Quellendokumentation Gruppe B3

Thema: A/B-Testing

Gruppenmitglieder:

- Julia Eberhardt (388742)
- Valentin Koob (392628)
- Judith Mengel (392609)
- Margarita Mishinova (396046)
- Meret Vollmann (347528)

Erstellt im Rahmen der Veranstaltung: Methoden des Usability Testing & Engineering Wintersemester 2017/18
Technische Universität Berlin

Autor(en): Beasley, M.

Titel: Practical Web Analytics for User Experience: How Analytics Can Help You Understand Your Users

Jahr: 2013

Laufende Nummer im Literaturverzeichnis: 01

Schlagwörter (3 bis 5): Einführung, Experimentplanung, Instrumente

Kurzbeschreibung:

Der Kapitel über A/B Testing bietet einen Einstieg ins Thema, indem es veranschaulicht, wie Experimente zu planen sind. Instrumente wie Google Analytics werden vorgestellt und es werden Tipps gegeben, wann ein A/B Testing Sinn macht (ab ungefähr 5000 Besucher im Monat) und ab wann das Experiment Ergebnisse liefert.

Erstellt von: Margarita Mishinova, 396046

Autor: Chopra, P.

Titel: The Ultimate Guide to A/B Testing

Jahr: 2012

Laufende Nr. im Literaturverzeichnis: 02

Schlagwörter: A/B Testing, Evaluation, Usability Methode

Kurzbeschreibung: Das Buch gibt einen grundlegenden Einstieg in die Thematik A/B Testing. Ziel eines A/B Tests ist die Entscheidung zwischen zwei Alternativen, die sich lediglich in einem Aspekt voneinander unterscheiden. Wenngleich die Logik sich nicht primär auf ein Anwendungsgebiet beschränkt, so findet sie die größte Anwendung im Web-Bereich. Folgende Schritte sind zu beachten: 1) Bestimmen der Ziele nach der angestrebten Optimierung. Beispielsweise die generelle Anzahl von gekauften Artikel soll steigen. 2) Ableiten relevanter, objektiv messbarer Variablen. Beispielsweise die generelle Anzahl von gekauften Artikeln. 3) Entwickeln der Testvarianten. Beispielsweise eine veränderte Aufforderung gegenüber dem herkömmlichen Text. 4) Zuteilung der Versuchspersonen. 5) Erhebung der Daten. 6) Statistische Absicherung der Unterschiede und 7) darauf aufbauend eine Entscheidungsfindung. Beispielsweise die Entscheidung, dass ein veränderter Text zu einer höheren Rate gekaufter Artikel führt. Weiter geht das Kapitel auf einige Tipps bezüglich der Wahrung von Objektivität und Validität ein. So sollten z.B. die Alternativen gleichzeitig getestet werden oder die Entscheidung lediglich auf Basis der Daten und nicht auf Basis einer persönlichen Vorliebe gefällt werden. Ferner wird empfohlen, sich durch kontinuierliche und sukzessive Durchführung von A/B Tests einer stetigen Optimierung anzunähern.

Stärken und Schwächen des Konzepts: Da A/B Testing vor allem eine experimentelle Investigationslogik ist, bietet sich eine sehr breite Einsetzbarkeit. Ob Websites, Einkaufsläden oder Prototypendesigns getestet werden spielt prinzipiell keine Rolle. Das gleiche gilt für die erhobenen Metriken (bspw. die Fehlerrate in einem Usability-Test). Die grundlegende Idee bleibt immer die Gleiche. Vor allem hervorzuheben ist, dass bei kontrollierten externen Einflussfaktoren kausale Schlüsse gezogen werden können. Ferner kann ein Test je nach Ausführung recht simpel gehalten werden und dennoch eine objektive Grundlage für die Entscheidung zwischen zwei Alternativen liefern.

Den Vorteilen stehen auch einige Schwächen gegenüber. Dabei gilt, dass es schwierig sein kann, die Validität und Belastbarkeit der Studie zu gewährleisten, da es potenziell viele konfundierende Variablen gibt (z.B. Kohavi, Deng, Frasca, Longbotham, Walker, & Xu, 2012). Auch sollte die durchführende Person grundlegende Kenntnisse der Statistik mitbringen und vor allem wissen, wie diese Ergebnisse zu interpretieren sind. Hinzu kommt, dass A/B Tests nur Reaktionen auf Veränderungen liefern. Dabei bleibt immer unklar, warum die Veränderung zu einem gewissen Effekt geführt hat (z.B. „warum wird die Fehlerrate geringer?). Somit ist es mit A/B Tests nicht möglich, tiefere bzw. psychologische Gründe für Veränderungen zu erforschen.

Erstellt von: Valentin Koob 392628

Autor(en): Christian, B.

Titel: The A/B Test: Inside the Technology That's Changing the Rules of Business

Jahr: 2012

Laufende Nummer im Literaturverzeichnis: 03

Schlagwörter: A/B-Testing, Anwendung in der Praxis, standardisierte Messinstrumente

Kurzbeschreibung: Der Artikel beschreibt, wie A/B-Testing Einzug in alltägliche Verfahren nicht nur in großen Unternehmen im Silicon Valley, sondern auch in kleinen und mittelständischen Unternehmen erhielt. Die Entwicklung von Programmen, die es Mitarbeitern ohne Programmier- und Statistikkenntnisse erlaubt, Websites zu designen und diese gegeneinander zu testen, hatte daran erheblichen Anteil. Es werden einige Prinzipien und Vorteile herausgestellt: Unternehmen müssen sich bezüglich der Website-Gestaltung nicht von vornherein für ein Produkt entscheiden, sondern können verschiedene ausprobieren. Die gewonnenen Daten der Tests entscheiden dann über das weitere Vorgehen. Des Weiteren macht es A/B-Testing möglich, Echtzeitdaten zu erhalten. Außerdem sind die Gründe, warum das eine oder andere Design bessere Daten erzeugt, zu vernachlässigen, solange das Outcome den Verkauf erhöht. Die dynamische und teilweise automatisierte Anpassung der Gestaltung, sobald signifikante Unterschiede zwischen Produkten entdeckt werden, ermöglicht eine weitgehend von der ständigen Aufsicht von Mitarbeitern unabhängige Veränderung des Produktes.

Stärken und Schwächen: Standardisierte Verfahren zur Gestaltung von Websites und A/B-Testing erhöhen die Zugänglichkeit und Anwendbarkeit für kleine und mittelständische Unternehmen. Allerdings werden auf diese Weise durch mangelnde methodologische Kenntnisse häufig mehrere voneinander völlig unterschiedliche Websites gegeneinander getestet. Dies lässt keine nachträgliche Rückführung der Kausalität auf einzelne Unterschiede zu. Ein Nachteil von stetiger Veränderung kleiner Details an Produkten ist, dass womöglich zu kleinteilig gedacht wird und große Veränderungen ausgebremst oder verhindert werden. Auch ersetzt A/B-Testing generell keine Innovationen.

Erstellt von: Judith Mengel 392609

Autoren: Dixon, E., Enos, E., & Brodmerkle, S.

Titel: United States Patent: A/B Testing Of A Webpage

Jahr: 2011

Laufende Nummer im Literaturverzeichnis: 04

Schlagwörter (3-5): Technische Funktionsweise, Einteilung der Testgruppen, IP-Adresse, Cookies

Kurzbeschreibung: Der Patentschein beschreibt die Funktionsweise der Einteilung von Personen in die Testgruppen für die verschiedenen Versionen der zu testenden Website. Beim Zugriff auf die Website wird jedem Nutzer ein Code zur Identifizierung zugewiesen, der über die Zuteilung zu einer Websiteversion bestimmt. Dieser Code kann durch die IP-Adresse des Geräts, das auf die Website zugreift, zugewiesen werden, oder durch den Server, auf dem die Website liegt. Im ersten Fall wird die IP-Adresse des Nutzers geprüft und damit entschieden, welche Version der Website dem Nutzer angezeigt wird. Dies kann über die letzte Ziffer erfolgen, wobei entscheidend ist, ob die Ziffer gerade oder ungerade ist. So wird beispielsweise allen IP-Adressen mit der Endung 0 die Version A zugewiesen und IP-Adressen mit der Endung 1 die Version B. Serverseitig funktioniert die Einteilung, indem ein sog. „tracking ID cookie“ an den Computer des Nutzensendend gesendet wird, der anstelle der IP-Adresse zur Identifizierung dient. Der Nutzer fragt über ein Endgerät (Laptop, Computer, Smartphone etc.) eine Website an. Es wird eine Zielwebsite aufgebaut, die ein Java Skript enthält, das die IP-Adresse des Endgeräts prüft, die IP-Adresse bzw. den tracking cookie in das Skript einsetzt und darüber den Nutzer auf die ihm zugeteilte Version der Zielwebsite umleitet. Die Versionen A und B unterscheiden sich in ihrer URL, sodass die Weiterleitung durch das Skript entscheidend ist. Im weiteren Verlauf wird das Verhalten der Nutzer auf den verschiedenen Versionen der Website verfolgt und gespeichert, um zu ermitteln, ob die Versionen A und B sich verschieden auf das Nutzungsverhalten auswirken. Durch Analyse größerer Teile der IP-Adresse können Nutzer in mehrere Gruppen eingeteilt werden (z.B. Versionen A bis E). Die Versionen sollten nur kleine Unterschiede haben, sodass ein bestimmtes Verhalten eindeutig einer Ursache zugewiesen werden kann. Der A/B-Test trägt so durch mehrfaches Wiederholen zur sukzessiven Verbesserung der Website bei.

Stärken und Schwächen: In dem Patentschein werden Grafiken gezeigt, die übersichtlich den technischen Aspekt des A/B-Testings darstellen. So wird die Funktionsweise von A/B-Tests auf einen Blick klar. Zusätzlich wird die Vorgehensweise sehr ausführlich und für jeden möglichen Fall zur Identifikation der Nutzer erläutert. Als Schwäche des Patentscheins als Quelle lässt sich sagen, dass nur der technische Aspekt erläutert wird, nicht aber die Anwendungsbereiche und Einsatzmöglichkeiten von A/B-Testing. Da es sich aber um einen Patentschein handelt, war letzteres auch nicht der Anspruch der Quelle.

Erstellt von: Meret Vollmann 347528

Autor(en): Dmitriev, P., Frasca, B., Gupta, S., Kohavi, R., & Vaz, G.

