

MANAGEMENT REVIEW

Management ohne Theater.

STRATEGIE

PILOTPROJEKTE UND EXPERIMENTE VOR DER GROSSEN VERÄNDERUNG

Eine Idee, die sechs Monate wartete, wie wenige getestete Ideen etwas verbessern, das Model for Improvement, warum Effekte im Großen schrumpfen, wann man einer Zahl traut, und eine Karte für den Pilotplan.



12 %

mehr Umsatz bei Bing durch eine kleine Änderung an Anzeigenüberschriften, eine Idee, die über sechs Monate auf einen Test gewartet hatte.

KOHAVI & THOMKE, 2017

IN DIESER AUSGABE

TITELTHEMA

**Sechs Monate in der
Schublade**

SEITE 03

WAS DIE FORSCHUNG SAGT

**Warum Ergebnisse im
Großen schrumpfen**

SEITE 06

WERKZEUG DER AUSGABE

Die Karte für den Pilotplan

SEITE 08

NR. 69 · STRATEGIE

In dieser Ausgabe

Große Veränderungen werden meist aus Überzeugung beschlossen: ein Plan, ein Budget, ein Starttermin. Diese Ausgabe handelt davon, die Veränderung zuerst im Kleinen zu erproben, mit Vergleichsgruppe und vorab vereinbarter Entscheidungsregel, damit das Ergebnis entscheidet, nicht die lauteste Meinung.

Bei Bing wartete eine kleine Idee über sechs Monate und steigerte dann den Umsatz um 12 %. Bei Microsoft verbesserte nur jede dritte getestete Idee, was sie verbessern sollte. Das Model for Improvement testet eine Änderung im Kleinen: ein Tag, eine Einheit. In 126 Behördenstudien waren die Effekte viel kleiner als in veröffentlichten Studien. Vier Prüfungen zeigen, wann man einem Ergebnis traut, und eine Karte hilft beim Planen.

Stiven Janaqi HERAUSGEBER · STIVEN CATALYST

03 TITELTHEMA
Sechs Monate in der Schublade

04 DIE ZAHLEN
Die meisten Ideen wirken nicht

05 DAS MODELL
Drei Fragen, kleine Zyklen

06 WAS DIE FORSCHUNG SAGT
Warum Ergebnisse im Großen schrumpfen

07 WIE MAN ES MISST
Zahlen, denen man traut

08 WERKZEUG DER AUSGABE
Die Karte für den Pilotplan

09 QUELLEN
Quellen und Methode

SO LESEN SIE DIESE AUSGABE

Zahl

Jede Zahl hat Quelle und Jahr unten auf ihrer Seite.

Unsere Lesart

Wo die Redaktion deutet und nicht die Forschung, steht es dabei.

Praxis

Schritte und Karte sind Vorschläge zum Ausprobieren, keine Forschungsergebnisse.

TITELTHEMA

Sechs Monate in der Schublade

2012 schlug ein Microsoft-Mitarbeiter, der an der Suchmaschine Bing arbeitete, vor, Anzeigenüberschriften anders darzustellen. Ein Entwickler hätte dafür wenige Tage gebraucht, doch es war eine von Hunderten Ideen, und die Programmmanager gaben ihr niedrige Priorität.

Die Idee wartete über sechs Monate. Dann startete ein Entwickler, der sah, wie wenig der Code kosten würde, einen einfachen A/B-Test: Die Nutzer wurden zufällig den alten oder den neuen Überschriften zugeteilt. Binnen Stunden stieg der Umsatz so schnell, dass ein Alarm „zu gut, um wahr zu sein“ auslöste. Meist steckt dahinter ein Fehler. Diesmal nicht.

6+

Monate wartete die Idee, mit niedriger Priorität

12 %

mehr Umsatz, ohne die wichtigsten Nutzerkennzahlen zu schädigen

100+ Mio. \$

pro Jahr allein in den USA, nach Schätzung der Autoren

UNSERE LESART

Keine Besprechung hätte diese Idee richtig eingestuft. Ein billiger Test beantwortete in Stunden, was monatelanges Priorisieren nicht geschafft hatte.

