

**Untersuchung der Wirksamkeit unterschiedlicher
A/B-Test-Methoden zur Optimierung der User
Experience in Onlineshops: Fixed-Horizon-Tests vs.
sequential A/B-Tests**

Bachelorarbeit

im Studiengang

Mobile Medien

vorgelegt von

Isabella Saibert

Matr.-Nr.: 38642

am 31.08.2023

an der Hochschule der Medien Stuttgart

Erstprüfer/in: Prof. Dr. Gottfried Zimmermann

Zweitprüfer/in: Melina Hess

Ehrenwörtliche Erklärung

„Hiermit versichere ich, Isabella Saibert, ehrenwörtlich, dass ich die vorliegende Bachelorarbeit mit dem Titel: „Untersuchung der Wirksamkeit unterschiedlicher A/B-Test-Methoden zur Optimierung der User Experience in Onlineshops: Fixed-Horizon Tests vs. sequenzielle A/B Tests“ selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden.

Ich habe die Bedeutung der ehrenwörtlichen Versicherung und die prüfungsrechtlichen Folgen (§ 26 Abs. 2 Bachelor-SPO (6 Semester), § 24 Abs. 2 Bachelor-SPO (7 Semester), § 23 Abs. 2 Master-SPO (3 Semester) bzw. § 19 Abs. 2 Master-SPO (4 Semester und berufsbegleitend) der HdM) einer unrichtigen oder unvollständigen ehrenwörtlichen Versicherung zur Kenntnis genommen.“

Stuttgart, 31.08.2023

Unterschrift _____

Kurzfassung

Die vorliegende Arbeit befasst sich mit dem Thema A/B Testing, insbesondere im Kontext der Optimierung der User Experience in Onlineshops. In einer quantitativen empirischen Untersuchung, die in Zusammenarbeit mit der DRIP AGENCY und SNOCKS durchgeführt wurde, wurden zwei A/B-Test-Methoden – Fixed-Horizon Tests und sequentielle Tests – anhand von drei spezifischen Testideen in realen Onlineshop-Szenarien angewendet und auf ihre Effektivität verglichen sowie bewertet. Die Ergebnisse zeigen signifikante Unterschiede in der Performance und liefern daraus resultierende Empfehlungen für Onlineshops. Diese Erkenntnisse bieten Unternehmen, die mit den beiden A/B-Test-Methoden ihre UX optimieren und dadurch ihre Conversion Rates steigern möchten, wertvolle Einblicke. Dabei wird auch die Barrierefreiheit als ein zentrales und immer relevanter werdendes Thema hervorgehoben, um sicherzustellen, dass alle Nutzer eine zugängliche und positive Shopping-Erfahrung erleben können.

Abstract

The thesis deals with the topic of A/B Testing, especially in the context of optimizing the User Experience in online shops. In a quantitative empirical study, conducted in collaboration with DRIP AGENCY and SNOCKS, two A/B testing methods - Fixed-Horizon Tests and sequential tests - were applied to real online shop scenarios using three specific test ideas and compared and evaluated for their effectiveness. The results show significant differences in performance and provide resulting recommendations for online shops. These insights offer valuable perspectives to companies looking to optimize their UX with the two A/B testing methods and thereby increase their conversion rates. Additionally, accessibility is highlighted as a central and increasingly relevant topic to ensure that all users can experience an accessible and positive shopping experience.

Inhaltsverzeichnis

Ehrenwörtliche Erklärung	II
Kurzfassung	III
Inhaltsverzeichnis	IV
Abbildungsverzeichnis	VI
Abkürzungsverzeichnis	VIII
1 Einleitung, Zielsetzung & Aufbau dieser Arbeit	1
2 Methodik	2
3 Theoretische Grundlagen	6
3.1 Grundlegende Begriffe	6
3.1.1 A/B-Testing	6
3.1.2 KPIs	9
3.1.3 Monitoring und Analysetools.....	12
3.1.4 Statistik Basiswissen.....	14
3.1.5 User Experience	18
3.2 Einführung in die A/B Tests	19
3.2.1 A/B Testing Prozess: Vorgehensweise, Ziele, Vorteile und Methoden.....	19
3.2.2 Fixed-Horizon Tests vs. sequential A/B-Tests: Unterschiede und Anwendungsbereiche	23
3.2.3 Bedeutung und Positionierung von A/B-Tests im E-Commerce.....	27
3.3 A/B-Teststatistik	30
3.3.1 Notwendigkeit zur Aufstellung systematischer Hypothesen und Auswahl der	
Testvariable	30
3.3.2 Festlegung des Stichprobenumfangs und der Testdauer	31
3.4 User Experience im E-Commerce.....	34
3.4.1 Bedeutung und Einflussfaktoren auf die UX in Onlineshops.....	34
3.4.2 Barrierefreiheit im E-Commerce	37
4 Empirische Untersuchung	44
4.1 Research - Identifizierung von Optimierungspotenzialen im Shop.....	46
4.2 UI/ UX Designumsetzung und Prototyping der priorisierten Testideen.....	51
4.3 Durchführung & Auswertung der Testidee 1: Feature „Hover-Image“ zum Vergleich der	
Produkte“	59
4.4 Durchführung & Auswertung der Testidee 2 „Platzierung einer Announcement Bar (Banner)	
auf der Landing Page“	65

4.5	Durchführung & Auswertung der Testidee 3 „Bedienungshilfe: Text- und Anzeigeeinstellungen“.....	70
5	Interpretation der Testergebnisse/ Handlungsempfehlungen und abschließendes Fazit	75
6	Anhang	82
7	Literaturverzeichnis	85

Abbildungsverzeichnis

Abbildung 1: Beispiel eines A/B/n-Tests	7
Abbildung 2: Beispiel eines Multivariatentests	9
Abbildung 3: Berechnung der Conversion Rate	10
Abbildung 4: Berechnung der Click-Through-Rate	10
Abbildung 5: Berechnung der Average Revenue per User	11
Abbildung 6: Berechnung der Average Order Value	11
Abbildung 7: Die Usability ist ein Teil der User Experience	18
Abbildung 8: Beispiel eines A/B-Testing-Prozesses	23
Abbildung 9: Umsatz durch E-Commerce (B2C) in Deutschland in den Jahren 1999 bis 2022 sowie eine Prognose für 2023 (in Milliarden Euro)	28
Abbildung 10: Stichprobenrechner mit Beispielwerten zur Berechnung der Stichprobengröße	33
Abbildung 11: Stichprobenrechner mit Beispielwerten zur Berechnung der Testdauer	33
Abbildung 12: Konzept des User Experience Honeycombs von Peter Morville	36
Abbildung 13: Barrierefreiheit im Internet kaum vorhanden	38
Abbildung 14: Webseiten mit den häufigsten WCAG-Fehlern 2023 (% der Webseiten)	38
Abbildung 15: Vergleich 2019-2023 - Webseiten mit den häufigsten WCAG-Fehlern 2023 (% der Webseiten)	39
Abbildung 16: Subsysteme des menschlichen Organismus in der HCI	41
Abbildung 17: Feature "Hover-Image" zum Vergleich der SNOCKS Produkte	52
Abbildung 18: Platzierung des Newsletter Banners auf der Startseite (Desktop Version)	53
Abbildung 19: Platzierung des SNOCKS App Banners auf der Startseite (Desktop Version)	54
Abbildung 20: Platzierung des Newsletter Banners auf der Startseite (Mobile Version)	54
Abbildung 21: Platzierung des SNOCKS App Banners auf der Startseite (Mobile Version)	54
Abbildung 22: Smoke Test Bedienungshilfe: Text- und Anzeigeeinstellungen - Icon	55
Abbildung 23: Smoke Test Bedienungshilfe: Text- und Anzeigeeinstellungen - Icon aufgeklappt Light Version	55
Abbildung 24: Smoke Test Bedienungshilfe: Text- und Anzeigeeinstellungen - Icon aufgeklappt Dark Version	56
Abbildung 25: Hell-Dunkel Kontrastmodus mit Textgröße S	57
Abbildung 26: Hell-Dunkel Kontrastmodus mit Textgröße M	57
Abbildung 27: Hell-Dunkel Kontrastmodus mit Textgröße L	58
Abbildung 28: Hell-Dunkel Kontrastmodus mit Textgröße XL	58
Abbildung 29: Sequential Testing - Testidee 1: Allgemeine Informationen	59
Abbildung 30: Sequential Testing - Testidee 1: Statistische Parameter	60
Abbildung 31: Sequential Testing - Testidee 1: Testplanung	61
Abbildung 32: Sequential Testing - Testidee 1: Darstellung des Testverlaufs	61
Abbildung 33: Sequential Testing - Testidee 1: Finale Testauswertung	63
Abbildung 34: Fixed-Horizon-Test - Testidee 1: Testplanung	64
Abbildung 35: Fixed-Horizon-Test - Testidee 1: Finale Testauswertung	65
Abbildung 36: Sequential Testing - Testidee 2: Möglicher Testplan zur Erreichung des MDE	66
Abbildung 37: Sequential Testing - Testidee 2: Testplanung	66
Abbildung 38: Sequential Testing - Testidee 2: Darstellung des Testverlaufs	66
Abbildung 39: Sequential Testing - Testidee 2: Finale Testauswertung	67

Abbildung 40: Fixed-Horizon-Test - Testidee 2: Testplanung	68
Abbildung 41: Fixed-Horizon-Test - Testidee 2: Testplanung zur Erreichung des MDE.....	68
Abbildung 42: Fixed-Horizon-Test - Testidee 2: Finale Testauswertung.....	69
Abbildung 43: Sequential Testing - Testidee 3: Testplanung	70
Abbildung 44: Sequential Testing - Testidee 3: Dartellung des Testverlaufs.....	71
Abbildung 45: Sequential Testing - Testidee 3: Finale Testauswertung.....	72
Abbildung 46: Fixed-Horizon-Test - Testidee 3: Testplanung	73
Abbildung 47: Fixed-Horizon-Test - Testidee 3: Finale Testauswertung.....	74
Abbildung 48: Heatmap-Analyse der Announcement Bar (Newsletter & SNOCKS App) für die Desktop-Version	77
Abbildung 49: Heatmap-Analyse der Announcement Bar (Newsletter & SNOCKS App) für die Mobile Version	77
Abbildung 50: Klickhäufigkeiten auf die unterschiedlichen Text- und Anzeigeneinstellungen (Testidee 3).....	79
Abbildung 51: Dunkel-Hell Kontrastmodus mit Textgröße S.....	82
Abbildung 52: Dunkel-Hell Kontrastmodus mit Textgröße M.....	82
Abbildung 53: Dunkel-Hell Kontrastmodus mit Textgröße L.....	82
Abbildung 54: Dunkel-Hell Kontrastmodus mit Textgröße XL	83
Abbildung 55: Farben Kontrastmodus mit Textgröße S	83
Abbildung 56: Farben Kontrastmodus mit Textgröße M.....	83
Abbildung 57: Farben Kontrastmodus mit Textgröße L.....	84
Abbildung 58: Farben Kontrastmodus mit Textgröße XL	84