Titel: Pitfalls of long-term online controlled experiments

Jahr: 2016

Laufende Nr. im Literaturverzeichnis: 05

Schlagwörter: konfundierende Variablen, Messmethodik, Validität, A/B Testing

Kurzbeschreibung: Das Paper erörtert mögliche Gefährdungen der internen und externen

Validität bei der Durchführung von langfristigen, kontrollierten A/B Tests im Online-Bereich. Dabei stehen insbesondere sechs Einflussfaktoren im Vordergrund, welche die Interpretierbarkeit der Ergebnisse beeinflussen können: A) Cookies identifizieren einen User. Da ca. 25-30% der Cookies innerhalb eines Monats gelöscht werden, kann die Zugehörigkeit einer Person zu ihrer Bedingung nicht über das gesamte Experiment gewährleistet werden. Ferner gilt, dass mit Hilfe eines Cookies lediglich ein Gerät und nicht die Versuchsperson an sich identifiziert werden kann. B) Weiter kann die Eigenschaft einer Bedingung dazu führen, dass nur gewisse User das Experiment komplett beenden, welche dann nicht mehr populationsrepräsentativ sind. C) Nutzer könnten auch erst gegen Ende des Versuchs einer der Gruppen beitreten und so den Effekt verwässern. D) Ferner könnten sich Eigenschaften der Population über die Zeit verändern. Somit sind Effekte nicht primär auf die Manipulation einer Variable zurückzuführen. E) Weitergehend kann eine Bedingung die Eigenschaften (z.B. Erfahrung, Zugang zu anderen Funktionen) von Nutzern ändern, sodass diese eine systematische Verzerrung in weiteren Studien mit sich bringen. F) Zuletzt können äußere Ereignisse die Ergebnisse unabhängig von der Manipulation verändern. Summa summarum kann jeder dieser Punkte einen verzerrenden Einfluss auf die Interpretierbarkeit der Ergebnisse haben. Zumal es nie die Möglichkeit gibt nach einer Studie ein (unerwartetes) Ergebnis auf eins der sechs Probleme eindeutig zu attribuieren. Vielmehr gilt das Verständnis, diese konfundierenden Variablen nach Möglichkeit auszuschließen.

Stärken und Schwächen des Papers: Das Paper gibt einen beispielhaften, knappen und methodisch-theoretischen Überblick über die Thematik der externen und internen Validität von kontrollierten Online-Studien (= u.a. A/B Testing). Im Zentrum steht die oft übersehene Mahnung, nicht jedes Ergebnis direkt als solches hinzunehmen, sondern dieses vor dem Hintergrund möglicher Einflussfaktoren kritisch zu evaluieren. Wenngleich die jeweiligen Faktoren auf inhaltlich logischen Überlegungen fußen und anhand von Praxisbeispielen erläutert werden, so bleiben diese jedoch auf einer theoretischen Ebene und mangeln somit an konkreter empirischer Überprüfung.

Erstellt von: Valentin Koob 392628

Autor(en): Courtney Eckerle

Titel: Case Study: Email Marketing: Extra Space Storage uses a customer-first approach for a 50% boost in email conversion rate

Jahr: 2016

Laufende Nr. im Literaturverzeichnis: 06

Schlagwörter: Fallstudie, E-Mails, Praxisbezug, iterativer Prozess

Kurzbeschreibung: Die Firma Extra Space Storage begann ihr E-Mail-System für Reservierungen zu überarbeiten, da dieses nicht nutzeroptimiert war.

Die Firma bietet ihren Kunden Möglichkeiten ihren Besitz zu lagern. Sie besitzen 1400 Lagereinrichtungen. Die beiden Hauptprobleme, waren dass 45% der Kunden ihre Reservierungen nicht wahrnahmen und dass viele Kunden mit dem Lagerungssystem überfordert waren.

Als Maßnahme dagegen wurden den Kunden per Mail Informationsvideos zur Verfügung gestellt und personalisierte auf den Ort bezogene Informationen gesendet. Den Verantwortlichen wurde klar, dass Nutzer gerne auf dem Kommunikationsweg, auf dem sie die Anfrage senden, auch eine Antwort erhalten möchten, sodass sie auf E-Mails nicht

weiterhin mit Anrufen reagierten. Außerdem erkannten sie, dass die gesendeten E-Mails auf dem Handy aufgrund des Formats kaum lesbar waren und auch die Webseite nicht für die mobile Anwendung geeignet war.

Für die Analyse verwendeten sie Nutzungsdaten der potentiellen Kunden durch A/B Tests und den Myers-Briggs Personality-Type Indicator. Die E-Mails wurden weiterhin auf den aktuellen Stand im Buchungsprozess optimiert und die gegebene Information für die aktuelle Situation angepasst. Das E-Mail Verfahren wurde in einem iterativen Prozess über drei Stufen hinweg angepasst.

Dadurch steigerte sich die Anzahl an Konversionen um 50 Prozent und auch die Nutzung der Webseite auf Mobilgeräten stieg deutlich an.

Stärken und Schwächen: Die Darstellung eines Fallbeispiels ermöglicht es A/B Testing nicht nur an theoretischen Beispielen, sondern auch an einem konkreten Fall nachzuvollziehen. Es wird jedoch in der Fallstudie deutlich mehr auf die speziellen Anforderungen der Firma eingegangen als auf die tatsächlichen Evaluationsverfahren. Tatsächlich wird nur angegeben, dass Verhaltensaufzeichnungen, der Myers-Briggs Personality-Type Indicator (MBTI) und „data“ verwendet werden. Die gibt nahezu keinen Einblick in das methodische Vorgehen. Die Verwendung des MBTI ist zudem fragwürdig, da dieser nicht zu den etablierten Persönlichkeitstests mit nachgewiesener Validität gehört. Mögliche Probleme in Evaluationsprozess sind auch nicht aufgelistet und der Bericht klingt sehr positiv, sodass es an Werbung für das eigene Vorgehen grenzt. Die Beschreibung des iterativen Entwicklungsprozesses gibt jedoch einen interessanten Einblick in die Praxis.

Erstellt von: Julia Eberhardt 388742

Autor: Farmer, J.

Titel: Statistical Analysis and A/B Testing

Jahr: 2009

Laufende Nr. im Literaturverzeichnis: 07

Schlagwörter: z-Test, Entscheidungsfindung bei A/B Tests, statistische Auswertung

Kurzbeschreibung: Die Website gibt eine kurze Einführung in die statistische Testung von relativen Häufigkeiten (z.B. Kaufrate oder Fehlerrate relativ zur Gesamtanzahl der Besucher) bei A/B Tests anhand eines Zahlenbeispiels. Als Ausgangspunkt steht die Frage, ob Unterschiede im Kriterium zwischen (zwei) Gruppen rein zufällig zustande hätten kommen können. Die Antwort wiederum bildet dann eine statistische Grundlage zur Entscheidung, welche der (beiden) Varianten besser ist. Hierfür wird zuerst die Differenz im Kriterium der zu testenden Gruppen gebildet, welche eine normalverteilte Zufallsvariable darstellt. Der dann mittels z-Transformation erhaltene z-Wert kann mit den für $\alpha = 5\%$ kritischen Werten von ± 1.96 bei zweiseitiger Testung oder -1.65 bzw. $+ 1.65$ bei einseitiger Testung verglichen werden. Je nachdem, ob der z-Wert die eben angesprochenen Grenzen überschreitet, spricht man von einer statistisch signifikanten bzw. nicht signifikanten Differenz. Dabei gilt, dass je höher die Differenz und je größer die Stichprobe, desto größer ist der z-Wert und desto wahrscheinlicher wird die Differenz signifikant. Die Formel hierzu lautet:

Dabei stellt p die relative Häufigkeit und N die Stichprobengröße der jeweiligen Gruppe (1 und 2) dar. Wichtig ist hierbei anzumerken, dass die Formel nur unter der Nullhypothese H_0 : gilt. Entschieden wird sich dann für die statistisch abgesicherte, bessere Variante. Analog kann man sich bei mehreren Gruppen für die beste Variante relativ zu einer

Vergleichsgruppe entscheiden.

Stärken und Schwächen: Der vorgestellte z-Test bietet ein einfaches Instrument zur Entscheidungsfindung/ Hypothesentestung bei relativen Häufigkeit bzw. Wahrscheinlichkeiten. Insbesondere angesichts dessen, dass bei A/B Tests im Online-Bereich oft relative Häufigkeiten analysiert werden (Kohavi & Longbotham, 2015). Ferner lässt sich das Prinzip mit einer Modifizierung auch auf Mittelwerte erweitern. Bedingt durch die Einfachheit ist das das Verfahren auch dann anwendbar, wenn Versuchsleiter weniger statistisches Hintergrundwissen besitzen bzw., wenn lediglich ein simpler A/B Test durchgeführt werden soll.

Entgegen dessen gibt es auch einige Schwächen: 1) Der Test gilt nur unter der Nullhypothese: H_0 : und bei relativen Häufigkeiten. Eine Testung außerhalb dieser Bedingungen ist also mit *diesem* z-Test nicht möglich. 2) Es gibt keine Entscheidungsgrundlage, wenn die Differenzen nicht signifikant werden (= „welche ist jetzt besser?“). 3) Der direkte Vergleich von z-Werten aus verschiedenen Tests ist nur bei gleicher Stichprobengröße der Gruppen sinnvoll. 4) Mehrfachtestung von Varianten gegenüber einer Kontrollgruppe kann zur Alphafehler-Kumulierung führen. 5) Die Normalverteilungsannahme der Zufallsvariable gilt nur dann, wenn n und p , wobei N die Gruppengröße beschreibt.

Erstellt von: Valentin Koob 392628

Autor: Kohavi, R., Crook, T., & Longbotham

Titel: Online Experimentation at Microsoft

Jahr: 2009

Laufende Nr. im Literaturverzeichnis: 08

Schlagwörter: A/B testing, firmenkulturelle Probleme, Umsetzungsprobleme

Kurzbeschreibung: Das Paper geht auf die Methode des A/B Testing sowie Anwendungsbeispiele und Tipps ein. Weiter werden Vorteile gegenüber klassischer Kundenforschung, Probleme in verschiedenen Domänen und firmenkulturelle Hindernisse diskutiert. Da ersteres bereits mehrfach zusammengefasst wurde (z.B. Chopra, 2012; Kohavi & Longbotham, 2015), wird hierauf nicht mehr eingegangen.

