dokument gelöst, das sie enthält, und wurde mit Beleg zurückgereicht.

Beides ist keine Halluzination in dem Sinn, in dem das Wort gebraucht wird. Die Zahl stand im Bestand. Irgendwo in der Kette zwischen Arbeit und Antwort war seit Jahren etwas falsch, und nichts in dieser Kette hatte einen Anlass, die Arbeit selbst zu öffnen.

Welche quantitativen Daten belegen den Mechanismus unkritischer Zitationsweitergabe?

„Die kostenlose Fassung gewinnt“ ist eine einleuchtende Erklärung und war bis vor Kurzem genau das – eine Erklärung, keine Messung. Also habe ich die Links gezählt.

Abbildung 1 Verweisende Domains: Working-Paper-URL gegen Verlagsfassung



DataForSEO Backlinks, Bulk Page Summary, Live-Index, 27. Juli 2026. Zählung auf URL-Ebene. Working-Paper-URLs: nber.org/papers/w31161; ssrn.com Abstract 4573321. Verlagsfassungen: academic.oup.com/qje/article/140/2/889/7990658; pubsonline.informs.org/doi/10.1287/orsc.2025.21838.

Das Verhältnis sagt das Ergebnis beider Fragen voraus. Bei der Verlagsfassung, die etwa ein Drittel so viele verweisende Domains gesammelt hat wie ihre Working-Paper-URL, fanden zwei von vier Systemen die Verlagsfassung – eine davon in allen drei Durchläufen, mit Nennung beider Werte und einer unaufgeforderten Erklärung der Korrektur. Bei derjenigen mit einem Zwanzigstel fand über alle dreizehn Durchläufe kein System die Verlagsfassung.

Der Mechanismus ist damit enger als „Bezahlschranken verlieren“ – enger, als ich zunächst schrieb. Die Fassung im Quarterly Journal of Economics ist geschlossen. Die in Organization Science ist es nicht; ihre Verlagsseite trägt einen Open-Access-Hinweis und „Copyright © 2026 The Author(s)“. Sie war vom Erscheinungstag an frei, und kein System hat sie dennoch gefunden. Was sie unterscheidet, ist Linkmasse und Alter. Die Dell’Acqua-Working-Paper-URL hatte knapp drei Jahre Zeit, Verweise zu sammeln, verteilt auf drei separat ver-

linkte Fundstellen, von denen jede einzelne den Zeitschriftenaufsatz um eine Größenordnung übertrifft. Die veröffentlichte Fassung hatte vier Monate.

Eine veröffentlichte Arbeit muss den Vorsprung ihrer eigenen Working-Paper-Seite möglicherweise erst aufholen, bevor der Abruf sie findet. Dieser Test zeigt nicht, wie lange das dauert, und verallgemeinert nicht auf jeden kürzlich veröffentlichten Befund.

Wie entwickelte die deutsche Rechnungslegungspraxis eine unbemerkte Lösung für das Problem?

An dieser Stelle lohnt der Blick nach Hause, weil er das Problem schärfer macht statt milder.

Über das Projekt DEAL hat die deutsche Wissenschaft mit drei der größten Verlage – Wiley, Springer Nature und Elsevier – Verträge geschlossen, die von 2024 bis 2028 laufen. Wer als Corresponding Author einer der über 500 teilnehmenden Einrichtungen angehört, publiziert in tausenden begutachteten Zeitschriften unmittelbar Open Access. Entscheidend für alles oben Gesagte: Open Access wird dabei die Verlagsfassung selbst, nicht ein Manuskript daneben.

Das nimmt dem Problem eine Hälfte – und, wie dieser Test zeigt, nicht die entscheidende. Die Fassung in Organization Science war vom ersten Tag an frei, unter CC BY, und gefunden hat sie über dreizehn Durchläufe keines der geprüften Systeme. Freier Zugang beseitigt die Schranke; er beseitigt nicht den Vorsprung von zwanzig zu eins im Linkgraphen, den eine Working-Paper-URL in drei Jahren aufbaut. DEAL sorgt dafür, dass die korrigierten Zahlen lesbar sind. Dass sie gefunden werden, sorgt es nicht.

Verhandelt wurde DEAL als Kosten- und Zugangsfrage. Dass es an einem Provenienzproblem überhaupt mitarbeitet, war kein erklärtes Ziel – und wird, soweit ich sehe, auch nicht so verbucht.