Quelle: Ron Kohavi & Stefan Thomke, Harvard Business Review 95(5), 2017 (über HBR reprint R1705E)

Kohavi und Thomke nennen sie die umsatzstärkste Idee in der Geschichte von Bing. Die 12 % wurden im Test gemessen; die 100 Millionen Dollar sind ihre Jahreszahl für die USA.

DIE ZAHLEN

Die meisten Ideen wirken nicht

Unternehmen, die fast jede Änderung testen, lernen dieselbe Lektion, schreiben Ron Kohavi und Stefan Thomke: Die meisten neuen Ideen scheitern im Experiment, und selbst Fachleute irren oft, welche sich auszahlen. Bei Microsoft insgesamt:

IN EXPERIMENTEN GETESTETE IDEEN BEI MICROSOFT, NACH ERGEBNIS



In gut optimierten Produkten wie Google und Bing liefern nur etwa 10–20 % der Experimente positive Ergebnisse, bei Slack etwa 30 % der Experimente zur Monetarisierung (2019). Thomke schätzt den Schnitt auf etwa 10 %.

UNSERE LESART

Wenn zwei von drei plausiblen Ideen nichts ändern oder schaden, ist eine große Veränderung ohne Test eine Wette, keine Entscheidung.

Quellen: Ron Kohavi & Stefan Thomke, Harvard Business Review 95(5), 2017 (über HBR reprint R1705E); Ron Kohavi, Diane Tang & Ya Xu, Cambridge University Press, 2020 (über Chapter 1, published by the authors); Stefan H. Thomke, Harvard Business Review Press, 2020 (über Michael Blanding, "Creating the Experimentation Organization", HBS Working Knowledge, 2019)

Online-Produkte, in denen Tests billig sind; für Betriebe fanden wir keine vergleichbaren Anteile. Slack: ein Beitrag, zitiert von Kohavi, Tang & Xu; Thomkes 10 %: ein Interview (2019).

DAS MODELL

Drei Fragen, kleine Zyklen

Das Model for Improvement von Associates in Process Improvement, dargestellt in The Improvement Guide (Langley und Kollegen, 2009), stellt drei Fragen, bevor etwas dauerhaft geändert wird.

01**Ziel**

Was wollen wir erreichen?

02**Messung**

Woran erkennen wir, dass eine Änderung eine Verbesserung ist?

03**Änderung**

Welche Änderung führt zu einer Verbesserung?

DIE ÄNDERUNG TESTEN: DER ZYKLUS PLAN-DO-STUDY-ACT

01**Planen****02****Tun****03****Studieren****04****Handeln**

Änderung planen, erproben, Ergebnisse studieren, nach dem Gelernten handeln. Der Test ist bewusst klein: Das Beispiel des Institute for Healthcare Improvement ist ein Patient, ein Tag. Nach mehreren Zyklen unter wechselnden Bedingungen, etwa Tag und Nacht, wird die Änderung dauerhaft eingeführt und ausgeweitet.

UNSERE LESART

Ein Pilot ist kein sanfter Start, sondern eine Frage mit Frist: Schneidet diese Änderung hier besser ab als das, was wir heute tun?

Quelle: Langley, Moen, Nolan, Nolan, Norman & Provost, Jossey-Bass, 2009 (über Institute for Healthcare Improvement, "Model for Improvement")

Fragen und Schritte folgen der Seite des Institute for Healthcare Improvement zum Modell; das Buch selbst lag nicht vor. Die Deutung stammt von der Redaktion.