Abkürzungsverzeichnis

Abb.	Abbildung
ARPU	Average Revenue Per Order
APU	Average per User
AOV	Average Order Value
bspw.	beispielsweise
CR	Conversion Rate
CTA	Call-To-Action
CTR	Click-Through-Rate
KPI	Key Performance Indicators
MDE	Minimum Detectable Effect
RPU	Revenue Per User
STDEV	Standardabweichung
UN	Unternehmen
UX	User Experience

1 Einleitung, Zielsetzung & Aufbau dieser Arbeit

Onlineshops haben die Art und Weise revolutioniert, wie Menschen weltweit einkaufen. Mit dem exponentiellen Wachstum des E-Commerce, wo nahezu jedes Unternehmen über eine eigene Webpräsenz verfügt, rückt insbesondere die kontinuierliche Optimierung und Weiterentwicklung dieser bestehenden Websites in den Mittelpunkt. Es geht darum, den Nutzern stets die bestmögliche Erfahrung zu bieten und sich den wandelnden Anforderungen und Trends anzupassen. (Wenz & Hauser, 2015, VII)

Daher sind sogenannte A/B-Tests zu einem zentralen Instrument geworden, um diese Optimierungen datenbasiert und effektiv durchzuführen, indem sie direkte Vergleiche zwischen verschiedenen Versionen einer Webseite oder eines Features ermöglichen. Dabei gibt es verschiedene Ansätze, wie den Fixed-Horizon sowie sequenziellen A/B Test. Die Wahl der richtigen A/B-Testmethode kann den Unterschied zwischen einem erfolgreichen und einem fehlgeschlagenen Test ausmachen. Wie bei jeder Methode haben sowohl Fixed-Horizon als auch sequenzielle A/B Tests ihre eigenen Vor- und Nachteile. Doch welche dieser Ansätze erweist sich im speziellen Umfeld von Onlineshops als tatsächlich effizienter?

Die DRIP Agency hat die entscheidende Bedeutung dieser Frage erkannt. Die Kunden der Agentur setzen auf deren Expertise, um die besten Ratschläge und Strategien für ihre E-Commerce-Plattformen zu erhalten. Es ist von großer Wichtigkeit, dass die Empfehlungen der DRIP Agency auf soliden Daten und gründlicher Forschung basieren.

Deshalb zielt diese Bachelorarbeit darauf ab, die Wirksamkeit von sequenziellen A/B Tests im Vergleich zu Fixed-Horizon Tests eingehend zu prüfen. Durch diese Untersuchung, die in Zusammenarbeit mit dem Unternehmen SNOCKS erfolgt, soll ermittelt werden, welche Methode am besten für die Optimierung der User Experience ist und zu einer Umsatzsteigerung im Onlineshop führt. Dies wird nicht nur dazu beitragen, die DRIP Agency als führende Experten in ihrem Bereich zu etablieren, sondern auch den Kunden zu vermitteln welche Ansätze am effektivsten sind, um ihre Online-Präsenz zu stärken.

2 Methodik

Das Kapitel "Methodik" widmet sich der systematischen Vorgehensweise der beiden prominenten A/B-Test-Methoden: Fixed-Horizon Tests und sequenzielle A/B-Tests.

Im Rahmen der Untersuchung der Wirksamkeit der beiden A/B-Test-Methoden zur Optimierung der User Experience in Onlineshops wurde daher zuerst eine umfangreiche Literaturrecherche als grundlegende wissenschaftliche Methode angewendet. Diese Recherche diente dazu, bestehende Literatur und Quellen zu diesem spezifischen Thema zu sammeln und zu analysieren.

Die wissenschaftliche Methode setzt zunächst das Verständnis der Grundlagen voraus und zielt darauf ab, dem Leser die Basisbegriffe zum Thema dieser Arbeit fundiert zu vermitteln. Dabei spielen der Begriff A/B Testing, die Auswahl passender KPIs, der Einsatz von Monitoring-Tools, ein fundiertes Verständnis der Statistik sowie der Begriff User Experience eine entscheidende Rolle, die in den Kapiteln 3.1.1-3.1.5 näher erläutert werden. Im Folgenden wird das Kapitel "3.2 Einführung in die A/B Tests" vertieft, welches den Prozess vorstellt, die es ermöglicht, zwei Varianten einer Webseite oder Anwendung direkt miteinander zu vergleichen, um die effektivere Option zu ermitteln. Es wird dargestellt, wie A/B-Tests durchgeführt werden, um den Lesern ein klares Verständnis der einzelnen Schritte zu vermitteln. Der Prozess beginnt mit der Zielsetzung der KPIs und der Formulierung einer Hypothese. Von dort aus werden die verschiedenen Aspekte der Testplanung, der Entwicklung von Testvarianten, der Durchführung des Tests bis hin zur Analyse und Monitoring der Ergebnisse in der Theorie behandelt. Das Kapitel betont auch die Ziele des A/B-Testings, insbesondere wie solche Tests dazu beitragen können, die User Experience sowie die KPI's CR und ARPU zu verbessern. Zudem werden die beiden Ansätze Fixed-Horizon Test und sequenzielle A/B-Tests vorgestellt und hinsichtlich ihrer Vor- und Nachteile näher betrachtet.

Nach dem Kapitel ist der Leser in der Lage, die Grundlagen des A/B-Testing-Prozesses zu verstehen und zwischen den Ansätzen des Fixed-Horizon-Tests und den sequenziellen A/B-Tests zu unterscheiden.

Um auch die Bedeutung der A/B-Tests im E-Commerce hervorzuheben, werden im darauffolgenden Kapitel auf die Positionierung und Vorteile der A/B-Tests im E-Commerce eingegangen.

Bereits in Kapitel 3.1.4 wurden grundlegende Begriffe der Statistik vorgestellt, während Kapitel 3.2 eine kurze Einführung in die Analyse und Interpretation von A/B-Testing-Ergebnissen lieferte. Einige spezifische „Schritte“, die zum A/B-Testing gehören, wie die Notwendigkeit zur

Aufstellung von Hypothesen, die Bestimmung der Stichprobengröße und die teilweise Festlegung der Testdauer (je nach Ansatz), werden daher in Kapitel 3.3.1 detaillierter erläutert. Der Grund für die Ausarbeitung eines gesonderten Kapitels hierzu liegt in der Komplexität und Bedeutung der A/B Teststatistik. Ein tieferes Verständnis dieser Aspekte ermöglicht es, Tests effizienter zu gestalten, Fehlerquellen zu minimieren und letztlich aussagekräftigere Ergebnisse zu erzielen.

Der Theorieteil findet seinen Abschluss in Kapitel 3.4, welches sich intensiv mit der User Experience im E-Commerce auseinandersetzt. Hierbei werden die Facetten der User Experience erläutert und die Bedeutung der User Experience für den Onlineshop beleuchtet. Ein besonderer Schwerpunkt liegt auf dem Thema Barrierefreiheit. In diesem Kontext wird diskutiert, wie Online-Shops gestaltet werden müssen, um für alle Nutzer, unabhängig von ihren physischen oder kognitiven Fähigkeiten, zugänglich und nutzbar zu sein. Dies umfasst sowohl technische Aspekte als auch Designprinzipien, die eine inklusive und positive Erfahrung für alle Besucher gewährleisten. Das Kapitel betont die Notwendigkeit, die User Experience ständig zu optimieren, da sie direkt mit der Kundenzufriedenheit und letztlich auch mit dem wirtschaftlichen Erfolg eines Online-Shops verknüpft ist.

Durch diese systematische Herangehensweise war es möglich, die Arbeit auf einem soliden theoretischen Fundament aufzubauen und dem Leser ein grundlegendes Verständnis für das Thema zu vermitteln. Dieser Schritt ist wichtig, um die nachfolgenden praktischen Untersuchungen und Analysen im Kontext der bestehenden Literatur durchzuführen

Das 4. Kapitel, welches sich der empirischen Untersuchung widmet, verfolgt in diesem Abschnitt einen Ansatz basierend auf einer Meta-Methodik. Dabei handelt es sich um eine statistische Methode, die dazu dient, die unterschiedlichen A/B-Test-Methoden auf ein angewendetes Praxisbeispiel quantitativ zu analysieren und zu bewerten. Dieser Ansatz ermöglicht eine tiefgreifende Auseinandersetzung mit den jeweiligen Methoden und deren Anwendbarkeit im Kontext der Optimierung der User Experience in Onlineshops. Ziel dabei ist es, die Ergebnisse einzelner quantitativer Testideen zusammenzufassen, diese zu interpretieren und auf dieser Basis eine fundierte Bewertung darüber abzugeben, ob der Fixed-Horizon- oder der sequenzielle Ansatz effektiver ist.

In den beiden Kapiteln 4.1 und 4.2 wurde speziell noch ein weiterer Ansatz verfolgt, der bestimmte Elemente des Design Thinkings integriert und reflektiert. Design Thinking ist dabei eine Methode, die darauf ausgerichtet ist die Bedürfnisse der Nutzer zu verstehen und Lösungen zu entwickeln, die sowohl technisch umsetzbar als auch wirtschaftlich rentabel ist. Im Kontext des Design Thinkings gibt es einen Prozess, der sich in fünf Phasen gliedert: Empathize (Empathiebildung), Define (Problemdefinition), Ideate (Ideenentwicklung), Prototype

(Prototypenerstellung) und Test (Evaluierung). Im Design Thinking-Prozess beginnt die Phase "Empathize" mit der Auseinandersetzung und dem Verständnis der Zielgruppe. In der darauf folgenden "Define"-Phase werden durch die gesammelten Erkenntnisse zentrale Problemstellungen identifiziert. In "Ideate" werden durch kreatives Brainstorming potenzielle Lösungsansätze generiert. Die "Prototype"-Phase dient der konkreten Ausarbeitung und Evaluierung dieser Lösungen, während in der abschließenden "Test"-Phase durch Benutzerfeedback die Adäquatheit des entwickelten Prototyps überprüft wird. (HPI ACADEMY) Vor allem die Phasen Define, Ideate und Prototype standen dabei im Fokus dieser Arbeit, da sie den Kern der Designumsetzung im SNOCKS Onlineshop repräsentieren.

Nach genauerer Betrachtung bedeutet dies, dass im SNOCKS Onlineshop anfänglich potenzielle Optimierungsbereiche identifiziert werden müssen, bevor der A/B-Testing Prozess starten kann, was der "Define"-Phase entspricht. Dazu zählen beispielsweise die Optimierung von Suchalgorithmen oder des Checkout-Prozesses, die Implementierung einer hochauflösenden Zoom-Funktion für Produktvisualisierungen oder die Etablierung von Anreizsystemen, um Kunden zur Newsletter-Registrierung zu motivieren. Darauf aufbauend fokussierte die "Ideate"-Phase auf kreatives Brainstorming, um vielfältige Lösungen für die identifizierten Herausforderungen zu entwickeln. Durch die Identifizierung dieser verschiedenen Bereiche können A/B Tests durchgeführt werden, um damit die User Experience zu verbessern, die CR & ARPU zu steigern oder andere relevante KPIs positiv zu beeinflussen. Das Ziel ist es, den Onlineshop effizienter, benutzerfreundlicher und letztlich erfolgreicher zu gestalten. Hinsichtlich dieser Arbeit wurden im SNOCKS Onlineshop 5 Optimierungspotenziale samt zugehörigen Hypothesen identifiziert, die in Kapitel 4.1 näher erläutert werden. Unter Berücksichtigung des Zeitaufwands und des prognostizierten Mehrwerts wurden letztendlich drei Hypothesen bzw. Testideen aus den erkannten Optimierungspotenzialen bevorzugt behandelt.