Und selbst diese halbe Hilfe greift hier nicht. Beide geprüften Arbeiten stammen aus den USA, und weder INFORMS noch Oxford University Press gehören zu den DEAL-Verlagen. Wer im deutschsprachigen Raum eine kommerzielle Entscheidung auf internationale Forschung stützt, arbeitet weiterhin in einem Bestand, in dem das Preprint gewinnt. Die eigene Infrastruktur schützt die Seite, auf der publiziert wird. Gelesen wird auf der anderen.

Wie begegnet man dem Einwand gegen eine rigorose Prüfung etablierter Zitate?

Vier Einwände, und einer davon trifft.

„Das Working Paper ist die Originalquelle. Sie haben nach dem Original gefragt.“ Das ist der gute. Bei der Brynjolfsson-Frage ist die Angabe von NBER 31161 eine vertretbare Lesart der Anweisung – das Preprint ist tatsächlich das Original. Unvertretbar ist, die überholten Zahlen des Preprints als Befund der Studie auszugeben, ohne zu erwähnen, dass es eine Korrektur gibt. Zwei Systeme haben das erwähnt. Die übrigen Durchläufe nicht.

„Zwei Arbeiten sind kein Korpus.“ Richtig, und der Beitrag behauptet keine Quote. Zwei Fragen, vier Systeme, je drei Durchläufe – dazu eine Momentaufnahme eines kommerziellen Crawler-Index, der nicht der Index ist, aus dem diese Systeme tatsächlich abrufen. Das ist ein Mechanismus, gestützt auf zwei übereinstimmende Beobachtungen.

„Die Zahlen haben sich kaum bewegt.“ Bei der ersten Arbeit stimmt das – 14 gegen 15. Bei der zweiten nicht, wo ein Hauptergebnis um rund ein Viertel fiel und ein zweiter Befund ersatzlos entfiel.

„Das erledigt sich, wenn die Indizes nachziehen.“ Irgendwann ja. Genau das zeigt Abbildung 1: Die ältere veröffentlichte Arbeit wird bereits gefunden. Das Problem ist die Länge des Fensters – und dass einem Lesenden nichts anzeigt, auf welcher Seite davon eine bestimmte Zahl steht.

Welche vier Prüfstufen muss eine publizierte Managementzahl vor dem Zitieren durchlaufen?

Die praktische Fassung besteht aus drei Fragen, und alle drei verlangen vom Assistenten etwas Überprüfbares – was etwas anderes ist, als ihm weniger zu vertrauen. Sie sind der Abruf-Sonderfall einer allgemeineren Prüfpraxis für jede Zahl, die in eine Entscheidung einzieht.

Abbildung 2 Drei Fragen, bevor eine Forschungszahl in eine Entscheidung eingeht

DIE ZAHL, WIE SIE SIE ERHALTEN HABEN	WAS DAFÜR ZITIERT WURDE	GIBT ES EINE VERLAGSFASSUNG?	ÜBERLEBT DIE ZAHL DORT?
Wörtlich, samt Nachkommastellen.	Ein DOI oder etwas anderes? Benennen.	Titel suchen. Zeitschrift, Jahr, DOI.	Gleich, korrigiert oder entfallen.

Eine Nachkommastelle ist eine Behauptung und braucht eine eigene Herkunft. Stammt die Zahl aus einem Preprint und existiert eine Verlagsfassung, dann ist die Verlagsfassung der Befund – und die Differenz zwischen beiden ist das,

Eigene Arbeitsvorlage.

Fragen Sie nach dem DOI statt nach der Quelle. Ein DOI löst auf genau ein Dokument auf; „die Originalquelle“ löst auf das auf, worauf das Netz am stärksten zeigt. Und behandeln Sie zusätzliche Genauigkeit als Warnung, nicht als Beruhigung – 40,2 Prozent war genauer und unwahrer als 40, und 40 war unwahrer als 33,9.

Warum entscheidet wissenschaftliche Zitationsintegrität über die Glaubwürdigkeit der Forschung?

Der erste Reflex bei einem solchen Befund lautet, die Werkzeuge seien unzuverlässig und man müsse eben selbst nachsehen. Das ist die falsche Lehre, und eine bequeme, weil sie das Problem dorthin verlegt, wo man selbst nicht ist.