Um Produkte optimal zu gestalten, benötigt es gute Kundenforschung. Dabei zeigt sich, dass herkömmliche Verfahren wie Fokusgruppensitzungen oder Prototypentests meist nur simulierte Situationen darstellen und nur mangelhaft die „wahren“ Bedürfnisse erkennen können bzw. zu teuer sind. An dieser Stelle können A/B Tests als eine kostengünstige Alternative zur Testung von Bedürfnisbefriedigung eingesetzt werden - insbesondere in den Software- Bereichen. Weiter stellen die Autoren fest, dass trotz der Sinnhaftigkeit von A/B Tests die Einführung dieser Methode oft auf firmenkulturelle Probleme, Hindernisse oder Gegenargumente trifft: 1) So besitzen Entwicklern meist nur geringe statistische Kenntnisse, um A/B Tests sinnvoll umzusetzen, was mangelhaft durchgeführte Tests zur Folge hat. 2) Entwickler können sich in ihrem Job oder Ansehen angegriffen fühlen, wenn Tests nicht die gewünschten Ergebnisse liefern – insbesondere vor dem Hintergrund, dass das Scheitern einer Idee eher die Regel ist. 3) Arbeitnehmer werden oft an der Quantität und nicht an der Qualität ihrer Ideen/Arbeit gemessen. Da A/B Tests zwar Arbeit in Anspruch nehmen, aber auch zur besseren Qualität neuer Produkte führen, verringert sich die Quantität zu Gunsten der Qualität – was angesichts des eben beschriebenen Bewertungssystems zu

Widerständen führen kann. 4) Ferner werden A/B Tests oft als zu teuer empfunden bzw. 5) als nicht einflussreich genug, da lediglich kleine Unterschiede untersucht werden können. Resümierend gilt es bei solchen Problemen während der Einführung von A/B Tests eine argumentative Überzeugungsarbeit zu leisten.

Stärken und Schwächen: Im Paper wird argumentiert, dass A/B Testing gegenüber herkömmlicher Kundenforschung den Vorteil hat, in einer echten Situation die Bedürfniserfüllung zu testen. Da jedoch immer nur Reaktionen auf Veränderungen gemessen werden (die in Form einer Bedürfnisbefriedigung interpretiert werden), wird es nie möglich sein, mittels A/B Tests Wünsche von Kunden explorativ zu erforschen. Denn zuerst müssen Entwickler Ideen generieren, um diese dann zu testen. Salopp ausgedrückt muss ein Entwickler zuerst darauf kommen, was der Kunde will, bevor er es testen kann.

Erstellt von: Valentin Koob 392628

Autor(en): Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., & Xu, Y.

Titel: Trustworthy online controlled experiments: Five puzzling outcomes explained.

Jahr: 2012

Laufende Nr. im Literaturverzeichnis: 09

Schlagwörter: beispielhafte A/B Studien, Validität von Daten, Interpretation unerwarteter A/B Ergebnisse

Kurzbeschreibung: Das Paper adressiert anhand fünf verschiedener Studien Problematiken bei der Interpretation und Bewertung von Ergebnissen bei kontrollierten Online-Experimenten (= A/B Testing). Dabei werden auf folgende Aspekte eingegangen: 1) Es kann sein, dass das *OEC (Overall Evaluation Criterion)* unterschiedliche langfristige und kurzfristige Implikationen für den Erfolg des Unternehmens haben kann (z.B. Maximierung von Gewinn vs. Kundenzufriedenheit). 2) Es kann sein, dass technische, methodenfremde Aspekte (z.B. die Datenübertragung zum Server) einen Einfluss auf die gesammelten Metriken haben. 3) Erschwerung der Interpretierbarkeit erster Ergebnisse, durch eine erhöhte Variabilität in den Metriken aufgrund von *Novelty Effects* (= Personen erkunden die Testumgebung) bzw. *Primacy Effects* (Erfahrungen von Personen verhindern die effiziente Nutzung des neuen Features, Elements etc.) zu Beginn einer Studie. 4) Es kann bei gewissen Metriken nicht von Nutzen sein, weitere Personen zu erheben, da trotz dessen keine höhere Power erzielt wird. 5) Durch *Carryover Effects* (= Verhalten von Personen wird durch vorherige Studien beeinflusst) kann es zur systematischen Konfundierung von Ergebnissen in einer nachfolgenden Studie kommen.

Als Beispiel für die erste Problematik stellte Bing trotz schlechter Suchergebnisse durch einen technischen Fehler fest, dass die Anzahl der Suchanfragen und Aufrufe von Werbung deutlich anstiegen. Wenngleich dies kurzfristige monetäre Vorteile dem Unternehmen bringt, sinkt langfristig die Kundenzufriedenheit, was auf Dauer gesehen monetäre Nachteile mit sich zieht. Resümierend betonen die Autoren, dass sich Operationalisierungen nicht immer objektiv mit dem decken, was sie augenscheinlich messen sollen und, dass Ergebnisse sowohl von technischen als auch psychologischen Effekten beeinflusst sein können. Umso wichtiger sei es zu erkennen, wenn Ergebnisse unglaubwürdig sind.

Stärken und Schwächen des Papers: Das Paper betont wichtige und oft vernachlässigte psychologische, technische und statistische Einflussfaktoren auf die Validität von Ergebnissen in kontrollierten Online-Studien. Dabei stehen Aspekte der Operationalisierung

im Vordergrund. Wichtig ist hierbei, dass sich insbesondere die Punkte 1, 3, 4, 5 nicht nur auf den Online-Bereich beschränken. Negativ ist zu bewerten, dass sich Vorschläge und Mahnungen aus der Analyse vergangener Studien ergeben, d.h. lediglich auf post-hoc Überlegungen basieren. Weitergehend kann aus praxisorientierter Sicht kritisiert werden, dass sich die oben beschriebenen Einflussfaktoren schwer abschätzen, entdecken oder gar messen lassen.

Erstellt von: Valentin Koob 392628

Autor(en): Kohavi, R., Deng, A., Longbotham, R. , Xu, Y.

Titel: Seven Rules of Thumb for Web Site Experimenters

Jahr: 2014

Laufende Nummer im Literaturverzeichnis: 10

Schlagwörter (3 bis 5): Einführung, Experimentplanung, Ergebnisanalyse

Kurzbeschreibung: Basierend auf vielen durchgeführten Experimenten, stellen Kohavi und seine Kollegen sieben Daumenregeln dar. Sie empfehlen vor allem, dass man sich Ergebnisse genau anschaut, keine sofortige Veränderungen erwartet, und betonen, dass Ergebnisse repliziert werden müssen. Die Wichtigkeit eines schnellen Servers sowie der Benutzung von simplen Design werden hervorgehoben. Eine Stärke dieses Papers sind die zahlreichen zitierten wissenschaftlichen Quellen und die praktischen Beispiele und Illustrationen.

Erstellt von: Margarita Mishinova, 396046

Autor(en): Ron Kohavi, Randall M. Henne, Dan Sommerfield

Titel: Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO

Jahr: 2007

Laufende Nr. im Literaturverzeichnis: 11

Schlagwörter: Leitfaden aus der Praxis, Statistik, Stichprobengröße, Technische Anforderungen, Ergebniskommunikation

Kurzbeschreibung: Der Artikel bietet einen Leitfaden für die Durchführung von Online Usability Tests. Die Wichtigkeit von kontrollierten Experimenten wird hervorgehoben, da die erhobenen Daten der menschlichen Intuition überlegen sind. Dies ist anhand des Beispiels der Microsoft Help Funktion nachvollziehbar.

In einem A/B Test wird jedem Nutzer zufällig eine von zwei Varianten zugeordnet. Wichtig ist, vor dem Beginn der Datenaufzeichnung das Evaluationskriterium zu wählen, anhand dessen die Daten ausgewertet werden. Das Evaluationskriterium ist die abhängige Variable. Als Faktorstufen oder Levels werden die unterschiedlichen Versionen bezeichnet. Andere wichtige Begriffe sind: Experimentaleinheit, Null-Hypothese, Power und Standardabweichung. Unter einem A/A Test versteht man einen Test, bei dem beide Versionen gleich sind. Er wird verwendet um das System des Experiments zu testen und um die Varianz zu schätzen.

Das für den Test verwendete Alpha-Level (meist 5%), die gewünschte Power, der Standardfehler und die Standardabweichung sind wichtig, um die benötigte Stichprobengröße zu ermitteln. Hierbei sollte auch bedacht werden, dass nur ein kleiner

Prozentsatz der Nutzer Käufe tätigt, was häufig das Evaluationskriterium ist.

Mögliche Erweiterungen bei Online Tests sind: Treatment Ramp-Ups, bei denen schrittweise der Prozentsatz der Nutzer in der Experimentalbedingung erhöht wird; Automation, die schnelle Optimierung ermöglicht; und Experimente, die bei der Software Migration helfen.

Die Begrenzungen der A/B Tests liegen darin, dass die Gründe für die Effekte unbekannt bleiben, dass es schwer ist Langzeit-Effekte zu evaluieren, dass psychologische Biases einen Einfluss haben können (z.B. Primacy Effekt) und dass User Unterschiede zu anderen Versionen bemerken können.

Für die technische Umsetzung ist besonders die Randomisierung relevant, hierbei ist auch wichtig, dass ein Nutzer immer dieselbe Version sieht. Dies funktioniert mit einem Random Number Generator und der Nutzung von Cache. Alternativ ist die Nutzung einer Hash Funktion möglich in Kombination mit einer einzigartigen Identifizierung jedes Nutzers, oft durch Cookies.

Abschließend werden Ratschläge für die Praxis gegeben. Im Bereich der Datenanalyse: Die Daten genau auszuwerten, auch in der Hinsicht auf Teilgruppen; nicht zu lange zu testen; jeweils nur eine Veränderung zu testen, aber sehr unterschiedliche Designs auszuprobieren. Im Bereich der Testausführung: A/A Tests verwenden; die minimale Stichprobengröße vorher festzulegen; gleichgroße Gruppen zu verwenden; und lange genug zu testen, um Wochentageffekte auszuschließen.

Im Bereich des Unternehmens: Das Evaluationskriterium sollte von Anfang an festgelegt sein; Veränderungen, die als nicht signifikant getestet wurden, sollten nicht genutzt werden. Außerdem sollten Designentscheidungen zunehmen aufgrund von durch Experimenten generierte Daten getroffen werden.