[Mehr im Essay](#)

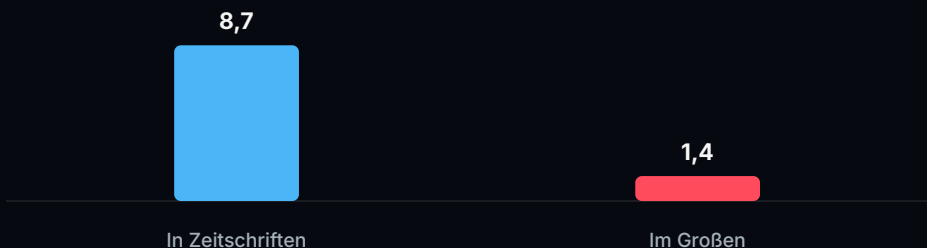
DMAIC ohne Black Belt

WAS DIE FORSCHUNG SAGT

Warum Ergebnisse im Großen schrumpfen

Stefano DellaVigna und Elizabeth Linos sammelten 126 randomisierte Studien mit 23 Millionen Menschen, alle von zwei großen „Nudge Units“ der US-Regierung, und verglichen sie mit Nudge-Studien aus Fachzeitschriften.

MITTLERER EFFEKT EINES NUDGES AUF DIE TEILNAHME, PROZENTPUNKTE



Die veröffentlichten Studien waren klein: Der Median einer Behandlungsgruppe lag bei 484 Personen, in den Nudge Units bei 10.006. Kleine Studien erkennen nur große Effekte, und Zeitschriften drucken vor allem, was erkannt wird. Selektive Veröffentlichung, verstärkt durch geringe statistische Power, erklärt etwa 70 % der Lücke; Unterschiede der Nudges erklären den Großteil des Rests.

UNSERE LESART

Die Erfolgsgeschichte, die man liest, ist die veröffentlichte. Besser einplanen, dass die Einführung weniger bringt als der Pilot, der überzeugt hat.

Quelle: Stefano DellaVigna & Elizabeth Linos, *Econometrica* 90(1), 2022

Teilnahme: der Anteil derer, die tun, worum gebeten wird, etwa sich anmelden oder impfen lassen. Effekte und die 70 % aus *Econometrica* (2022); Gruppengrößen aus dem Arbeitspapier von 2020.

WIE MAN ES MISST

Zahlen, denen man traut

„Zahlen zu bekommen ist leicht; Zahlen, denen man trauen kann, sind schwer“, schreiben Kohavi und Thomke. Vier Prüfungen, bevor ein Pilotergebnis entscheidet:

1

Mit einer Kontrollgruppe vergleichen

Bei Yahoo bezifferte eine Beobachtungsstudie den Effekt von Werbebannern auf Markensuchen auf 871–1.198 %; ein kontrolliertes Experiment fand 5,4 %.

2

Eine Hauptkennzahl und Leitplanken festlegen

Vor dem Start, nicht danach; Leitplanken dürfen nicht schlechter werden.

3

Mindestens eine volle Woche laufen lassen

LinkedIn misst mindestens eine Woche: Ein einzelner Tag überbewertet Vielnutzer.

4

Aufteilung prüfen, Überraschungen misstrauen

Falsche Gruppengrößen machen das Ergebnis oft wertlos. Twymans Gesetz: Jede Zahl, die interessant oder anders aussieht, ist meist falsch.

HYPOTHETISCHES BEISPIEL, EINE NEUE KOMMISSIONIERROUTE IM LAGER

Vergleich Zone A neue Route, Zone B wie bisher, dieselben zwei Wochen

Hauptkennzahl Positionen pro Stunde: A +6 %, B +4 %

Leitplanke Pickfehler: in beiden unverändert

Gegenüber Zone B bringt die Route etwa 2 Punkte, nicht 6. Die Zahlen sind erfunden.

Quellen: Ron Kohavi & Stefan Thomke, Harvard Business Review 95(5), 2017 (über HBR reprint R1705E); Ron Kohavi, Diane Tang & Ya Xu, Cambridge University Press, 2020 (über Chapter 1, published by the authors); Ya Xu, Weitao Duan & Shaochen Huang, KDD '18 (ACM), 2018

Prüfungen nach Kohavi & Thomke (2017), Kohavi, Tang & Xu (2020) und, für die Woche, Xu et al. (2018). Das Beispiel stammt von der Redaktion.