Folgende Testideen wurden dabei priorisiert:

- Feature „Hover-Image“: ermöglicht es den Nutzern, durch einfaches Überfahren des Produktbildes mit dem Mauszeiger einen direkten und intuitiven Vergleich der Produktvarianten zu erhalten.
- Platzierung einer Announcement Bar (Banner) auf der Landing Page: dadurch wird die Aufmerksamkeit gezielt auf bestimmte Aktionen, Rabatte oder Neuigkeiten gelenkt und kann somit die Interaktion und Konversion auf der Webseite fördern.
- Bedienungshilfe: Text- und Anzeigeeinstellungen: ermöglicht es den Nutzern, die Darstellung von Schriftgrößen und Kontrastmodi anzupassen, um eine optimale Lesbarkeit und Benutzerfreundlichkeit zu gewährleisten.

Im darauffolgenden Schritt wurden diese 3 Testideen designkonzeptionell in Form von digitalen Mockups umgesetzt, die mittels des Designtools Figma als Prototypen ausgearbeitet wurden (Prototype-Phase). Durch iterative Feedback-Zyklen, die in enger Kooperation zwischen der Agentur DRIP und dem Onlineshop SNOCKS durchgeführt wurden, konnte das Design verfeinert und optimiert werden. Detaillierte Ausführungen zu den priorisierten Testideen finden sich im Abschnitt 4.2.

Im Anschluss daran erfolgte die Umsetzung und Analyse der priorisierten Testkonzepte unter Anwendung der untersuchenden Testmethoden: Fixed-Horizon und sequenzielles A/B-Testing. Die Tests wurden quantitativ durchgeführt, wobei jede Testidee mit beiden Ansätzen durchgeführt wurden. Abhängig vom gewählten Ansatz wurden die Daten über einen definierten Zeitraum erfasst, um letztendlich den erhofften signifikant positiven Effekt festzustellen. Die Resultate der durchgeführten Tests wurden anschließend ausgewertet, um fundierte Entscheidungen über die Implementierung der Testideen im Onlineshop zu treffen und zu erörtern, welche der beiden Ansätze Fixed-Horizon oder sequential effektiver ist.

Nach Durchführung der Tests schließt die Arbeit mit der Interpretation der Testergebnisse, sowie mögliche Handlungsempfehlungen und einem abschließenden Fazit, das die gewonnenen Erkenntnisse zusammenfasst, ab. Diese sind in Kapitel 5 detailliert dargestellt.

3 Theoretische Grundlagen

3.1 Grundlegende Begriffe

Bevor man sich mit den spezifischen Ansätzen und Methoden des A/B-Testens beschäftigt, ist es wichtig, ein solides Grundverständnis der zugrundeliegenden Konzepte zu erlangen. Infolgedessen widmet sich dieses Kapitel der eingehenden Untersuchung und Analyse zentraler Konzepte, die für das Verständnis und die Anwendung von A/B-Testing unerlässlich sind. Dies umfasst die Identifikation und Interpretation relevanter Leistungsindikatoren (KPIs), die Nutzung von Monitoring- und Analysetools wie Google Analytics & Analytics Toolkit, die Vermittlung grundlegender statistischer Kenntnisse und die Erörterung der Rolle der User Experience.

3.1.1 A/B-Testing

Der Begriff „A/B Testing“ findet vorrangig Anwendung im E-Commerce-Sektor, insbesondere im Bereich des Online-Marketings. Dabei gibt es verschiedene Begrifflichkeiten des „A/B Testings“, die je nach Kontext oder Fachgebiet verwendet werden. Dazu gehören Begriffe wie „A/B-Test, A/B Testing, Split-Testing, Webseiten-Testing oder (Online-)Testing“, wobei alle aus dem gleichen Bedeutungsfeld stammen. (Witzenleiter, 2021, S. VI) Unter A/B Testing versteht man somit eine Methode, bei der zwei Varianten oder Versionen eines Elements verglichen werden, um festzustellen, welche Variante besser performt bzw effektiver ist. (Birkett, 2022) Oft werden auch die Begriffe "Experiment" oder "experimentieren" in Verbindung gebracht, da sie einen Ansatz darstellen, bei dem spezifische Änderungen an einem oder mehreren Elementen vorgenommen werden, um zu sehen, wie sich diese auf das Verhalten oder die Reaktion der Benutzer oder Kunden auswirken. Im Bereich des Marketings wird die Zielgruppe eines UN nach dem Zufallsprinzip in zwei Gruppen aufgeteilt, wobei die erste Gruppe die Originalversion, auch Kontrollversion genannt, sieht, während die zweite Gruppe eine abgeänderte Version erhält. (FUSEON) A/B-Testing kann somit als Experiment betrachtet werden, da die Auswirkungen der Änderungen auf das Verhalten der Benutzer gemessen werden können.

Im Bereich des A/B-Testings gibt es zudem verschiedene Testmethoden, die je nach Anforderungen und Zielen unterschiedlich eingesetzt werden. Sie sind grundlegend gleich aufgebaut, unterscheiden sich aber beispielweise in Bezug auf die Dauer oder die Anzahl der Versionen. Im nachfolgenden Abschnitt werden zwei weitere A/B-Testing-Methoden erläutert.

A/B/n Testing ist eine Form des A/B Testings, bei dem mehr als zwei Varianten (A und B) gegeneinander getestet werden. Das "n" steht hierbei für eine beliebige Anzahl von Varianten, die zusätzlich eingeführt werden und mit der Kontrollversion (A) verglichen werden, um die Wirksamkeit verschiedener Design- oder Funktionsänderungen zu untersuchen (siehe Abb. 1). (Witzenleiter, 2021, S. 3) Vorteil dieser Methode ist der gleichzeitige Vergleich mehrerer Varianten. So wird jede Variante an eine separate Gruppe von Benutzern ausgeliefert, um zu ermitteln, welche Version die besten Ergebnisse erzielt. Zusätzlich werden nicht nur die effektivsten Varianten herausgefiltert, sondern auch diejenigen identifiziert, die am wenigsten performen. (Witzenleiter, 2021, S. 4) Dies unterstützt Unternehmen dabei, auf Basis der Tests datengesteuerte Entscheidungen zu treffen, kontinuierliche Optimierungen durchzuführen und gleichzeitig negative Auswirkungen zu vermeiden. Obwohl A/B/n-Tests einige Vorteile bietet, gibt es auch potenzielle Nachteile, die zu betrachten sind. Die Durchführung von A/B/n-Tests, bei denen mehrere Varianten getestet werden, erfordern normalerweise einen höheren Zeitaufwand und mehr Ressourcen im Vergleich zu herkömmlichen A/B-Tests. Mit einer steigenden Anzahl von Varianten erhöht sich auch die Komplexität der Analyse. Es erfordert eine ausreichende Datenerhebung und eine größere Stichprobengröße, um statistisch signifikante Ergebnisse zu erzielen. (Optimizely)

„Daher sollte man abwägen, ob man stattdessen nicht mehrere A/B-Tests als Alternative dazu durchführt, bei denen dann immer die Gewinnervariante gegen eine neue Variante antritt.“ (Witzenleiter, 2021, S. 3)

Dies ermöglicht eine schrittweise Verbesserung der Versionen, da jedes Teststadium auf den vorherigen Ergebnissen aufbaut.

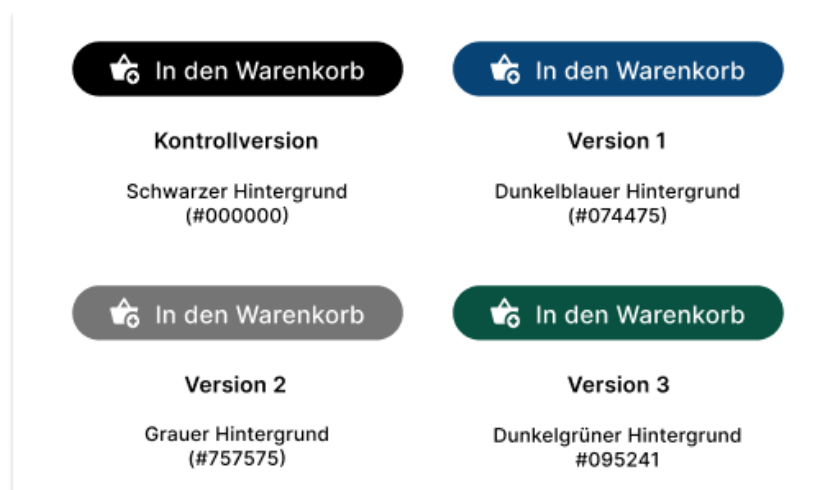


Abbildung 1: Beispiel eines A/B/n-Tests

(Vgl. Witzenleiter, 2021, S. 4)

Multivariates Testing

Im Gegensatz zum klassischen A/B-Testing ermöglicht das Multivariate Testing die gleichzeitige Testung mehrerer Variablen, um herauszufinden, welche Kombination von Variablen das beste Ergebnis liefert. (Witzenleiter, 2021, S. 4) Das heißt, es werden verschiedene Elemente betrachtet und in Varianten aufgeteilt. Elemente sind beispielsweise Überschriften, Call-to-Action-Buttons, verschiedene Farbschemata oder die Auswahl von Bildern. Durch die Kombination dieser Varianten entstehen verschiedene Versionen der Seite, die gleichzeitig getestet werden. Die Abb. 2 verdeutlicht den Prozess des Multivariates Testing anhand eines Beispiels. „Es werden zwei Textvarianten des Call-to-Action-Buttons mit drei Farbvariationen getestet, wodurch sich im Endeffekt sechs mögliche Kombinationen ergeben.“ (Vgl. Witzenleiter, 2021, S. 4)

Über eine Formel, lässt sich die Menge der möglichen Testvariationen berechnen:

„Anzahl Varianten Element A X Anzahl Varianten Element B X Anzahl Varianten Element C ... = Gesamtanzahl der Varianten“ (Witzenleiter, 2021, S. 5)

Multivariate A/B-Tests bieten zwei Vorteile aufgrund der Vielzahl an Variablenkombinationen, die getestet werden können. Zum einen erlauben Multivariate Tests Zusammenhänge und Abhängigkeiten zwischen den getesteten Variablen leichter festzustellen. (Witzenleiter, 2021, S. 5) Das heißt, es können Effekte identifiziert werden, die nur auftreten, wenn bestimmte Variablen gemeinsam variiert werden. Durch die Identifizierung ist es möglich, eine Analyse durchzuführen, um zu untersuchen, wie sich die Variablen gegenseitig beeinflussen und ob es eine optimale Kombination gibt, die zu den besten Ergebnissen führt. (ADVIDERA) Zum anderen, können Ressourcen effizienter genutzt werden, da die Tests in einem Experiment durchgeführt werden. Somit können schneller umfangreiche Informationen über verschiedene Variablenkombinationen gewonnen werden, die im besten Fall zu einem optimalen Ergebnis führen.