Stärken und Schwächen: Der Leitfaden ist eine einfache Einführung, die nur geringes psychologisches und technisches Vorwissen benötigt wird. Grundlegende Konzepte werden nachvollziehbar erklärt und definiert. Viele der Themen werden allerdings nur oberflächlich behandelt. Punkte, die nur kurz angeschnitten werden, sind meist mit Quellen hinterlegt, sodass man mehr dazu erfahren kann. Einige Quellen verweisen jedoch auf Websites, die nicht mehr existieren. Sie sind also nicht überprüfbar.

Es werden Praxistipps gegeben und die Unterschiede zwischen Theorie und Praxis werden berücksichtigt. Der Leser erhält klare Anweisung zum Vorgehen bei der Datenerhebung und Datenanalyse. Die kritische Betrachtung von A/B Tests fällt kurz aus. Es werden Begrenzungen, jedoch keine klaren Kritikpunkte an A/B Tests genannt.

Erstellt von: Julia Eberhardt 388742

Autor(en): Kohavi, R., Longbotham, R.

Titel: Unexpected results in online controlled experiments

Jahr: 2010

Laufende Nummer im Literaturverzeichnis: 12

Schlagwörter (3 bis 5): Experimentplanung, Ergebnisanalyse

Kurzbeschreibung: Kohavi und Longbotham gehen in diesem Paper auf zahlreichen unerwarteten Ergebnisse in Experimenten ein, indem sie auf subtile Fehler wie fehlendes Verständnis für Benutzerverhalten, Software und Design probleme hinweisen. Eine skeptische Haltung gegenüber überraschenden Resultate wird empfohlen. Die detaillierte Beschreibung der Experimente stellt eine Stärke des Papers.

Erstellt von: Margarita Mishinova, 396046

Autor(en): Kohavi, R., & Longbotham, R.

Titel: Online controlled experiments and A/B tests

Jahr: 2015

Laufende Nr. im Literaturverzeichnis: 13

Schlagwörter: Schritte von A/B Testing, kontrollierte Online-Studien, Versuchsplan

Kurzbeschreibung: Der Artikel schildert wichtige Kernaspekte der Planung, Auswertung und Kontrolle von A/B Tests bzw. kontrollierten Online-Studien. Darunter versteht sich der Test zweier bzw. mehrerer Varianten eines Produkts (= Website, App-Design etc.), meist mittels einer experimentellen Bedingung und einer Kontrollgruppe. Wichtig ist hierbei, dass sich die Varianten nur in einem einzigen Aspekt unterscheiden. Schritte bei der Planung sind hierbei: 1) Entwicklung der Varianten, 2) des Randomisierungsalgorithmus (= Auswahl relevanter Personen) sowie 3) der messbaren Variablen zur Untersuchung der Unterschiede. Dabei wird insbesondere empfohlen eine Schlüsselvariable (= *OEC*, *Overall Evaluation Criteria*) zu erstellen. Diese soll möglichst gut die Unternehmensziele repräsentieren sowie die Grundlage der späteren Entscheidung (= „welche Variante ist jetzt besser?“) bilden. Weitergehend benötigt es die Planung der 4) statistisch notwendigen Stichprobengröße und 5) Dauer der Studie. 6) Zuletzt sollten möglichst externe Einflussfaktoren ausgeschlossen werden. 7) Nach der Durchführung werden meist Konfidenzintervalle berechnet und auf Basis dieser eine Entscheidung getroffen. Zur Gewährleistung der Validität werden im Online-Bereich diagnostische Tools empfohlen, welche bspw. signalisieren, wenn Probanden ungleichmäßig auf die Bedingungen verteilt werden. Zusätzlich kann mittels Vorstudien ohne Variantenmanipulation (sog. A/A Tests) geprüft werden, ob systematische Einflüsse existieren.

Stärken und Schwächen des Konzepts: A/B Tests bieten eine breit einsetzbare, einfache, leicht interpretierbare und aussagekräftige Möglichkeit entwickelte Ideen gegeneinander zu testen. Dies ist insofern wichtig, da diese Ideen selten den Effekt erzielen, für den sie eigentlich entwickelt worden sind. Dabei ist es vor allem im Software-Bereich einfach, eine große Stichprobe zu erheben, sowie die Testvarianten zu implementieren.

Problematisch im Online-Bereich ist, dass die Zuordnung auf Basis von Cookies stattfindet. Dies bedeutet, dass die konsistente Messung einer Person auf einem Gerät oder über dieses Gerät hinaus nicht gewährleistet werden kann. Das Untersuchungsobjekt ist somit ein Cookie und nicht eine Person. Bezogen auf den Offline-Bereich stellt sich kostentechnisch vor allem das Problem der ausreichenden Akquirierung von Probanden und der Implementierung der Testvarianten (z.B. die Entwicklung eines neuen Touchscreens als Prototypen). Vergleicht man aus diesem Grund nur wenige Probanden pro Testversion miteinander, ist lediglich eine deskriptive Analyse möglich. Ferner sind A/B Tests durch die Veränderung lediglich eines einzigen Aspekts kleinschrittig. In der Folge benötigt es viele Tests bis eine optimale Lösung gefunden wurde. Auch dies steht dem aufwändigen und kostenintensiveren Einsatz im Offline-Bereich entgegen. Zuletzt gilt, dass der einmalige „Verlierer“ in einem Test nicht nochmal betrachtet wird. Sieht man dies vor möglichen falsch positiven Testentscheidungen durch mangelnde Power, so könnte es sein, dass eine eigentlich bessere Variante nicht noch einmal getestet wird.

Erstellt von: Valentin Koob 392628

Autoren: Kohavi, R., & Thomke, S.

Titel: The Surprising Power of Online Experiments

Jahr: 2017

Laufende Nummer im Literaturverzeichnis: 14

Schlagwörter(3-5): finanzielle Auswirkungen, Einsatzbeispiele, Qualitätssicherung der Daten

Kurzbeschreibung: Der Artikel startet mit einer Anekdote: Ein Mitarbeiter von Microsoft Bing hatte einen Plan zur Veränderung der Werbeanzeige, die aufgrund anderer Prioritäten jedoch nicht zeitnah umgesetzt wurde. Ein halbes Jahr später wurde die Umgestaltung doch umgesetzt und mit einem kostengünstigen A/B-Test geprüft. Die Auswertung ergab eine Steigerung der Einnahmen durch die Werbeanzeige um 12%, was zeigt, wie unterschätzt das A/B-Testing bis dahin war. Heute wird A/B-Testing von vielen großen Firmen online eingesetzt, und die Wirkung ist groß. Sowohl die Nutzerzahlen als auch der Gewinn durch Werbeanzeigen von Bing konnten durch A/B-Testing gesteigert werden. Durch A/B-Testing werden Entscheidungen von Unternehmen wissenschaftlich begründet, was Zeit und Ressourcen spart. A/B-Testing kann von jedem Unternehmen als kostengünstige, schnelle und effektive Methode zur Datenerfassung genutzt werden. Die Veränderungen der B-Version gegenüber der ursprünglichen Version A können dabei sehr klein sein und trotzdem sehr große Wirkungen haben. So wurde durch A/B-Testing herausgefunden, dass eine Verbesserung der Ladezeiten von Suchergebnissen um 100 Millisekunden einen jährlichen Mehrgewinn von 18 Mio. US-Dollars für Bing bedeutet. Für Unternehmen bedeutet der Einsatz von A/B-Testing auch die Minimierung von Risiken, da jede Veränderung auf einer Website auch negative Folgen haben kann. Eine Möglichkeit, die Qualität der ermittelten Daten zu prüfen, ist A/A-Testing. Dabei wird eine Version A gegen sich selbst getestet, wobei zu zeigen ist, dass es keine statistisch signifikanten Unterschiede in der Datenerhebung der zwei identischen Websites gibt. Bei überraschenden Ergebnissen sollte der Test wiederholt werden, um die Validität zu prüfen. So geschehen bei sehr kleinen Veränderungen von Schriftfarben bei Bing. Durch leicht dunkleres Blau und Grün bei Titeln sowie helleres Schwarz bei Bildbeschriftungen ergaben sich so starke Verbesserungen, dass der A/B-Test erneut durchgeführt wurde. Da dasselbe Ergebnis erzielt wurde, wurden die Veränderungen übernommen und der jährliche Gewinn dadurch liegt für Bing bei über 10 Mio. US-Dollar. Dies zeigt, dass die durch A/B-Testing generierten quantitativen Daten sehr wichtig für ein Unternehmen sein können, obwohl keine Gründe für das Verhalten der Nutzer ermittelt werden. Für die Sicherung der Datenqualität sollten auch Ausreißer in der Statistik geprüft werden, sowie auffallende Unterschiede in den Stichproben, da diese einen direkten Einfluss auf das Ergebnis des A/B-Tests haben. Auch die Erfahrung der Testpersonen kann Ergebnisse verzerren, weshalb es wichtig ist, die Stichprobe zu randomisieren. Die Anzahl der Variablen sollte geringgehalten werden, um Kausalitäten eindeutig bestimmen zu können. Generell ist es entscheidend, A/B-Tests in kontrollierten Bedingungen umzusetzen, um Abhängigkeiten ausmachen und die gewonnenen Daten richtig interpretieren und nutzen zu können.

Stärken und Schwächen: In dem Beitrag werden multiple Beispiele für Einsatz von A/B-Testing von großen Unternehmen vorgestellt (Ladezeiten von Websites, Platzierung von Werbung und Buttons, Schriftfarbe etc.). Diese machen die vielen Einsatzbereiche von

A/B-Testing sichtbar und zeigen, wie wichtig A/B-Tests für Unternehmen sind. Viele Beispiele beinhalten dabei auch die Gewinnsummen für die Unternehmen. Dadurch wird die Reichweite kleiner Veränderungen deutlich, die Auswirkungen selbiger greifbar und die Relevanz für Unternehmen, auch in der Risikoabschätzung, erkennbar.

Erstellt von: Meret Vollmann 347528

Autor(en): Lennarz, H.

Titel: AB-Testing

Jahr: 2017

Laufende Nummer im Literaturverzeichnis: 15

Schlagwörter: Objektivität, Ideenfindung, Testing-Tools

Kurzbeschreibung: Das Kapitel gibt praktische Tipps für die eigene Durchführung von A/B-Tests. Beim A/B-Test werden zwei Versionen einer (meist) Webseite gegenübergestellt und anhand von User-Reaktionen praktisch überprüft, welche erfolgreicher ist. Damit umgeht man durch das Verlassen auf Daten subjektive Entscheidungen und überflüssige Diskussionen. Die Ergebnisse sind somit manchmal auch überraschend kontraintuitiv. Das Beispiel einer Autovermietung zeigt, dass sich über Jahre hinweg und nach hunderten A/B-Tests eine Abbildung eines älteren Cabrio-Modells auf der Startseite des Unternehmens konsequent gegen neuere Modelle durchsetzt.