Erfunden wurde hier nichts. Jede Zahl wurde korrekt aus einem realen Dokument übernommen, von einem System, das seine Aufgabe erfüllt hat. Das Versagen liegt weiter oben: in einem Bestand, in dem die Pressemitteilung der beteiligten Beratung mehr Verweise hat als die Zeitschrift, in dem das Missverständnis einer Zeitung auffind-

barer ist als die Arbeit, die sie missverstanden hat, und in dem die Korrekturen des Peer-Review zu spät kommen, um noch zu beeinflussen, was irgendjemandem gesagt wird. Eine Abrufschicht auf diesem Bestand liegt zuversichtlich, präzise und gut belegt daneben.

Das Protokoll für diesen Test wurde vor den Durchläufen festgeschrieben, einschließlich der Frage, welches Ergebnis die Behauptung eingeschränkt hätte. Die gesammelten Antwortaufzeichnungen, die ausgezählte Kodierung, die Zählung hinter Abbildung 1 und das Skript erscheinen gemeinsam mit dem Text. Das Paket ist eine Hilfe zum Wiederholen, keine vollständige Studie aller sechs Fragen: Drei Fragen sind ausgezählt, und beide Dell'Acqua-Fassungen liegen jetzt als getrennte Quittungen vor. Bei diesem Thema wäre jede andere Form der Veröffentlichung ihre eigene Widerlegung.

Wo liegen die empirischen Grenzen forensischer Zitationsanalysen?

Grenze. Die Aussage gilt für die geprüften Konfigurationen und Zeitpunkte, nicht für jedes Modell oder Interface. Wiederhole den Test bei Corpus-, Endpoint- oder Modellwechsel.

Evidenzbasis. Der analytische Rahmen stützt sich außerdem auf diese zusätzlichen Quellen: Boston Consulting Group 2023; DEAL-Konsortium. (o. J.). Für Publizierende . Abgerufen am 27. Juli 2026.https://deal-konsortium.de/publizierende; Martinez and Mezitis 2023. Die Links identifizieren die konkreten Werke; sie stützen die hier diskutierten Mechanismen und Geltungsgrenzen, nicht jede einzelne Aussage isoliert.

Boston Consulting Group. (2023). How people create and destroy value with generative AI. https://www.bcg.com/publications/2023/how-people-create-and-destroy-value-with-gen-ai

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work (Working Paper Nr. 31161). National Bureau of Economic Research. https://doi.org/10.3386/w31161

Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at work. The Quarterly Journal of Economics, 140(2), 889–942. https://doi.org/10.1093/qje/qjae044

DEAL-Konsortium. (o. J.). Für Publizierende. Abgerufen am 27. Juli 2026. https://deal-konsortium.de/publizierende

Dell'Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality (Harvard Business School Working Paper Nr. 24-013). https://www.hbs.edu/ris/download.aspx?name=24-013.pdf

Dell'Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. Organization Science, 37(2), 403–423. https://doi.org/10.1287/orsc.2025.21838

Martinez, C. J., & Mezitis, T. A. (13. Oktober 2023). Harvard Business School partners with BCG on AI productivity study. The Harvard Crimson. https://www.thecrimson.com/article/2023/10/13/jagged-edge-ai-bcg/

ENDE DES BEITRAGS

ZITIERWEISE

Isoglu, S. (2026, 27. Juli). Was das Peer-Review strich, ist im Netz geblieben. Sinan Isoglu. https://isoglu.com/de/insights/journal/was-peer-review-strich/

WEITERLESEN

Aus der Forschung
Ein System für immaterielle Werte, zwei nützliche Sichten
https://isoglu.com/de/insights/journal/ein-system-fuer-immaterielle-werte-zwei-nuetzliche-sichten/

Wachstum, das sich potenziert
Die Playbook-Studie fragte nicht, wer die Werkzeuge baute
https://isoglu.com/de/insights/journal/die-playbook-studie-fragte-nicht-wer-die-werkzeuge-baute/

Aus der Forschung
Die Kennzahl starb nicht. Die Kohorte starb.
https://isoglu.com/de/insights/journal/die-kennzahl-starb-nicht-die-kohorte/

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

Online-Fassung: https://isoglu.com/de/insights/journal/was-peer-review-strich/