Aufgrund verschiedener Faktoren, Ressourcen oder Zielen ergeben sich beim Multivariaten Testing auch Nachteile, die in Betracht gezogen werden müssen.

Multivariate A/B Tests sind oft komplexer im Gegensatz zu den klassischen A/B-Tests, da es sich schwieriger gestaltet, die Auswirkungen jeder einzelnen Variablenkombinationen zu verstehen und zu interpretieren. (ADVIDERA) Es erfordert eine sorgfältige Planung und statistische Techniken sowie Analysen, um die Ergebnisse korrekt zu deuten.

„Zudem benötigt man zur Durchführung von Multivariaten Tests in der Regel sehr große Besuchermengen.“ (Vgl. Witzenleiter, 2021, S. 5) Für jede Kombination müssen genügend Daten gesammelt werden, um signifikante Ergebnisse zu erhalten. Daher besteht die Möglichkeit, dass der Test länger andauert, was wiederum zu höheren Kosten führt.

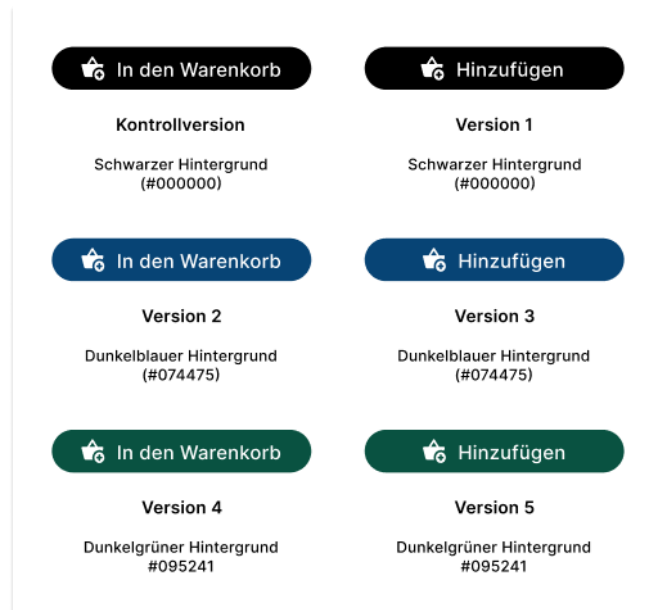


Abbildung 2: Beispiel eines Multivariatentests
 (Vgl. Witzel, 2021, S. 5)

Letztendlich hängt die Entscheidung für die geeignete Methode von einer Vielzahl von Faktoren ab. Je nach Ziel, Ressourcen oder Testumfang muss sorgfältig abgewogen werden, welche Methode am besten geeignet ist.

Das grundlegende Ziel eines A/B-Testings ist jedoch die bestmögliche Variante eines Elements oder einer Webseite zu finden und diese über einen mehrstufigen Testansatz und mehrere Testzyklen so zu optimieren, dass am Ende eine bestmögliche Version entsteht.

3.1.2 KPIs

Aus einer wirtschaftlichen und unternehmerischen Perspektive besteht das Ziel des A/B-Testings darin, die Key Performance Indicators (KPIs) oder auch als „Schlüsselkennzahlen“ bezeichnet, zu optimieren. Für eine detailliertere Betrachtung der unterschiedlichen Kennzahlen des A/B Testings ist es erforderlich, zunächst eine allgemeine Definition des Begriffs vorzunehmen. KPIs werden laut Gabler „in der Betriebswirtschaftslehre als allgemeine Kennzahlen bezeichnet, die sich auf den Erfolg, die Leistung oder Auslastung des Betriebs, seiner einzelnen organisatorischen Einheiten oder einer Maschine beziehen.“ (Springer Gabler) Die KPIs dienen somit als Messgrößen, um die gesetzten Ziele eines Unternehmens zu messen, die Effektivität von Marketingmaßnahmen zu beurteilen und Entscheidungen zu treffen, welche Marketingstrategien eingesetzt werden. KPIs werden im Kontext des A/B-Testing verwendet,

um die Leistung und Wirkung unterschiedlicher Testvarianten zu vergleichen und zu bewerten. Dadurch ermöglichen sie die Ermittlung der Auswirkungen von Änderungen in einem Experiment. Im Rahmen des A/B-Testings werden folgende relevante Kennzahlen näher erläutert, um eine detaillierte Bewertung durchzuführen.

Conversion Rate: Die CR ist eine wichtige Schlüsselkennzahl im E-Commerce Bereich, die dazu dient, den Erfolg von Online-Marketingaktivitäten zu messen und Verbesserungspotentiale zu identifizieren oder aufzuweisen. Sie misst das Verhältnis zwischen der Anzahl der Besucher, die eine bestimmte Aktion (Conversion) durchführen und der Gesamtanzahl der Besucher einer Webseite. (Lemodo) Die CR wird in der Regel als Prozentsatz ausgedrückt. Je höher der Prozentsatz liegt, desto erfolgreicher ist eine Webseite. Wenn beispielweise von 100 Besuchern einer Webseite zehn tatsächlich eine Bestellung aufgeben, beträgt die CR 10%. Sie bezieht sich somit darauf, wie viele Besuche letztendlich zu einer Bestellung führen. Die Abb. 3 zeigt wie man die CR berechnet.

$$= \frac{\text{Anzahl der Conversions einer Internetseite}}{\text{Gesamtanzahl der Nutzer/ Besucher}} \times 100\%$$

Abbildung 3: Berechnung der Conversion Rate

Im Rahmen des A/B-Testings wird die CR genutzt, um die Leistung sowie den Erfolg der verschiedenen Testvarianten zu vergleichen und schlussendlich zu ermitteln, welche Variante die gewünschten Aktionen am effektivsten fördert und dabei gleichzeitig eine höhere Konversionsrate erzielt.

Click-Through-Rate: Die Click-Through-Rate (CTR) ist eine weitere KPI im E-Commerce, die anzeigt, wie erfolgreich eine Anzeige oder ein Element die Aufmerksamkeit der Benutzer erregt und sie dazu motiviert, darauf zu klicken. Sie misst das Verhältnis der Anzahl der Klicks auf ein Werbemittel zur Gesamtzahl der angezeigten Impressionen. (Lemodo) Die Abb. 4 zeigt wie man die CTR berechnet. Die CTR wird in der Regel als Prozentsatz ausgedrückt. Eine höhere Klickrate deutet darauf hin, dass mehr Benutzer auf das Werbemittel reagiert und darauf geklickt haben..

$$= \frac{\text{Anzahl der Klicks}}{\text{Anzahl der Impressionen}} \times 100\%$$

Abbildung 4: Berechnung der Click-Through-Rate

Im Kontext zum A/B-Testing wird die CTR verwendet, um festzustellen welche Testvariante eine höhere Klickrate erzielt und somit attraktiver für die Benutzer ist. Eine höhere CTR kann auf eine bessere Gestaltung, eine ansprechendere Botschaft oder eine effektivere Platzierung hinweisen.

Average Revenue per User: Der Average Revenue per User (ARPU), dt. "durchschnittlicher Erlös pro Nutzer" wird verwendet, um den durchschnittlichen Umsatz pro Kunde zu berechnen. Um den ARPU zu ermitteln, ist es zunächst notwendig, eine spezifische Zeitperiode festzulegen. Der ARPU wird berechnet, indem der Gesamtumsatz während eines bestimmten Zeitraums durch die durchschnittliche Anzahl der Benutzer in diesem Zeitraum geteilt wird (siehe Abb. 5). (Ollmann, 2020)

$$= \frac{\text{Gesamtumsatz einer bestimmten Zeitdauer}}{\text{Gesamtanzahl der Nutzer einer bestimmten Zeitdauer}}$$

Abbildung 5: Berechnung der Average Revenue per User

In Bezug auf das A/B Testing handelt es sich um eine Kennzahl, die verwendet wird, um den Vergleich des durchschnittlichen Umsatzes pro Nutzer zwischen den Testvarianten zu ermöglichen.

Average Order Value: der Average Order Value (AOV), dt. "durchschnittlicher Bestellwert" ermittelt den durchschnittlichen Warenkorbwert der Kunden. Der AOV wird berechnet, indem der Gesamtumsatz durch die Anzahl der Bestellungen in einem bestimmten Zeitraum geteilt wird (siehe Abb. 5). (Lemodo) Ein höherer Average Order Value (AOV) zeigt an, dass Kunden entweder mehr Produkte kaufen oder sich für teurere Artikel entscheiden.

$$= \frac{\text{Gesamtumsatz einer bestimmten Zeitdauer}}{\text{Anzahl der Bestellungen einer bestimmten Zeitdauer}}$$

Abbildung 6: Berechnung der Average Order Value

Im A/B-Testing spielt der Average Order Value (AOV) eine bedeutende Rolle, um den monetären Wert der verschiedenen Testvarianten zu bewerten. Dadurch kann ermittelt werden, welche Variante einen höheren durchschnittlichen Bestellwert erzielt und somit wirtschaftlich erfolgreicher ist. Die Auswahl der relevanten KPIs ist von der Art des Tests und den spezifischen Zielen abhängig. Es ist daher von großer Bedeutung, KPIs auszuwählen, die mit den Zielen des UN übereinstimmen.

3.1.3 Monitoring und Analysetools

Die KPIs und das Monitoring stehen eng miteinander in Verbindung. „Monitoring, dt. „Dauerbeobachtung“ ist eine fortlaufende ständige Überwachung von Prozessen und Vorgängen.“ (ONLINEMARKETING.DE)

Das Monitoring basiert auf der Definition und Auswahl geeigneter KPIs und bildet eine wichtige Grundlage bei der Bewertung und Überwachung der Leistung von Unternehmen.

Im Marketing spielen die KPIs eine zentrale Rolle und dienen im Monitoring als Instrumente zur fortlaufenden Überwachung und Messung des Fortschritts. Durch die gezielte Beobachtung und Analyse der KPIs ermöglicht das Monitoring den Unternehmen ihre Marketingstrategien anzupassen, Verbesserungspotenziale aufzuweisen und fundierte Entscheidungen zu treffen. (Versa commerce, 2023b) Das Monitoring der KPIs bietet somit eine Grundlage für eine datenbasierte und erfolgsorientierte Herangehensweise im Marketingbereich.