Wichtig bei der Umsetzung von A/B-Testing ist es, die Ziele des Tests und die Erwartung an das Verhalten der User vorher festzulegen, zu formulieren und im Prozess nicht aus den Augen zu verlieren, um valide Ergebnisse zu erzielen. Des Weiteren empfiehlt es sich, pro Test nur eine Veränderung vorzunehmen, damit die Ergebnisse nicht konfundiert sind. Auch ist es wichtig, genügend Traffic auf der Seite zu erzielen, damit statistisch relevante und valide Ergebnisse in möglichst kurzer Zeit generiert werden können. Ein weiterer Punkt, der Beachtung finden sollte, ist die Nichtkopierbarkeit von Testergebnissen. Diese sind individuell für jede Website anders und sollten lediglich als Ideen für das eigene System aufgefasst werden.

Wichtig ist auch, herauszufinden, welche Elemente der Webpage den größten Einfluss auf z.B. die Konversionsrate haben. In einigen Webtracking-Tools sind sogenannte Konversionstrichter enthalten, die genau dies tun. So kann es helfen, genau zu filtern, wann User z.B. am häufigsten abspringen, um an dieser Stelle anzusetzen und Veränderung zu schaffen. Auch das ICE-Modell (Impact, Confidence, Ease) kann bei der Ideenfindung helfen. Es ordnet Testideen nach den Kriterien Einfluss, Wahrscheinlichkeit und Einfachheit und gibt so einen guten Überblick über die zu testenden Elemente.

Es gibt inzwischen viele Testing-Tools, die als Demoversion kostenlos zur Verfügung stehen. So ist es auch Anwendern, die ohne statistische Kenntnisse testen, möglich, A/B-Tests durchzuführen. Dies vereinfacht die Handhabung enorm.

Es wird betont, dass zu viel Testen von einer endgültigen Entscheidungsfindung ablenken kann und User verwirren könnte, wenn sie immer wieder mit verschiedenen Layouts konfrontiert werden.

Stärken und Schwächen: Die wichtigsten Punkte, die es beim A/B-Testing zu beachten gilt, werden genannt. Das Kapitel bietet allerdings nur eine grundsätzliche Einführung in die Thematik und klammert viele Probleme, die beim Testen entstehen können, sowie methodische Bedenken, aus. Auch wird nicht darauf eingegangen, das bei der Verwendung

von Testing-Tools ohne Statistikkenntnisse möglicherweise invalide Ergebnisse entstehen können.

Erstellt von: Judith Mengel 392609

Autor(en): Florian Lettner, Clemens Holzmann, and Patrick Hutflesz

Titel: Enabling A/B Testing of Native Mobile Applications by Remote User Interface Exchange

Jahr: 2013

Laufende Nr. im Literaturverzeichnis: 16

Schlagwörter: Mobile Apps, Native Interfaces, technische Umsetzung, Android

Kurzbeschreibung: A/B Testing ist bei Webseiten sehr populär und einfach umsetzbar. Bei mobilen Apps hingegen ist es schwieriger. Das Paper präsentiert eine Möglichkeit das Nutzer Interface bei Android Apps zu variieren und dadurch A/B Tests durchzuführen.

Der aktuelle Stand der Technik wurde anhand von sechs aktuellen Ansätzen für A/B Testing bei mobilen Apps ermittelt. Vier der getesteten Services nutzen eine Web-View Ansatz bei dem Webseiten in der App in einem integrierten Browser ausgeführt werden. Dadurch lässt sich das Design leicht verändern, die Methode geht jedoch mit einigen Sicherheitsproblemen einher und ermöglicht nicht das individuelle Nutzungsgefühl der App zu erhalten. Zwei Services nutzten einen Native-Ansatz, bei dem die Variation in der App stattfindet. Das Nutzungsgefühl der App bleibt für den Anwender hierbei bestehen, jedoch muss für jede Änderung in der App eine neue Version veröffentlicht werden.

Um A/B Testing durchzuführen, muss im ersten Schritt die App entwickelt und veröffentlicht werden. Im zweiten Schritt weist ein Algorithmus den unterschiedlichen Nutzern eines der beiden Layouts zu. Im dritten Schritt wird ein methodischer Ansatz zur Analyse und zum Vergleich der Nutzungsdaten benötigt.

Die Autoren haben eine Methodik und einen Prototyp entwickelt, die es ermöglichen A/B Testing durch Native User Interfaces auf Android-Geräten durchzuführen, ohne dabei die App erneut veröffentlichen zu müssen.

Zuerst sollte ein neues Layout entwickelt werden, das auf XML basiert. Dieses wird dann auf einen Server geladen, wo es in ein binäres Format übersetzt wird. Um die Layouts auf die Apps zu verteilen, kann ein Cloud-to-Device Datentransfer genutzt werden. Die Autoren selbst nutzen dazu die Google Cloud Messaging API. Dadurch erhalten die Nutzer die neuen Layouts, wenn sie verfügbar sind. Die Layouts werden im zeitlich begrenzten Datenverzeichnis im Android File System gespeichert. Dort kann durch die Nutzung von Java Reflection auf sie zugegriffen werden. Durch den LayoutInflater können diese dann angezeigt werden.

Dieser Ansatz ist sicherer, erhält das eigene Nutzungsgefühl der App und benötigt keine erneute Veröffentlichung der App. Der Prototyp ermöglicht es die User Interfaces auszutauschen und deren Funktionsweise zu ändern, sowie das Hinzufügen oder Entfernen von einzelnen Elementen wie Buttons oder Textfelder.

Stärken und Schwächen: Der beschriebene technische Ansatz scheint dem Native-Ansatz mit erneuten Veröffentlichung der App und dem Web-View Ansatz überlegen. Die vorgeschlagene Vorgehensweise ist exakt und schlägt konkrete Tools für die einzelnen Schritte des Austauschs und der Veränderung der Layouts dar. Das Paper wurde jedoch im Jahr 2013 veröffentlicht und aufgrund der schnellen technischen Entwicklung ist das

Vorgehen nicht mehr auf dem aktuellen Stand der Technik, auch wenn der weiterhin funktioniert. Trotzdem kann es Einblick in die grundlegenden Prozesse geben, die aus technischer Sicht beim A/B Testing bedacht werden sollen. Bei den genannten Tools und spezifischen Vorgehensweisen sollten jedoch neuere Alternativen in Betracht gezogen werden.

Erstellt von: Julia Eberhardt 388742

Autoren: Martin, B., & Hanington, B.

Titel: A/B-Testing. In Universal Methods of Design – 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions

Jahr: 2012

Laufende Nummer im Literaturverzeichnis: 17

Schlagwörter (3-5): Quantitative Methode, iterative Anwendung, Websitegestaltung

Kurzbeschreibung: A/B Testing wird als Methode vorgestellt, die dazu dient herauszufinden, welche Version des Designs statistisch besser ein zuvor definiertes Ziel erreicht. Die erhobenen Daten sind dabei rein quantitativ und liefern keine Gründe über das Verhalten der Testpersonen. Es wird betont, dass A/B Testing als Ergänzung zu qualitativen Methoden eingesetzt werden sollte. Als Beispiel für den Einsatz dient eine Website mit einem Anmeldeformular. Um die Anmeldezahlen zu erhöhen, werden mehrere potentiell relevante Faktoren genannt, die sich durch A/B Tests prüfen lassen: der Text der Website und seine Gestaltung, das Layout und die Gestaltung und Platzierung des Anmelde-Buttons. A/B Testing wird als quantitative Methode zur Evaluation des Nutzerverhaltens im Designprozess von Websites o.ä. eingestuft. Am Beispiel von eBay wird gezeigt, dass es wichtig ist, immer wieder A/B Tests durchzuführen. Obwohl vorherige Tests gezeigt hatten, dass User eine größere Anzahl von Artikeln pro Seite präferieren, ergab der A/B Test mit verschiedenen Bildgrößen, dass größere Artikelbilder zu mehr Verkäufen führt. Auch wenn dadurch weniger Artikel pro Seite gezeigt werden können, hat die Bildgröße einen größeren Einfluss auf das Kaufverhalten als die Artikelanzahl.

Stärken und Schwächen: Der Beitrag liefert ein leicht verständliches Beispiel für den Einsatz von A/B Testing, das sich als Einstieg in das Thema eignet und zur groben Einordnung der Methode dient. Für tiefergehendes Verständnis der Methode ist dieser Einblick allerdings zu oberflächlich.

Erstellt von: Meret Vollmann 347528

Autor(en): Jakob Nielsen

Titel: Putting A/B Testing in Its Place

Jahr: 2005

Laufende Nr. im Literaturverzeichnis: 18

Schlagwörter: Vorteile von A/B Testing, Nachteile von A/B Testing, UX-Research, Webseiten-Testing

Kurzbeschreibung: A/B Testing wird häufig bei Webseiten durchgeführt. Hierbei werden die Nutzer in zwei Gruppen aufgeteilt und jeder Gruppe wird einer Version gezeigt. Die Gruppen können gleichgroß oder unterschiedlich groß sein. In multivariaten Test werden auch mehr als ein Designelement verändert.

Die vier großen Vorteile von A/B Testing sind, dass (1) echtes Nutzerverhalten im Feld gemessen wird, (2) auch kleine Performanzunterschiede statistisch signifikant erkannt werden können, (3) bei im Widerspruch stehenden qualitativen Guidelines getestet werden kann, welche für den konkreten Fall von größerer Bedeutung ist und (4) es günstig ist.

In den meisten Fällen überwiegen die Nachteile jedoch. A/B Testing kann nur in einem limitierten Bereich angewendet werden, nämlich wenn es nur ein klar definiertes Ziel gibt und sich dieses auch durch Interaktion mit dem Computer messen lässt. Dies ist in der Praxis nur selten der Fall. Ein weiterer Nachteil ist, dass A/B Testing nur mit fertigen Designs möglich ist und dadurch kein iterativer Designprozess stattfindet. A/B Testing wird im Regelfall nur dazu verwendet kurzzeitige Effekte zu messen. Langzeitmessungen sind möglich, jedoch müssten dazu beide Versionen über Monate bis Jahre getestet werden. Da dies kaum passiert, ist ein Langzeiteffekt nicht messbar, der z.B. auf Nutzungszufriedenheit hindeuten könnte.