WERKZEUG DER AUSGABE

Die Karte für den Pilotplan

Vor dem Start ausfüllen, mit denen, die den Pilot durchführen. Dann in Stufen einführen: erst eine kleine Gruppe, dann eine Messphase, dann alle.

01 Hypothese wenn wir X ändern, verbessert sich Y um etwa Z, weil ...

02 Was sich ändert eine Änderung, so klar, dass jeder sie wiederholen kann

03 Wo und wie lange ein Bereich, eine Schicht, ein Team; mindestens eine normale Woche

04 Was gemessen wird eine Hauptkennzahl, wie sie gezählt wird, und die Leitplanken

05 Vergleichsgruppe wo es an denselben Tagen weiterläuft wie bisher

06 Entscheidungsregel vor dem Start: einführen, anpassen oder stoppen, und wer entscheidet

Quellen: Langley, Moen, Nolan, Nolan, Norman & Provost, Jossey-Bass, 2009 (über Institute for Healthcare Improvement, "Model for Improvement"); Ya Xu, Weitao Duan & Shaochen Huang, KDD '18 (ACM), 2018; Stefan H. Thomke, Harvard Business Review Press, 2020 (über Michael Blanding, "Creating the Experimentation Organization", HBS Working Knowledge, 2019)

Eine Praxis, die die Redaktion vorschlägt, nach dem Model for Improvement, Xu et al. (2018) und Thomkes Regel: eine prüfbare Hypothese, eine Kontrollgruppe und sich an die Ergebnisse halten.

Tool öffnen: Sigma & Regelkarte

stivencatalyst.com/de/tools/sigma-control-chart/

QUELLEN UND METHODE

Jede Zahl hat eine Quelle.

Die Zahlen dieser Ausgabe stammen aus den folgenden Quellen. Das Jahr zeigt, wie aktuell jede ist.

The Surprising Power of Online Experiments

Ron Kohavi & Stefan Thomke, Harvard Business Review 95(5), 2017 · über HBR reprint R1705E

<https://hbr.org/2017/09/the-surprising-power-of-online-experiments>

Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing

Ron Kohavi, Diane Tang & Ya Xu, Cambridge University Press, 2020 · über Chapter 1, published by the authors

<https://experimentguide.com/>

Experimentation Works: The Surprising Power of Business Experiments

Stefan H. Thomke, Harvard Business Review Press, 2020 · über Michael Blanding, "Creating the Experimentation Organization", HBS Working Knowledge, 2019

<https://www.library.hbs.edu/working-knowledge/creating-the-experimentation-organization>

The Improvement Guide: A Practical Approach to Enhancing Organizational Performance (2nd ed.)

Langley, Moen, Nolan, Nolan, Norman & Provost, Jossey-Bass, 2009 · über Institute for Healthcare Improvement, "Model for Improvement"

<https://www.ihl.org/resources/how-improve-model-improvement>

RCTs to Scale: Comprehensive Evidence from Two Nudge Units

Stefano DellaVigna & Elizabeth Linos, Econometrica 90(1), 2022

<https://doi.org/10.3982/ECTA18709>

SQR: Balancing Speed, Quality and Risk in Online Experiments

Ya Xu, Weitao Duan & Shaochen Huang, KDD '18 (ACM), 2018

<https://doi.org/10.1145/3219819.3219875>

REDAKTIONELLE METHODE

Jede Zahl wurde auf Jahr, Herausgeber und genaue Bedeutung geprüft. Wo die Seite des Herausgebers nicht erreichbar war, wurde sie an unabhängigen Zusammenfassungen geprüft und mit „über“ markiert. Die Deutung der Redaktion ist als „Unsere Lesart“ markiert. Zahlen, die sich nicht bestätigen ließen, stehen nicht in der Ausgabe.



MANAGEMENT REVIEW

Management ohne Theater.

Jede Ausgabe eine Führungsfrage, geprüft an der besten Forschung.

ALLE AUSGABEN

stivencatalyst.com/de/magazine/management-review.html