Das Monitoring im A/B-Testing bezieht sich auf die kontinuierliche Überwachung und Bewertung der Testergebnisse, um den Fortschritt des Tests, die Leistung der verschiedenen Testvarianten und die erzielten Auswirkungen zu beobachten. (Boßow-Thies, Hofmann-Stöltig & Jochims, 2020, S. 167) Es dient dazu, mögliche Probleme frühzeitig zu erkennen und fundierte Entscheidungen für den weiteren Verlauf des Tests zu treffen. Das Monitoring im A/B-Testing ist somit von großer Bedeutung, um den Testprozess effizient zu steuern und optimale Ergebnisse zu erzielen.

Zur laufenden Kontrolle gibt es dafür verschiedene Tools und Softwarelösungen, die für das Monitoring im A/B-Testing eingesetzt werden können. In diesem Abschnitt werden lediglich die beiden praxisrelevanten Tools betrachtet, die für diese Bachelorarbeit von besonderer Relevanz sind, da es heutzutage eine umfangreiche Auswahl an verfügbaren Tools gibt.

Google Analytics: Google Analytics ist ein clientbasierter Tracking-Dienst von Google, der speziell für die Analyse von Websites entwickelt wurde. Durch die Nutzung von Google Analytics können Websitebetreiber wertvolle Daten über ihre Website-Besucher erfassen und analysieren. Durch das Tracking kann beispielsweise der Gesamtverkehr auf der Website verfolgt werden, einschließlich der Anzahl der Besucher, Seitenaufrufe, Besuchszeiten oder Absprungraten. (Ertel & Venzke-Caprarese, 2014, S. 182) Zudem bietet Google Analytics viele weitere Funktionen an, um das Tracking und die Analyse einer Website zu unterstützen. Die Funktionalität von Google Analytics basiert wesentlich auf der Verwendung von Cookies. Durch das clientbasierte Verfahren arbeitet der Tracking-Dienst mit Daten, die auf dem Endgerät des Nutzers gesammelt werden und von dort aus an den Tracking-Dienst gesendet werden. „Besucht also ein Nutzer eine Website, wird ein Cookie mit einer sogenannten Client-ID auf dem verwendeten Endgerät abgelegt. Hierdurch ist es möglich, das verwendete Endgerät

wiederzuerkennen.“ (Ertel & Venzke-Caprarese, 2014, S. 182–183) Bei jedem Zugriff auf die Webseite erfolgt eine Aktualisierung des Cookies, das spezifische Daten enthält. Diese Daten werden über eine HTML-Verbindung an Google Analytics gesendet und dort gespeichert. Der Webseiten-Betreiber hat später die Möglichkeit, die gespeicherten Daten in Google Analytics abzurufen und sie in Form von anschaulichen Grafiken aufzubereiten. (Steidle & Pordesch, 2008, S. 325) Bei der Erfassung von Nutzerdaten ist es zudem von großer Bedeutung, dass Websitebetreiber die Privatsphäre und den Datenschutz ihrer Besucher gewissenhaft berücksichtigen. Dies erfordert die Einhaltung von Richtlinien zur Datenerfassung und -nutzung, bspw. durch die Implementierung einer Datenschutzerklärung. (Ertel & Venzke-Caprarese, 2014, S. 183)

Analytics Toolkit: Analytics Toolkit ist im Gegensatz zu Google Analytics ein kosten-pflichtiges Tool, das speziell für das A/B-Testing und die allgemeine Webanalyse für Websites entwickelt wurde. Es stellt viele Tools & Funktionen zur Verfügung wie bspw. den statistischen Signifikanzrechner oder Stichprobengrößenrechner, um bei der Planung, Durchführung und Analyse von A/B-Tests zu helfen. Analytics Toolkit stellt darüber hinaus auch erweiterte Analyse- und Berichtsfunktionen zur Verfügung, um die Ergebnisse der durchgeführten Experimente zu überwachen und detailliert zu analysieren.

Im Vergleich zu Google Analytics liegt der Fokus von Analytics Toolkit somit speziell auf die Durchführung von A/B-Tests und Experimenten. Es bietet eine benutzerfreundliche Oberfläche, die es einfach macht, verschiedene Varianten zu erstellen und zu testen. (Analytics Toolkit) Hingegen konzentriert sich Google Analytics eher auf die Analyse von Website-Daten. Es bietet eine breite Palette von Analysefunktionen, mit denen Nutzer Einblicke in das Verhalten der Besucher, die Traffic-Quellen und andere wichtige Metriken erhalten können. Das Analytics Toolkit ist daher eine hervorragende Plattform, die es auch Nutzern ohne tiefere statistische Kenntnisse ermöglicht, Einblicke in ihre Daten zu gewinnen.

Die Auswahl eines geeigneten Tools hängt letztendlich von den spezifischen Anforderungen, dem Budget und den technischen Fähigkeiten eines Unternehmens ab. Aus diesem Grund ist es von großer Bedeutung, das Tool zu wählen, das am besten den Bedürfnissen und Zielen des Unternehmens entspricht.

3.1.4 Statistik Basiswissen

Im Rahmen des A/B-Testings spielt die statistische Analyse eine entscheidende Rolle, um auf Grundlage der erhobenen Daten objektive und fundierte Entscheidungen zu treffen. Ihr Ziel besteht darin sicherzustellen, dass die beobachteten Unterschiede zwischen den Varianten nicht dem Zufall geschuldet sind, sondern statistisch signifikant sind.

Um ein grundlegendes Verständnis zu erlangen, ist es daher von Bedeutung, die elementaren Begriffe der Statistik zu kennen.

Statistische Signifikanz: Die Statistische Signifikanz ist ein wichtiger Begriff in der Statistik, um zu bestimmen, ob ein beobachteter Unterschied oder Effekt real ist und nicht auf Zufall oder Stichprobenfehler zurückzuführen ist. Um es in anderen Worten zu verdeutlichen heißt es: „Eine Signifikanz von 90% bedeutet, dass bei einer Wiederholung des Experiments mit einer 90%igen Wahrscheinlichkeit mit demselben Ergebnis zu rechnen ist, das Resultat also nicht dem Zufall geschuldet ist.“ (Witzenleiter, 2021, S. 119)

In A/B-Tests bezieht sich die statistische Signifikanz auf die Aussagekraft der Ergebnisse und zeigt an, ob die beobachteten Unterschiede zwischen den Varianten A und B tatsächlich auf die unterschiedlichen Bedingungen oder Änderungen zurückzuführen sind oder ob sie rein zufällig auftreten könnten. (SurveyMonkey) Wenn ein statistisch signifikanter Unterschied festgestellt wird, bietet dies die Grundlage, um anzunehmen, dass die Änderungen oder Bedingungen in den Varianten tatsächlich einen Einfluss auf das Nutzerverhalten haben.

Nullhypothese/Alternativhypothese: Die Nullhypothese ist in der Statistik eine Annahme und besagt, dass es keinen signifikanten Unterschied oder keinen Zusammenhang zwischen den untersuchten Variablen (Original und Variante) gibt. Es wird versucht, durch die Analyse der Daten die Nullhypothese zu widerlegen. (Witzenleiter, 2021, S. 63) Wenn die Daten stark genug sind, um die Nullhypothese zu widerlegen, deutet dies darauf hin, dass es tatsächlich einen echten Unterschied oder Zusammenhang gibt, der nicht auf Zufall oder Stichprobenfehler zurückzuführen ist. Die Ablehnung der Nullhypothese ist ein Hinweis darauf, dass weitere Untersuchungen oder Maßnahmen gerechtfertigt sein könnten, um den beobachteten Effekt genauer zu untersuchen.

Daher gibt es die Alternativhypothese die ausdrückt, dass es einen signifikanten Unterschied oder Zusammenhang zwischen den untersuchten Variablen gibt. Durch die Alternativhypothese wird gezeigt, dass die Daten oder Varianten nicht auf Zufall oder Stichprobenfehler zurückzuführen sind. Deshalb werden in einem A/B-Test beide Hypothesen gegeneinander getestet, wobei in der Regel das Ziel verfolgt wird, die Nullhypothese abzulehnen und die

Alternativhypothese anzunehmen. Dies geschieht, um zu beweisen, dass die (Alternativ-) Variante eine höhere Anzahl von Conversions mit sich bringt. (Witzenleiter, 2021, S. 63)

Fehler der 1. Art und Fehler der 2. Art: Jede Entscheidung basierend auf einer Hypothese kann falsch sein. Beim Testen von Hypothesen gibt es daher zwei Arten von Fehlern.

Fehler der 1. Art (α -Fehler): Ein Fehler der 1. Art tritt auf, wenn man fälschlicherweise die Nullhypothese ablehnt, obwohl sie tatsächlich wahr ist. Mit anderen Worten, man zieht den Schluss, dass es einen Effekt oder einen Unterschied gibt, obwohl es in der Grundgesamtheit keinen echten Effekt gibt. Die Wahrscheinlichkeit für einen Fehler der 1. Art wird durch das gewählte Signifikanzlevel (α) bestimmt. (Studyflix) Wenn das Signifikanzlevel beispielsweise 0,05 ist, bedeutet dies, dass es eine 5%ige Wahrscheinlichkeit für einen Typ-I-Fehler gibt. (Lecturio)

Fehler der 2. Art (β -Fehler): Ein Fehler der 2. Art tritt auf, wenn man fälschlicherweise die Nullhypothese akzeptiert, obwohl sie tatsächlich falsch ist. (Studyflix) Mit anderen Worten, man zieht den Schluss, dass es keinen Effekt oder keinen Unterschied gibt, obwohl es in der Grundgesamtheit einen echten Effekt gibt. Die Wahrscheinlichkeit für einen Fehler der 2. Art wird als β (Beta) bezeichnet und steht in direktem Zusammenhang mit der statistischen Power ($1-\beta$) des Tests. (Lecturio)

Signifikanzlevel: Das Signifikanzlevel oder auch Signifikanzniveau genannt, wird in statistischen Hypothesentests verwendet. Es ist ein vorab festgelegter Schwellenwert, der bestimmt, wie hoch die Wahrscheinlichkeit sein soll, dass ein statistischer Unterschied als signifikant angesehen wird und nicht auf Zufall oder Stichprobenfehler zurückzuführen ist. (EUPATI) Im Gegensatz zur statistischen Signifikanz wird das Signifikanzlevel vor der Analyse festgelegt und dient dazu, die Entscheidung über die statistische Signifikanz eines Ergebnisses zu treffen. Oft wird mit dem Signifikanzlevel der p-Wert in Betracht gezogen. „Der sogenannte p-Wert ist nun das Signifikanzlevel im tatsächlichen Test, also sozusagen der gemessene Zufall im Experiment. Damit lässt sich das Signifikanzniveau überprüfen. Einfach ausgedrückt, ist der Test signifikant, sobald der p-Wert kleiner als das Signifikanzlevel ist.“ (Witzenleiter, 2021, S. 64)