Als größter Nachteil wird genannt, dass A/B Testing keine Erkenntnis darüber liefert, warum eine Version besser funktioniert als die andere. Allgemeine Erkenntnisse werden nicht generiert und jede Veränderung muss einzeln getestet werden. Zusätzlich gibt es keine Möglichkeit Daten über andere Aspekte der Seite, die nicht variiert wurden, zu erhalten. Vielleicht liegt der Grund, warum eine Seite nicht erfolgreich ist, jedoch an einem ganz anderen Punkt, den Nutzer verbal leicht ausdrücken könnten, der durch A/B Testing jedoch nicht erkennbar ist. Dies birgt die Gefahr große Verbesserungspotentiale zu übersehen.

Deshalb wird empfohlen A/B Testing nicht als erste und/oder einzige Methode anzuwenden, sondern es als Ergänzung zur qualitativen Forschung zu sehen.

Stärken und Schwächen: Die jahrelange Berufserfahrung des Autors im Bereich Usability Research spricht für eine fundierte fachliche Grundlage der aufgelisteten Vor- und Nachteile von A/B Testing. Diese sind prägnant dargestellt und mit Anwendungsbeispielen unterlegt, sodass sie gut in der Praxis nachvollziehbar sind. Die erwähnte Schwäche, dass keine Aussagen über nicht gemessene Parameter getroffen werden können, erscheint jedoch trivial. Die meisten Fakten sind mit Quellenangaben unterlegt, diese verweisen jedoch nur auf weitere Artikel derselben Firmenseite, deren Autoren vermutlich ähnliche Ansichten und Interessen haben. Der Mangel an externen Quellen stellt die Güte der aufgelisteten Fakten in Frage. Zusätzlich ist der Artikel von einem kommerziellen UX-Research Unternehmen. Die Objektivität ist deshalb möglicherweise nicht gegeben, da das Verkaufsinteresse der eigenen Methoden Grund für die überwiegend kritische Darstellung von A/B Testing sein kann.

Erstellt von: Julia Eberhardt 388742

Autoren: Quirnbach, S. M.

Titel: Suchmaschinen: User Experience, Usability und nutzerzentrierte Website-Gestaltung

Jahr: 2012

Laufende Nummer im Literaturverzeichnis: 19

Schlagwörter (3-5): Voraussetzungen, Vorteile, Nachteile

Kurzbeschreibung: A/B-Testing ist ein statistisches Verfahren zum Testen von verschiedenen Versionen einer Website in Bezug auf ein oder mehrere bestimmte Ziele, z.B. wirtschaftliche Faktoren oder die Steigerung der Wiederkehrrate. Es ist eine zeit- und

kostenintensive Methode, da Software benötigt wird und die Tests über einen langen Zeitraum durchgeführt werden müssen. Die Kosten sollen durch die Verbesserung der Website wieder ausgeglichen werden. Je nach Anzahl der verschiedenen Versionen wird in A/B-Testing (bei zwei Varianten) und multivariates Testing (mehrere Varianten) unterschieden. Zur Vorbereitung der Tests müssen Hypothesen formuliert sein, die während des Tests überprüft werden sollen. Ob eine Version das definierte Ziel besser erreicht als eine andere, wird an zuvor festgelegten Zielwerten ermittelt. Für das Testing selbst ist eine sogenannte Webanalytic-Software notwendig, um die Nutzerdaten statistisch zu erfassen. Zum Vergleich muss eine Messung der unveränderten Website gemacht werden. Zudem muss vor Beginn festgelegt werden, ob die Nutzerdaten über eine Software oder über Cookies gesammelt werden. Im Anschluss werden die Vor- und Nachteile dieser Methode genannt. Die Vorteile beinhalten: Einbindung des Testings in reale Nutzung der Website durch echte User; zügige Sammlung der Daten in Abhängigkeit der realen Nutzungsdaten; unbewusstes Verhalten des Nutzers durch Unkenntnis über die andere(n) Websiteversion(en) führt zur Authentizität der gesammelten Daten. Zudem können mehrere Versionen zeitgleich getestet und optimiert werden, sowie kleinste Veränderungen der Website auf die Zielwerte geprüft werden. Nachteile der Methode hingegen sind: hohe Anfangskosten durch Software und Anfangsmessung, die bei Widerlegung der Arbeitshypothese nicht amortisiert werden können; ein notwendigerweise hohes Maß an Wissen zur Umsetzung der Methode; Gründe für das Verhalten der User werden aus der Methode nicht geklärt. Damit ist es ein rein vergleichendes Ergebnis. Der Einsatz der Methode ist erst im Betrieb der Website möglich, nicht schon in der Entwicklung. Zum Abschluss des Kapitels über A/B-Testing werden einige Beispiele für den Einsatz von A/B-Testing und multivariatem Testing genannt. Diese gehen über den Einsatz von Werbung über die Gestaltung sowie Anordnung von Schrift und Buttons.

Stärken und Schwächen: In dem Artikel werden die Voraussetzungen für A/B-Testing genannt. Es wird deutlich gemacht, dass eine Hypothese zu Beginn des Testings absolut notwendig ist, um das Testing zielgerichtet durchführen zu können. Zur Veranschaulichung werden beispielhafte Hypothesen genannt, die verdeutlichen, dass eine Hypothese prägnant formuliert und ausführlich sein muss, um effektives Testing zu gewährleisten, was durchaus positiv zu bewerten ist. Die Relevanz der Basismessung wird in dieser Quelle sehr deutlich gemacht, da ohne sie keine Vergleichswerte zur Verfügung stehen. Die Stärken der Methode sind ebenso wie die Schwächen prägnant erläutert. Während die genannten Stärken nachvollziehbar und überzeugend sind, wird als Schwäche der Methode unter anderem „detailliertes Wissen zur Umsetzung der Hypothesen in die Varianten“ genannt. Dies ist ein streitbares Argument, weil jede Methode auf jedem Fachgebiet detailliertes Wissen voraussetzt, um effektiv zu sein. Positiv anzumerken ist jedoch, dass der Einsatzzeitpunkt der Methode als Schwäche genannt wird. Weil es Kosten verursacht, die Gestaltung nachträglich testen und gegebenenfalls ändern zu müssen, wäre ein Einsatz von A/B-Testing in der Entwicklungsphase der Website von Vorteil. Dies ist aufgrund der Art dieser Methode jedoch nicht möglich. Auch, dass durch die Methode keine Gründe für das Verhalten der Nutzer ermittelt werden, ist ein wichtiger Kritikpunkt von A/B-Testing. Dies führt nämlich dazu, dass die Methode iterativ auf verschiedene Versionen einer Website angewendet werden muss, was viele Testdurchläufe und damit hohe Kosten verursacht. Die Einsatzbereiche, die Quirnbach zum Ende des Kapitels nennt, sind recht informativ, aber keinesfalls vollständig. Hier wäre es sinnvoll, die Beispielhaftigkeit der Einsatzbereiche

zu betonen, oder weitere allgemeine Bereiche zu nennen.

Erstellt von: Meret Vollmann 347528

Autor(en): Ronicke, Ph., Kleinknecht, J., Krüger, J.D.

Titel: Leitfaden A/B-Tests auf Landingpages

Jahr: 2016

Laufende Nummer im Literaturverzeichnis: 20

Schlagwörter (3 bis 5): Landingpages, Experimentplanung, Testelemente, Testvarianten

Kurzbeschreibung: Der praktische Leitfaden einer Plattform für Conversion-Optimierung konzentriert sich auf Landingpages also Seiten, auf der ein Besucher „landet“, wenn er ein Banner oder einen Textlink für zum Beispiel Angebotsanfragen, Teilnahme an einem Gewinnspiel oder Bestellung in einem Online-Shop. Der Leitfaden geht auf die Elemente ein, die bei einem A/B-Test getestet werden: Elemente ausblenden, verschiedene Argumentationen, Überschriften, Bilder und Fotos, Texte verkürzen oder verlängern, Schaubilder und Prozesse visualisieren, Call-to-Action, Formulare und Rechner, Telefonnummer.

Der geschätzte Aufwand der Umsetzung sowie der geschätzte Einfluss auf Conversion-Rate werden als einfach, mittel oder hoch hervorgehoben. Weitere Stärke des Leitfadens sind die Illustrationen, die die Beispiele erklären. Eine Schwäche ist, dass keine Quellen für die sogenannten "Hypothesen" ("Weniger Text (und damit Informationen) führt zu einer höheren Conversion-Rate.") genannt werden, sie kommen wohl aus der Praxis.

Erstellt von: Margarita Mishinova, 396046

Autor(en): Schubart, M.

Titel: A/B-Test und Multivariates Testing

Jahr: 2012

Laufende Nummer im Literaturverzeichnis: 21

Schlagwörter: A/B-Testing, Multivariates Testing, Attention-Analyse-Tools

Kurzbeschreibung: Die Internetseite erläutert die Relevanz des A/B-Testing sowie multivariaten Testings für den e-Commerce. Diese Verfahren sind Ergänzungen zu Attention-Analyse-Tools, die auf neurotechnologischen Verfahren und Erfahrungswerten basieren und dazu dienen, die Aufmerksamkeit der Nutzer auf bestimmte Aspekte zu erhöhen. A/B-Testing ist vergleichsweise leicht durchzuführen, da jeweils nur zwei Versionen getestet werden und die Ergebnisse einfach interpretierbar sind. Dabei wird betont, dass man dennoch abwägen sollte, inwieweit nur kleine entdeckte Unterschiede ökonomisch zu rechtfertigen sind. Da jeweils nur eine Veränderung getestet werden kann, sollte eine Priorisierung hinsichtlich der Umsetzung erfolgen.

Multivariates Testing dagegen hat den Vorteil, dass viele verschiedene Versionen gleichzeitig getestet werden können. Dabei ist es auch möglich, Abhängigkeiten zu entdecken. D. h., dass z.B. ein Element A in Kombination mit den Elementen B und C bessere Ergebnisse erzielt als mit Element D. Allerdings ist multivariates Testing um einiges aufwändiger als A/B-Tests und in der Umsetzung um einiges schwieriger. Der Erfolg des multivariaten Testings ist abhängig vom auf der Seite zur Verfügung stehenden Traffic. Wichtig ist es also, im Vorhinein abzuwägen, wie viel Traffic die Seite bietet, wie hoch die

Conversion-Rate ist, und was die erwarteten Veränderungen sind. Davon wird abhängig gemacht, wie viele Versionen getestet werden können und wie lange die Laufzeit sein soll. Durch die Komplexität des multivariaten Testings bietet es sich hier besonders an, professionelle Beratung in Anspruch zu nehmen.

Stärken und Schwächen: Der Artikel liefert lediglich einen groben Überblick. Somit eignet er sich gut für einen ersten Einstieg in die Thematik und eine grobe Abgrenzung der beschriebenen Verfahren. Für weiterführende und tiefere Informationen ist er allerdings zu oberflächlich. Auf statistische Methoden und Probleme wird nicht eingegangen.

Erstellt von: Judith Mengel 392609

Autor(en): Siroker, D., Koomen, P.

Titel: A/B Testing: The Most Powerful Way to Turn Clicks Into Customers

Jahr: 2013

Laufende Nummer im Literaturverzeichnis: 22

Schlagwörter (3 bis 5): Experimentplanung, Instrumente, Statistik

Kurzbeschreibung: Die Autoren sind die Gründer von Optimizely.com, einer Webplattform für Testing und Optimierung von Websites. Das Buch basiert auf mehr als 300 000 durchgeführten Tests und gibt Einblicke, wie A/B Tests zur personalisierten Nutzererfahrung auf der Website führen können. Die Kapitel über den statistischen Hintergrund heben die wissenschaftliche Grundlage der Methode hervor.

Erstellt von: Margarita Mishinova, 396046

Autor(en): Speicher, M., Both, A., & Gaedke, M.

Titel: Was That Webpage Pleasant to Use? Predicting Usability Quantitatively from Interactions

Jahr: 2013

Laufende Nummer im Literaturverzeichnis: 23

Schlagwörter: Klassifikation, Nutzer-Intention, Usability, Interaktionstracking

Kurzbeschreibung: Der Beitrag beschreibt eine Methode, die die Usability einer Webseite durch die Nutzerinteraktionen und implizites Nutzerverhalten quantitativ erfassen und bewerten soll. Die Autoren argumentieren, dass klassisches A/B-Testing davon ausgeht, dass eine höhere Conversion-Rate für eine hohe Usability steht, dies aber nicht unbedingt so sein muss. Die Entwicklung und Anwendung der neuen Methode soll dazu beitragen, dass A/B-Testing-Ergebnisse nicht nur auf Conversion-Rate oder Klicks allein beruhen, sondern auch auf Usability. Die Autoren nehmen an, dass Webseiten des gleichen Typs ein ähnliches Nutzerverhalten vorhersagen können.

Eine Pilotstudie zeigte jedoch, dass das Nutzerverhalten bereits für sehr ähnliche Webseiten erheblich voneinander abwich. In der daraufhin durchgeführten Studie lasen 81 Personen einen von 4 Online-Artikeln zu einem eher unbekanntem Thema, mittels derer sie eine spezifische Frage beantworten sollten. Zwei der Artikel waren jeweils länger als die anderen. Außerdem enthielten nur zwei der Artikel jeweils die gesuchte Information. Die Nutzerinteraktion wurde durch Mausbewegungen, Scroll-Verhalten etc. erfasst. Im Anschluss wurden die Teilnehmer zur Usability der jeweiligen Seite befragt. Auch hier fanden sich erhebliche Unterschiede im Nutzerverhalten bei bereits sehr kleinen

Unterschieden zwischen den Artikeln. Auch unterschiedliche Nutzerintentionen (operationalisiert durch den Fakt, dass die spezifische Information nicht in allen Artikeln enthalten war), hingen mit verändertem Nutzerverhalten zusammen. Usability allein schafft also kein ähnliches Nutzerverhalten. Die anfängliche Annahme, dass es möglich sei, die Usability einer Webpage anhand von Trainingsdaten von ähnlichen Webseiten vorherzusagen, kann nicht gestützt werden. Es ist vielmehr notwendig, a) die Webseiten im Vorhinein hinsichtlich ihrer Struktur und ihres Aufbaus zu clustern, um ein möglichst ähnliches Nutzerverhalten erwarten zu können und b) die spezifische Nutzerintention für einzelne Webseiten zu erfassen. Diese zusätzlichen vorhergehenden Schritte können dazu beitragen, einen methodischen Rahmen zu schaffen, um A/B-Testing zu unterstützen und Nutzerverhalten vorherzusagen.

Stärken und Schwächen: Der Ansatz versucht, weitere Variablen in die Analyse von Nutzerverhalten und Usability mit einzubeziehen. Ein somit ganzheitlicherer Ansatz kann dazu beitragen, A/B-Tests zu vereinfachen und zu verbessern.

Allerdings ist die Beschreibung der Methode relativ unkonkret und bedarf weiterer Analysen, um ein gültiges Modell zu entwickeln. Auch ein Clustern von Nutzerinteraktion kann dazu beitragen, stärkere Korrelationen mit Usability zu finden.

Erstellt von: Judith Mengel 392609

Autor(en): Maximilian Speicher, Andreas Both, Martin Gaedke

Titel: Ensuring Web Interface Quality through Usability-Based Split Testing

Jahr: 2014

Laufende Nr. im Literaturverzeichnis: 24

Schlagwörter: Mixed-Model-Ansatz, Webseiten-Testing, qualitative Methoden, quantitative Methoden, neue Anwendung

Kurzbeschreibung: Auf Basis der Usability Norm haben die Autoren das WaPPu („Was the Page pleasant to use“) entwickelt, ein Produkt mit dem sich Split-Tests von Websites durchführen und evaluieren lassen. Das Tool orientiert sich an User Interaction Analysis, Website Checking und A/B Testing.

Das verwendete usabilitybasierte Split-Testing basiert auf einem zweistufigen Prozess. Zuerst werden durch Split-Testing Verhaltensparameter aufgezeichnet (beispielsweise Cursorbewegungen) und zwischen den beiden Varianten verglichen. Die Website kann dazu in verschiedene Komponenten aufgeteilt werden, die sich am Konzept der Areas of Interest orientieren. Bei der Datenauswertung werden bei signifikanten Unterschieden Heuristiken für die Interpretation der quantitativen Daten angewandt. Im zweiten Schritt werden die nicht signifikanten Ergebnisse anhand eines von Nutzer ausgefüllten Usability Fragebogens spezifischen Usability Modellen zugeordnet um genauere Aussagen zur Usability treffen zu können.

Die Nutzer eines des Untertests erhalten einen Fragebogen zur Usability-Messung basierend auf dem INUIT Fragen-Set. Der Usability Fragebogen besteht aus sieben Items auf einer dreistufigen Likert-Skala und ermöglicht die Berechnung eines Gesamtwertes der Usability der Webseite. Die Antworten werden durch Machine Learning mit den aufgezeichneten Website-Interaktionen verbunden und die dadurch erzeugten Modelle können dann auf die restlichen Gruppen zur Schätzung der Usability angewandt werden, dies ermöglicht eine effiziente kostengünstige Vorgehensweise.

Im vorgestellten WaPPU-Modell sind nur A/B Tests möglich, also höchstens zwei verschiedene Varianten. Es werden 27 verschiedene Interaktionsparameter aufgezeichnet. Auf eine Nutzeroberfläche angewandt, entwickelt WaPPU automatisch anhand der gesammelten Daten sieben verschiedene Nutzermodelle. Außerdem können die Verwendung eines Ad-Blockers und die Größe des verwendeten Bildschirms in der Auswertung berücksichtigt werden.

Das Produkt wurde an einer Internet-Suchmaschine der Unister GmbH getestet. Es wurde ein A/B Test durchgeführt bei dem beide Gruppen den Usability-Fragebogen beantworteten. Für den Test wurden die alte Version und eine frisch nach Usability Gesichtspunkten überarbeitete Version verwendet. Die 81 Versuchsteilnehmenden wurden instruiert nach einem Geburtstagsgeschenk für einen Freund zu suchen und dazu die Suchmaschine zu nutzen. Sie wurden in vier Gruppen nach den Variablen Vorerfahrung mit der Suchmaschine (gering/hoch) und Bildschirmart (HD/full HD) eingeteilt. Eine Verringerung der Verwirrung und Ablenkung des Nutzers und eine bessere Lesbarkeit und Informationsdichte werden angenommen, klare Hypothesen wurden nicht aufgestellt.

In der größten Nutzergruppe, der Gruppe ohne Vorerfahrung und mit HD Bildschirm wurde anhand des Fragebogens ein signifikanter Anstieg der Usability durch die neue Version gemessen, außerdem eine signifikant verbesserte Informationsdichte und geringere Ablenkung. Dies wird als Bestätigung für die Effektivität von WaPPU gesehen. In den anderen Gruppen gibt es widersprüchliche Ergebnisse, die dahingehend gedeutet werden, dass Nutzer mit Vorerfahrung weniger empfänglich für Verbesserungen in der Usability sind. Aus der Kombination der Fragebögen und der Interaktionsdaten der Variante A wurden anhand der Interaktionsdaten die Fragebogenergebnisse der Variante B vorhergesagt. Die Korrelation der Vorhersage mit den erhobenen Daten lag zwischen 0,296 und 0,518 in den Items.

Die Autoren planen WaPPU weiter auszubauen und auf die Intentionen der Nutzer zu messen.

Stärken und Schwächen: Das WaPPU-Tool baut sinnvoll auf dem aktuellen Stand der Praxis auf. Die Kombination aus quantitativen und qualitativen Daten bietet zusätzliche Möglichkeiten der Interpretation der Daten. Die dabei angewandten Heuristiken ermöglichen eine hohe Objektivität bei der Interpretation der Cursordaten. Dies bietet die Möglichkeit auch nicht signifikante Ergebnisse der Interaktionsparameter anhand der Nutzermodelle weiter zu erklären.

Die Evaluation des eigenen Verfahrens in Form eines Papers kann durch das Eigeninteresse der Bewerbung des eigenen Tools beeinflusst sein. Die Studie die dazu durchgeführt würde enthält keine klaren Hypothesen, sondern nur vage Annahmen. Bei der statistischen Auswertung wird die Alpha Inflation nicht berücksichtigt, es werden alle Gruppen auf Signifikanzunterschiede überprüft.

Die Interpretation der Ergebnisse ist inkonsistent. Ergebnisse, die mit den vorherigen Annahmen konform sind, werden als Beweis für Funktionalität von WaPPU gesehen. Ergebnisse, die nicht antizipiert wurden, werden als neue Erkenntnisse über die Website angesehen und nicht als Erkenntnis zur Effektivität von WaPPU. Diese Interpretation der Daten entspricht nicht dem Zweck der Studie, zu testen ob WaPPU bereits getätigte Verbesserungen auch als Verbesserungen erkennt.