Irrtumswahrscheinlichkeit: Sie ist eng mit dem Signifikanzlevel verbunden und bezieht sich auf die Wahrscheinlichkeit, die Nullhypothese abzulehnen, wenn sie tatsächlich wahr ist. Sie wird durch das Signifikanzlevel α (Alpha) angegeben, das üblicherweise auf einen vordefinierten Wert (p-Wert festgelegt wird (z. B. 0,05 oder 0,01). Ein Signifikanzlevel von 0,05 bedeutet eine Irrtumswahrscheinlichkeit von 5 %. Das heißt, wenn der berechnete p-Wert, der die

Wahrscheinlichkeit für das Auftreten der beobachteten Daten angibt, kleiner ist als das Signifikanzlevel, wird die Nullhypothese abgelehnt. (Still, 2021)

Konfidenzlevel: Das Konfidenzlevel oder auch Konfidenzniveau genannt, ist eine sogenannte „Aussagewahrscheinlichkeit“. „Das Konfidenzlevel lässt sich als die Wahrscheinlichkeit, dass eine erneute Durchführung des Tests dasselbe Ergebnis bringt, beschreiben“ (Witzenleiter, 2021, S. 26) Das Signifikanzniveau und das Konfidenzniveau sind miteinander komplementär verbunden. Das bedeutet, bei einem A/B-Test mit einem Konfidenzlevel von 95 % lässt sich mit einer Wahrscheinlichkeit von 95 % feststellen, dass die Ergebnisse nicht allein durch Zufall entstanden sind. Es deutet darauf hin, dass es einen tatsächlichen Unterschied zwischen den getesteten Varianten gibt. Das Signifikanzlevel gibt an, dass man in diesem Fall eine Irrtumswahrscheinlichkeit von 5 % akzeptiert. (Witzenleiter, 2021, S. 64)

Konfidenzintervall: Das Konfidenzintervall und das Konfidenzniveau stehen in engem Zusammenhang. „Das Konfidenzintervall gibt den Bereich an, der mit einer gewissen Wahrscheinlichkeit (dem Konfidenzniveau) die Verteilung einer Zufallsvariable einschließt.“ (Witzenleiter, 2021, S. 158) – was nichts anderes bedeutet, dass der tatsächliche Wert in einem bestimmten Bereich liegt, der auf unseren Daten basiert.

Statistische Power: Bei der statistischen Power oder auch Teststärke/power genannt, „handelt es sich um die Wahrscheinlichkeit, dass die Nullhypothese widerlegt werden kann.“ (Witzenleiter, 2021, S. 27) – In anderen Worten, es ist die Wahrscheinlichkeit, dass der Test einen echten Effekt findet, falls es einen gibt. Eine hohe statistische Power verringert außerdem die Wahrscheinlichkeit für Fehler der 2. Art und erhöht die Zuverlässigkeit des Tests. (Hemmerich, 2016)

Varianz: „Die Varianz ist ein Maß für die Streuung, welches die quadratische Abweichungen der Stichprobenwerte vom Mittelwert quantifiziert.“ (Oestreich & Romberg, 2009, S. 96) Um es einfach zu erklären sagt die Varianz aus, wie sehr die Werte in einer Datenmenge voneinander abweichen. Es misst die Streuung oder Variation der Wert um den Durchschnitt. Um die Varianz berechnen zu können, muss zuerst der Mittelwert ermittelt werden. Anschließend wird für jeden Datenpunkt die Abweichung vom Durchschnitt berechnet. Diese Abweichungen werden quadriert, um die Streuung der Daten um den Durchschnitt zu erfassen. Schließlich werden alle quadrierten Abweichungen summiert und durch die Anzahl der Datenpunkte geteilt, um die Varianz zu erhalten. (DATAtab Team, 2023b)

Standardabweichung: Die Standardabweichung (STDEV) ist die Wurzel aus der Varianz, die die durchschnittliche Abweichung vom Mittelwert angibt. „Die Standardabweichung ist das

wichtigste Maß für die Streuung der Werte in einer Stichprobe“ (Oestreich & Romberg, 2009, S. 99) Im Zusammenhang mit A/B-Testing bezieht sich die Standardabweichung auf die Streuung der Ergebnisse innerhalb der einzelnen Varianten A und B. Sie gibt an, wie sehr die Ergebnisse innerhalb jeder Gruppe (A oder B) um den Mittelwert dieser Gruppe variieren. Eine geringe Standardabweichung in einer Variante deutet darauf hin, dass die Ergebnisse innerhalb dieser Variante konsistent sind und nur geringfügige Unterschiede aufweisen. Eine hohe Standardabweichung in einer Variante zeigt hingegen, dass es eine größere Abweichung in den Reaktionen der Benutzer innerhalb dieser Variante gibt. (Bhandari, 2020)

Ideale Stichprobe: In der Statistik ist die Stichprobe eine Teilmenge der untersuchten Grundgesamtheit. Diese Stichprobe besteht aus einer zufälligen Auswahl von Nutzern, die den Varianten A und B zugeordnet werden. Die Auswahl der Stichprobe sollte sorgfältig erfolgen, um eine Verzerrung der Ergebnisse zu vermeiden und eine repräsentative Verteilung der Merkmale in der Stichprobe sicherzustellen. Eine ausreichend große Stichprobe ist von weiterer Bedeutung, um statistische Signifikanz zu gewährleisten und verlässliche Schlussfolgerungen ziehen zu können. (Stotz, 2022, S. 112–113) Eine angemessene Stichprobengröße ermöglicht es, statistische Tests anzuwenden, um festzustellen, ob die beobachteten Unterschiede zwischen den Varianten statistisch signifikant sind. Die sorgfältige Auswahl einer repräsentativen Stichprobe und eine ausreichende Stichprobengröße sind somit entscheidend, um valide und aussagekräftige Ergebnisse im A/B-Testing zu erzielen.

Minimum Detectable Effekt: Im A/B-Testing bezieht sich der Begriff "Minimum Detectable Effect" (MDE), dt. minimal nachweisbaren Effekt, auf die kleinste messbare Veränderung oder den kleinsten Effekt, den man mit einem A/B-Test zuverlässig nachweisen möchte. Es ist der Mindestwert, ab dem man davon ausgeht, dass eine Änderung einen tatsächlichen Einfluss hat. Der MDE wird vor dem Test festgelegt und dient als Maßstab, um festzustellen, ob die beobachteten Unterschiede zwischen den Varianten tatsächlich statistisch signifikant sind. Wenn der beobachtete Effekt größer oder gleich dem MDE ist, wird angenommen, dass die Änderung oder Bedingung einen praktisch relevanten Einfluss hat. Ein kleinerer MDE erfordert eine größere Stichprobe, um den Effekt zuverlässig nachweisen zu können, während ein größerer MDE eine kleinere Stichprobe erfordert. Der MDE ist von großer Bedeutung, um die Testgröße zu bestimmen und sicherzustellen, dass der Test ausreichend leistungsfähig ist, um den gewünschten Effekt zu erkennen. (Analytics Toolkit)

3.1.5 User Experience

In diesem Kapitel wird der Begriff User Experience (UX) ausführlich erläutert, da die UX einen wesentlichen Einfluss auf die Zufriedenheit der Nutzer und den Erfolg von Produkten und Dienstleistungen hat und daher in deren Gestaltung immer wichtiger wird.

„Die DIN-Norm „Prozess zur Gestaltung gebrauchstauglicher interaktiver Systeme“ definiert User Experience allgemein als: „Wahrnehmungen und Reaktionen einer Person, die aus der tatsächlichen und/oder der erwarteten Benutzung eines Produktes, eines Systems oder einer Dienstleistung resultieren.“ (DIN EN ISO 9241-210: 2011-01, S. 7)“ (Gast, 2018, S. 13)

Während der Begriff „Usability“ sich auf die Effektivität und Effizienz der Benutzeraktion konzentriert und ein Aspekt der UX ist, geht die User Experience darüber hinaus und betrachtet das gesamte Nutzererlebnis, einschließlich ästhetischer, emotionaler und motivationsbezogener Faktoren. (Becker, 2015) Somit umfasst es alle Aspekte der Interaktion des Benutzers mit einem Produkt oder einer Dienstleistung - von der ersten Interaktion mit dem Produkt oder Dienstleistung bis hin zur Nachbereitung geht es darum ein positives Erlebnis für den Benutzer zu schaffen.

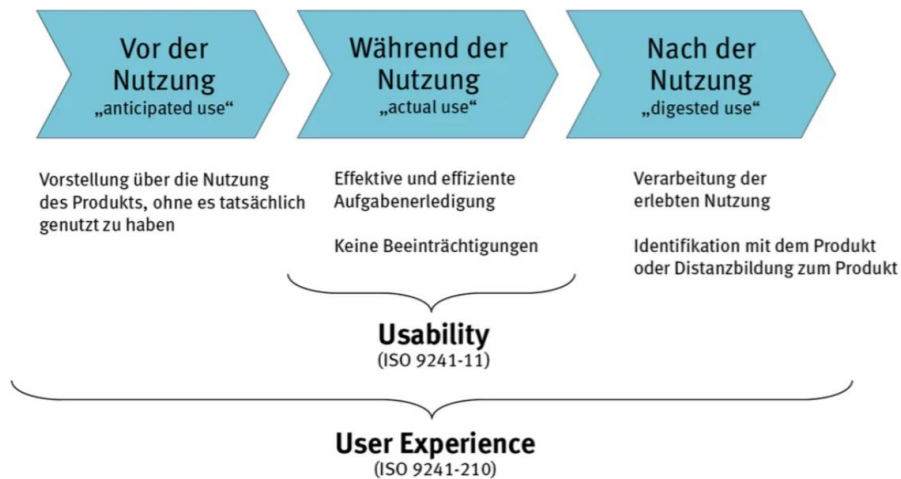


Abbildung 7: Die Usability ist ein Teil der User Experience (Vgl. Becker, 2015)

3.2 Einführung in die A/B Tests

In diesem Kapitel der Bachelorarbeit wird der A/B-Testing-Prozess, einschließlich seiner Vorgehensweise, Ziele und Vorteile untersucht. Die Vorgehensweise des A/B-Testings wird detailliert dargestellt, um ein klares Verständnis der einzelnen Schritte zu vermitteln. Beginnend mit der Zielsetzung der KPIs und der Formulierung einer Hypothese werden die verschiedenen Aspekte der Testplanung, der Entwicklung von Testvarianten, der Auswahl der Stichprobe, der Durchführung des Tests bis hin zur Analyse der Ergebnisse behandelt. Darüber hinaus werden die Ziele des A/B-Testings erläutert, um zu zeigen wie ein Test die Benutzererfahrung und Conversion Rate erhöhen kann. Zudem werden ebenfalls die Vorteile des A/B-Testings hervorgehoben, um seine Anwendung in der Geschäftswelt zu unterstreichen.

3.2.1 A/B Testing Prozess: Vorgehensweise, Ziele, Vorteile und Methoden

A/B-Tests gewinnen immer mehr an Popularität und werden von vielen Unternehmen genutzt, um fundierte Entscheidungen zu treffen und die Benutzererfahrung zu verbessern. Anhand des A/B-Testing-Prozess können gezielt Verbesserungen und Optimierungen vorgenommen werden.