Erstellt von: Julia Eberhardt 388742

Titel: From Infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks
Jahr: 2015

Laufende Nummer im Literaturverzeichnis: 25

Schlagwörter: kontrollierte Online-Experimente, Messmethode, Praxisbeispiele, Soziale Netzwerke

Kurzbeschreibung: Das Paper stellt exemplarisch die Vorgehensweisen und Anwendungen von A/B-Testing im sozialen Netzwerk LinkedIn dar und diskutiert die Herausforderungen und Limitationen, die damit einhergehen. A/B-Testing ist heutzutage tief in die Entscheidungskultur des Unternehmens verankert und finden in vielen verschiedenen Bereichen Anwendung. Jeden Tag werden ca. 400 Tests gleichzeitig durchgeführt. Das führt zu einer erheblichen Datenfülle, die Grenzen kennt. Die verwendete Testplattform, XLNT, ist auf die Bedürfnisse des Unternehmens zugeschnitten. In dieser Plattform können große Mengen an Daten erfasst, gespeichert und angemessen dargestellt werden. Außerdem berücksichtigt es bisherige Vorgehensweisen und Praxen, integriert die Ingenieurskultur des Unternehmens, ist flexibel einsetzbar und gebrauchstauglich.

Da viele Tests gleichzeitig laufen, verläuft die Randomisierung orthogonal. Jedes Experiment wird in Segmente eingeteilt, die jeweils eine Untergruppe von Unternehmensmitarbeitern (in der 1. Testphase) oder Gruppen von Usern darstellt, und jedes Segment erhält für den Überblick eine ID. Segmentierung ist gleichzeitig wichtig für die Einteilung der User in verschiedene Zielgruppen. Dies kann je nach Anforderung z.B. nach den Daten des letzten Log-ins, dem Land, externen Daten, dem verwendeten Browser oder dem genutzten Device geschehen. Nachdem das Experiment online gegangen ist, errechnet das Programm automatisch Statistiken, Konfidenzintervalle und p-Werte. Auch multidimensionale Darstellungen sind möglich.

Generell werden die Tests in Online-Experimente und Offline-Experimente gegliedert. Offline-Experimente sind z.B. Email-Kampagnen oder Tests mit Usern, die lange nicht online waren. Eine besondere Herausforderung in sozialen Netzwerken ist die Interdependenz der User. So ist es z.B. nur möglich, die Videochat-Funktion zu testen, wenn „befreundete“ User jeweils online sind. Diese Netzwerkeffekte können in einen Bias resultieren, wenn sie nicht in der Analyse berücksichtigt werden. Daher werden User in Cluster eingeteilt. Diese Cluster werden dann im Randomisierungsprozess als eine Einheit betrachtet.

Des Weiteren ist die Integration des Testens in die alltäglichen Abläufe des Unternehmens von zentraler Bedeutung. Auch ein genereller Leitfaden, wie mit unerwarteten Ergebnissen umgegangen werden soll, ist vorhanden, damit weniger geschultem Personal eine Orientierungshilfe gegeben wird. So werden unerwartete Ergebnisse beispielsweise mit kleineren -Niveaus betrachtet als erwartete Ergebnisse. Durch die große täglich generierte Datenfülle wird so vorab sortiert, welche Effekte von besonderer Bedeutung sind, was die Auswertung erleichtern soll.

Stärken und Schwächen: Das Paper gibt einen guten Einblick in die praktische Anwendung von A/B-Tests. Obwohl das Vorgehen spezifisch auf ein Unternehmen abgestimmt ist, gibt es doch Einsichten in A/B-Tests für soziale Netzwerke, die teilweise für Fragestellungen in anderen sozialen Netzwerken gelten können. Darunter zählen z.B. die Einteilung in Zielgruppen und die Vorgehensweise des Randomisierens sowie die parallele Teilnahme der User in mehreren Experimenten.

Das Paper geht nicht auf Fälle ein, die nicht mittels A/B-Tests getestet werden können,

erwähnt jedoch, dass dies in einer Social-Network-Umgebung häufig vorkommt. Außerdem kann es bei der Einteilung von Usern in mehrere Experimentalgruppen gleichzeitig zu Konfundierungen kommen. Dies wird in diesem Paper zumindest nicht berücksichtigt.

Erstellt von: Judith Mengel 392609

Autoren: Young, S. W. H.

Titel: Improving Library User Experience with A/B Testing: Principles and Process

Jahr: 2014

Laufende Nummer im Literaturverzeichnis: 26

Schlagwörter (3-5): sechs Schritte von A/B-Testing, Rolle des Users, click-through rate, drop-off rate, homepage-return rate

Kurzbeschreibung: Das Paper beschreibt die Methodik von A/B-Testing und verdeutlicht sie an einem Beispiel. Zunächst wird die wichtige Rolle des Users in der Gestaltung eines Produktes betont. A/B-Testing ist eine Möglichkeit, permanent User-Präferenzen abzufragen und das Produkt entsprechend anzupassen. A/B-Testing dient im Gegensatz zu Usability-Testing nicht dazu, Usability Probleme zu finden, sondern die beste von mehreren Lösungen zu einem bereits bekannten Problem zu finden. Daher sollte A/B-Testing mit Usability-Testing kombiniert werden. Für den Ablauf eines erfolgreichen A/B-Testings nennt Young folgende Schritte: „1. Forschungsfrage definieren, 2. Forschungsfrage durch Nutzerinterviews verfeinern, 3. Hypothese formulieren, angemessene Methoden und Testmetrik wählen, 4. Experiment aufstellen und durchführen, 5. Daten sammeln und Ergebnisse analysieren, 6. Ergebnisse teilen und Entscheidung treffen“ (wörtl. übersetzt). Diese Schritte werden am Beispiel der Montana State University Library Website angewendet. Auf der Website gibt es drei Kategorien, von denen die dritte Kategorie „Interact“ auffallend geringe Klickzahlen hat. Im ersten Schritt wurde entsprechend die Frage aufgeworfen, woran das liegt. Als Variationen für „Interact“ wurden „Connect“, „Learn“, „Help“, und „Services“ als neue Überschriften entwickelt. In Schritt zwei wurden Nutzer (Studierende) über ihr Verhalten auf der Website, die Gründe für selbiges und ihre Präferenzen für die Umbenennung des Buttons befragt. Die Interviews ergaben, dass „Help“ und „Services“ von Studierenden bevorzugt werden, und dass signifikante Unterschiede in den Ausprägungen der Varianten zu erwarten sind. Durch die Gespräche können auch Vorschläge der User in den Test einbezogen werden. Die Arbeitshypothese entsteht im dritten Schritt. Außerdem werden die Werkzeuge für den Test festgelegt, hier Google Analytics und Crazy Egg zur Auswertung. Analysiert werden sollen die Anzahl der Klicks (click-through rate), die Abbruchrate (drop-off rate), also die Stelle, an der User die Website verlassen, und die Rate der wiederkehrenden Nutzer auf die Hauptseite, nachdem eine Kategorie ausgewählt wurde (homepage-return rate). Letzteres gibt Aufschluss darüber, ob die Kategorie die Informationen enthielt, die der User gesucht hat. Im vierten Schritt wird nun das eigentliche A/B-Testing durchgeführt, indem die fünf Versionen der Website angelegt und hinterlegt werden, sodass die zufällige Verteilung an die Nutzer stattfinden kann. Im fünften Schritt werden nun die Daten gesammelt und analysiert. Das Ergebnis war, dass „Services“ in allen Bewertungsbereichen am besten abgeschnitten hat. Mit 0% drop-off rate und 0% homepage-return rate ist das Ergebnis sehr eindeutig. Schritt sechs beinhaltet nun die Darstellung der Daten und die Entscheidung für eine Variante. Hier wurden die Daten in übersichtlichen Diagrammen visualisiert, aus denen „Services“ als beste Lösung hervorgeht.

Daraufhin wurde die Website dahingehend verändert, dass der „Interact“-Button gegen den „Services“-Button ausgetauscht wurde.

Zum Abschluss des Papers gibt es neben der Zusammenfassung der Ergebnisse einen Ausblick, wofür die Universitätsbibliothek A/B-Testing noch verwendet werden kann, beispielsweise die Gestaltung des Newsletters oder die Wirkung von Benachrichtigungen in der App. Es wird deutlich gemacht, dass A/B-Testing als Ergänzung zu Usability-Testing und anderen Methoden sinnvoll ist, nicht aber als einzige Methode benutzt werden sollte, da A/B-Testing nur quantitative Daten liefert, keine qualitativen.

Stärken und Schwächen: Zu Beginn des Papers wird im Abschnitt „Literature Review“ ein Einblick in den aktuellen Stand der Literatur gegeben. Das Paper beschreibt ausführlich, worauf es beim A/B-Testing ankommt. Dabei wird einerseits die Wichtigkeit der Arbeitshypothese herausgestellt, andererseits den User als wichtigen Teil der Entwicklung. Dieser kann an der Generierung der Variationen beteiligt werden und liefert durch das Testing die Userpräferenzen. Es wird sowohl jeder einzelne Schritt des A/B-Testings theoretisch als auch am Beispiel erläutert. Das Beispiel ist sehr detailliert und verständlich, auch die Anwendung von geeigneter Software (Google Analytics und Crazy Egg) wird dargelegt. Die Ergebnisse des Testings werden vom Autor selbst kritisch betrachtet und die recht kurze Dauer des Testings (3 Wochen) als Faktor genannt, der die Ergebnisse beeinflusst. Unter längeren Testbedingungen hätte die Variante „Services“ wahrscheinlich nicht die drop-off rate und homepage-return rate von 0% behalten. Als Schwäche des Artikels muss genannt werden, dass in den Diagrammen der drop-off rate und homepage-return rate die Variante „Help“ nicht erscheint. In der Diskussion dazu werden dafür keine Gründe genannt, der Grund dafür findet sich erst in einer Anmerkung am Schluss des Papers. Ursache dafür ist eine Berichtsbeschränkung durch das Programm Google Analytics, die zur Unvollständigkeit geführt hat. Es finden sich in dem Paper viele aktuelle Quellen, die den Eindruck der hohen Wissenschaftlichkeit dieses Papers bestärken.

Erstellt von: Meret Vollmann 347528