Ein typischer Prozess beinhaltet die nachfolgend aufgeführten Phasen und lässt sich in sieben wesentlichen Schritten zusammenfassen:

1. **Research** – Ein A/B-Testing-Prozess beginnt meistens mit einem Research und bezieht sich auf den Prozess der Untersuchung und Datenerhebung. Dabei ist es von entscheidender Bedeutung, die vorhandenen Daten gründlich zu analysieren und zu verstehen, um potenzielle Verbesserungsbereiche identifizieren zu können. Mithilfe von User Researches (Nutzerforschung) oder Webanalytics-Daten, wie bspw. durch Online-Umfragen, Session Recordings, Heatmaps oder Google Analytics können UN das Nutzerverhalten ihrer Zielgruppe verstehen und bessere, datengestützte Entscheidungen treffen. (Me&company)
2. **Zielsetzung inklusive Festlegung der KPIs** – Basierend auf der Research Analyse sollten spezifische Ziele festgelegt werden, um Bereiche der Optimierung zu bestimmen. Zudem sollten KPIs festgelegt werden, um diese Optimierung messen zu können. Häufig besteht das Ziel für Unternehmen darin, die CR zu optimieren. „Statistisch ausgedrückt wäre der erste Schritt die Bestimmung einer oder mehrerer abhängiger Variablen.“ (Boßow-Thies et al., 2020, S. 167) Ein Experiment besteht stets aus

mindestens einer unabhängigen und einer abhängigen Variable. Das heißt, bei einem A/B-Test ist die abhängige Variable das Ergebnis, das man messen möchte, um die Auswirkungen der Änderung zu bestimmen. Die unabhängige Variable ist das Element, das man ändern oder testen möchte, um ihren Einfluss auf andere Variablen zu messen. (Voxco) Anhand eines Beispiels, in dem die Farbe einer Schaltfläche von Grün auf Rot geändert und die Klickrate gemessen wird, lässt sich das Konzept der unabhängigen und abhängigen Variablen verdeutlichen. In diesem Szenario ist die Farbe der Schaltfläche die unabhängige Variable, während die Klickrate die abhängige Variable darstellt. Durch die Veränderung der unabhängigen Variable (Farbe der Schaltfläche) wird gemessen, wie sich dies auf die abhängige Variable (Klickrate) auswirkt. „Von zentraler Relevanz für die Ergebnisqualität ist, dass jede Veränderung einer unabhängigen Variablen zu einer neuen Variante führt, die unter sonst gleichen Bedingungen getestet wird.“ (Boßow-Thies et al., 2020, S. 168) Daher ist es von Bedeutung, dass alle anderen Aspekte der getesteten Elemente gleich bleiben, da ansonsten nicht sichergestellt werden kann, welche Änderung zu den beobachteten Ergebnissen geführt hat. Natürlich besteht die Möglichkeit, einen multivariaten Test durchzuführen, der den zusätzlichen Vorteil bietet, dass er Interaktionen zwischen unabhängigen Variablen erfassen kann. Jedoch besteht der Nachteil darin, dass man eine große erforderliche Stichprobengröße pro Variante benötigt, um aussagekräftige Ergebnisse erzielen zu können. (Boßow-Thies et al., 2020, S. 168–169) Daher ist es im Anfangsstadium des A/B-Testings entscheidend, ein klares Verständnis der Ziele und der zur Verfügung stehenden Ressourcen zu haben.

- 3. Entwicklung von Ideen und begründeten Hypothesen** – Im zweiten Schritt werden die ersten Ideen und die dazugehörigen Hypothesen ausgearbeitet und entwickelt. „Ideen sowie Hypothesen können anhand kreativer Prozesse, auf Basis bisheriger Daten zum Nutzerverhalten oder auf Grundlage von Recherchen generiert werden.“ (Boßow-Thies et al., 2020, S. 163) Somit können die Ideen für den Test aus verschiedenen Quellen stammen wie beispielsweise aus Benutzerfeedbacks, Usability-Test, Analysen von Nutzerverhalten oder anderen Datenquellen. Mögliche Ideen könnten Änderungen an verschiedenen Elementen sein, wie z.B. die Farbe oder Position von CTA Buttons, die Formulierung von Texten oder die Ausarbeitung neuer Funktionen. Sobald die Idee entwickelt wurde, ist der nächste Schritt die Formulierung von Hypothesen. „Eine Hypothese ist eine Annahme, die weder bestätigt noch widerlegt ist. Im Forschungsprozess wird eine Hypothese gleich zu Beginn aufgestellt und das Ziel ist es, diese Hypothese entweder abzulehnen oder bei-zubehalten.“ (DATA-tab Team, 2023a)

Ein Beispiel einer Hypothesenformulierung könnte wie folgt aussehen: „Wenn die Farbe der Schaltfläche von Grün auf Rot geändert wird, dann steigt die Klickrate, weil Rot als Signalfarbe besser hervorsteht und dadurch die Aufmerksamkeit der Nutzer verstärkt auf sich zieht.“ Wichtig bei der Formulierung ist, dass die Hypothesen präzise, aussagekräftig, sinnvoll und messbar sind. (Dodt, 2020) Sie spielen eine entscheidende Rolle im A/B-Testing-Prozess und bilden das Fundament dafür.

4. **Erstellung der Testvarianten** – Nachdem die Ideen und die Hypothesen ausgearbeitet wurden, erfolgt die Erstellung der Testvarianten. In diesem Schritt werden zwei oder mehr Versionen des Elements, das man testen möchte erstellt. Die Kontrollversion (Version A) und die Testversion (Version B) sollten so identisch wie möglich sein, mit Ausnahme der spezifischen Änderung, die man testen möchte. Dies stellt sicher, dass alle Unterschiede in der Leistung zwischen den beiden Versionen auf die Änderung zurückzuführen sind und nicht auf andere Faktoren. Wie viele Versionen letztendlich getestet werden, hängt vom jeweiligen Test ab. Dies bedeutet, dass sie einerseits von der Anzahl der zu testenden Merkmale abhängt und andererseits von der erreichbaren Stichprobengröße. (Boßow-Thies et al., 2020, S. 165)

Oft erfordert die Erstellung der Testvarianten die Zusammenarbeit zwischen verschiedenen Teams. Designer, Entwickler, Produktmanager oder Projektmanager arbeiten oft zusammen und müssen gemeinsam die Testversionen sorgfältig erstellen, bevor der Test gestartet werden kann.

5. **Durchführung des Experiments** – Die Durchführung des Experiments stellt den vierten Schritt im A/B-Testing-Prozess dar, in dem die Testvarianten den Nutzern tatsächlich vorgeführt und Daten erfasst werden. Dabei werden die unterschiedlichen Versionen durch eine zufällige Zuteilung der Nutzer gleichmäßig verteilt, um eine Verzerrung der Ergebnisse durch eine ungleiche Verteilung zu vermeiden. Dabei verläuft die Verteilung über das A/B-Testing-Tool, welches über ein Code Snippet (wird in den Head des Shops eingebaut) verfügt. Dadurch ist eine 50:50 Verteilung bzw. eine gleichmäßige Verteilung der Besucher möglich. Während die verschiedenen Tests den jeweiligen Benutzern ausgespielt werden, werden die Daten über die Interaktionen der Benutzer gesammelt. Je nach Interesse könnten das Daten über Klickraten, ARPU's, CR's oder andere wichtige Metriken sein. Die erhobenen Daten dienen anschließend dazu, eine detaillierte Analyse der Leistungsfähigkeit verschiedener Versionen durchzuführen. Dabei ist es zu beachten, dass die Daten regelmäßig überprüft werden, damit sichergestellt werden kann, dass der Test wie geplant läuft. Zudem sollte darauf

geachtet werden, den Test nicht zu früh abubrechen, da es sonst zu voreiligen Schlussfolgerungen führen kann. Im Kontext von A/B-Tests wird der Begriff „Peeking“ dafür verwendet, was letztendlich bedeutet, dass die Ergebnisse zu früh interpretiert werden, bevor der Test statistisch signifikant ist. „Daher ist es ratsam, während der Datenerhebung, die zuvor als notwendig bzw. sinnvoll erachtete Stichprobengröße abzuwarten, bevor der Test beendet und die Ergebnisse interpretiert werden“ (Vgl. Boßow-Thies et al., 2020, S. 167)

Durch den Vergleich dieser Daten kann die vorliegende Hypothese entweder bestätigt oder widerlegt werden.

6. **Datenanalyse und Interpretation der Ergebnisse** – „Zur Datenanalyse und Interpretation der Ergebnisse bedarf es eines adäquaten statistischen Testverfahrens“ (Boßow-Thies et al., 2020, S. 167) Welches Verfahren letztendlich verwendet wird, hängt von verschiedenen Faktoren ab, beispielsweise der Art der Daten oder der Anzahl der zu testenden Versionen. „Je nach Skalenniveau (Konzept aus der Statistik, das die Art der Messung oder Kategorisierung von Daten beschreibt) und Verteilung der Variablen bieten sich unterschiedliche Verfahren an. (Boßow-Thies et al., 2020, S. 167)

7. **Verwendung und Monitoring der besten Variante** – Nach der Analyse der Tests besteht der abschließende Schritt im A/B-Testing-Prozess spezifisch darin, die überlegene Variante zu implementieren. „Die Erfolgskennzahlen sollten zudem zur Überprüfung der Validität weiterhin gemessen werden.“ (Boßow-Thies et al., 2020, S. 167) Das heißt, dass das Monitoring stets fortgesetzt werden sollte, da dies sicherstellt, dass die Verbesserungen, die während des Tests beobachtet wurden, auch in der Praxis beibehalten werden. Zudem dient das Monitoring auch mögliche unerwarteten Nebenwirkungen zu erkennen, aber auch gleichzeitig weitere Bereiche für Verbesserungen zu identifizieren. Somit ist das Monitoring bzw. das A/B-Testing ein kontinuierlicher Prozess, der dazu beiträgt, die Benutzererfahrung ständig zu optimieren.

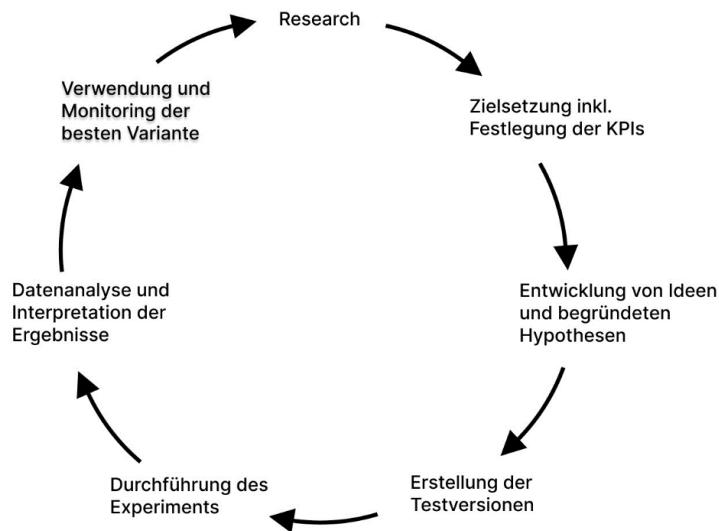


Abbildung 8: Beispiel eines A/B-Testing-Prozesses

(Vgl. Boßow-Thies et al., 2020, S. 163)

3.2.2 Fixed-Horizon Tests vs. sequential A/B-Tests: Unterschiede und Anwendungsbereiche

In Kapitel 3.1.1 wurde der Begriff A/B-Testing und die zwei gängigsten Testmethoden bereits erläutert. Durch das A/B-Testing bzw. den Vergleich von zwei oder mehr Varianten eines Elements können Unternehmen fundierte Entscheidungen darüber treffen, welche Änderungen die besten Ergebnisse liefern. Zur Durchführung von A/B-Tests stehen den Anwendern hauptsächlich die folgenden 2 Methoden zur Verfügung: Fixed-Horizon-Tests und sequential A/B-Tests. Da jeder dieser Ansätze seine eigenen Stärken und Schwächen sowie unterschiedliche Anwendungsbereiche hat, werden in diesem Kapitel die Unterschiede zwischen Fixed-Horizon-Tests und sequential A/B-Tests untersucht und diskutiert, wann und warum man den einen oder den anderen Ansatz verwenden sollte.

Fixed-Horizon-Tests

Fixed-Horizon-Tests sind eine bewährte statistische Methode, die dazu dienen die Wirksamkeit unterschiedlicher Varianten eines Elements oder eines Produkts zu bewerten. Der Begriff „Fixed-Horizon“ bezieht sich auf die Durchführung eines Tests über einen vorher festgelegten Zeitraum. Dieser spezifische Zeitraum, auch als „Horizont“ bekannt, wird im Voraus festgelegt und bleibt während des gesamten Tests unverändert. Erst am Ende des festgelegten Zeithorizonts werden die Daten einmalig analysiert sowie ausgewertet und auf dessen Grundlage

eine Entscheidung getroffen. Ein wichtiger Aspekt der beim Fixed-Horizon-Test zu beachten ist, ist die statistische Signifikanz. In der Theorie wird ein Signifikanzniveau von 0.05 verwendet, was bedeutet das man zu 95% sicher sein kann, dass die beobachteten Unterschiede nicht auf Zufall beruhen. Jedoch ist es wichtig, dass das Signifikanzniveau von den spezifischen Anforderungen eines Tests abhängen kann. Im Kontext der Marketing-forschung ist auch oft, ein Signifikanzniveau von 80% akzeptabel, da dies in vielen Fällen ausreichend ist, um fundierte Entscheidungen zu treffen.

Angesichts der vielen Vorteile, die der Fixed-Horizon-Test bietet, stellt er eine attraktive Methode für Unternehmen dar. Der Hauptvorteil von Fixed-Horizon-Tests ist die Einfachheit und Verständlichkeit. Sie sind im Vergleich zu anderen statistischen Testmethoden einfacher durchzuführen, da bei einem Fixed-Horizon-Test die Testparameter von Beginn an festgelegt sind, wodurch die Planung vereinfacht wird. Das heißt, UN können im Voraus planen wann der Test beginnt sowie endet und wann sie die Ergebnisse zu erwarten sind, ohne ständige Zwischenanalysen auch zu machen. (Pierce & Wu, 2023)

Ein weiterer Vorteil ist die einfache Analyse durch die Festlegung des Signifikanzniveaus im Voraus. Dadurch können Fehler der 1. Art besser kontrolliert werden. (Pierce & Wu, 2023) Im Kontext auf das A/B-Testing würde man durch einen Fehler erster Art irrtümlich zu dem Schluss kommen, dass eine Variation besser ist als die andere, obwohl das in Wirklichkeit nicht der Fall ist. Wird ein Signifikanzniveau von 0.05 gewählt bedeutet dies letztendlich, dass es eine 5%ige Chance gibt, dass die Nullhypothese fälschlicherweise abgelehnt wird, wenn sie tatsächlich wahr ist. Durch die Festlegung des Signifikanzniveaus beim Fixed-Horizon-Test, kann somit die Zuverlässigkeit und Genauigkeit der Testergebnisse gewährleistet werden. Nichtsdestotrotz gibt es auch beim Fixed-Horizon-Test Nachteile die zu erwähnen sind. Ein großer Nachteil von Fixed-Horizon-Tests ist das sogenannte „Peeking Problem“. Im Bereich des A/B-Testings bezieht der Begriff darauf, dass man wiederholt auf die Daten eines laufenden Experiments schaut und vorzeitig Entscheidungen trifft, bevor das Experiment abgeschlossen ist. (Pekelis, 2015)

Um das „Peeking Problem“ zu verdeutlichen, wird der Begriff anhand eines Beispiels erläutert. Bei einem A/B-Test werden zwei verschiedene Designvarianten auf einer Webseite getestet, um die Leistungsfähigkeit der Layouts zu vergleichen. Angenommen, ein Datenanalyst wirft nach der Hälfte der Testdauer einen Blick auf die vorliegenden Daten und stellt fest, dass Variante B eine höhere Anzahl von Klicks verzeichnet. In dieser Situation kann das sogenannte Peeking-Problem auftreten, da der Datenanalyst in Versuchung geraten könnte, den Test vorzeitig zu beenden und Variante B als den überlegenen Gewinner zu deklarieren. Es ist jedoch zu beachten, dass die Variante A möglicherweise besser abgeschnitten hätte, wenn der Test bis zum Ende der geplanten Testdauer durchgeführt worden wäre.

Wird das Peeking-Problem somit ignoriert kann es „eine erhebliche Bedrohung für die Validität jedes online kontrollierten Experiments sein, da es die Fehlerrate erster Art unkontrolliert erhöhen und jegliche Signifikanzberechnungen oder Konfidenzintervalle bedeutungslos machen kann.“ (Analytics Toolkit) Einfach gesagt, bedeutet dies, dass man keine aussagekräftigen Ergebnisse erzielt, was letztendlich auch dann zu falschen Schlussfolgerungen und Entscheidungen führen kann.

Ein weiterer Nachteil dieser Methode, ist die nicht Beachtung von unvorhergesehenen Ereignissen oder Schwankungen, die während eines Tests auftreten können. Aufgrund von externen Faktoren oder saisonalen Einflüssen kann es vorkommen, dass Testergebnisse beeinflusst werden. Durch den festen Zeitraum kann es dann vorkommen, dass Ergebnisse verzerrt werden und diese zu falschen Ergebnissen führen. Zwar ist es schwierig, unvorhergesehene Ereignisse oder Schwankungen miteinzubeziehen, dennoch ist es hilfreich bei der Anlegung des Tests mögliche potentielle unvorhergesehene Faktoren mit-zuberücksichtigen. (Pierce & Wu, 2023)

Sequential A/B -Tests

Sequential A/B-Tests, dt. „Sequenzielle A/B-Tests“ werden im Gegensatz zu Fixed-Horizon-Tests kontinuierlich in denselben zuvor festgelegten Zeitabständen überwacht. Wenn ausreichende Daten und Beweise vorliegen, um eine Entscheidung zu treffen, kann der Test beendet werden. Das heißt, wenn der Test eine statistische Signifikanz erreicht hat, kann dieser gestoppt und daraus sinnvolle Ergebnisse abgeleitet werden. Statt also einem kompletten Zeitfenster abzuwarten, ermöglicht das sequenzielle Testing einen früheren Gewinner der vergleichenden Varianten zu ermitteln. (Split) Wie auch bei der Fixed-Horizon-Test Methode hat der sequenzielle A/B-Test seine Vor- und Nachteile. Der größte Vorteil des sequenziellen A/B-Testing Ansatzes ist die Effizienz. „Durch das häufige Blicken auf die Daten können sequenzielle Tests die Experimentdauer verkürzen, wenn die Effektgröße größer ist als der minimal nachweisbare Effekt (MDE).“ (Prakash, 2022) Sobald die Daten zeigen, dass der Unterschied zwischen den Varianten größer ist als der MDE kann der Test gestoppt werden. Somit muss nicht gewartet werden, bis der gesamte Test abgeschlossen ist, wodurch Ressourcen eingespart werden können. Des weiteren können während der Durchführung des Tests eine Schätzung der gültigen Konfidenzintervalle und p-Werte angegeben werden. Diese sind wichtige Metriken in der Statistik, um die Zuverlässigkeit und Signifikanz von Testergebnissen zu bewerten. Durch die kontinuierliche Sammlung der Daten, kann unter anderem geschätzt werden, wo der wahre Effekt der getesteten Varianten liegen könnte. Diese Informationen können nützlich sein, um zu bestimmen, ob genügend Daten gesammelt wurden, um den Test zu beenden, oder ob die Datenerhebung fortgesetzt werden sollte. (Prakash, 2022)

Da bei einem sequenziellen A/B-Test keine Stichprobengröße im Voraus festgelegt werden muss, ist dies ein weiterer Vorteil. Die Testmethode ermöglicht eine dynamische Anpassung der Stichprobengröße während des Tests. Im Gegensatz zum Fixed-Horizon-Test, bei dem eine bestimmte Stichprobengröße im Voraus erforderlich ist, bedarf es beim sequenziellen A/B-Test keiner Notwendigkeit eines Stichprobengrößenrechners oder der Berechnung der Stichprobengröße mit einer Formel. Oft haben „nicht-technisch“ versierte Personen Schwierigkeiten, die Stichprobengröße mit einem Rechner bspw. zu ermitteln und verstehen nicht wie sie die erforderlichen Zahlen berechnen können. „Das Wissen über die Standardabweichung einer Metrik ist beispielsweise etwas, das die meisten Menschen nicht aus dem Stegreif wissen“ (Vgl. Prakash, 2022)

Da somit beim sequenziellen A/B-Testing kein umfangreiches Vorwissen über das Festlegen verschiedener Parameter erforderlich ist, wird die Eintrittsbarriere gesenkt und ermöglicht es auch Personen ohne spezifisches statistisches Fachwissen, Tests durchzuführen. (Prakash, 2022)

Obwohl das sequenzielle A/B-Testing viele Vorteile bietet, gibt es auch Nachteile dieser Testmethode, die berücksichtigt werden sollten. Das sequenzielle Testing erfordert mehr Vorsicht und Sorgfalt, da genau überwacht werden muss, um eine frühes oder zu spätes Beenden der Tests zu vermeiden. Zudem können sequenzielle Tests dazu führen, dass echte Unterschiede zwischen den Gruppen möglicherweise nicht so leicht erkannt werden wie bei der Analyse aller Daten auf einmal. Das heißt, wenn die Daten Schritt für Schritt während des Tests beobachtet werden, könnten die Muster, die man sieht, stärker von zufälligen Veränderungen beeinflusst werden. (Fantaye, 2021)

Des Weiteren kann es vorkommen, dass der Test eine größere Stichprobengröße erfordert, um eine statistische Signifikanz zu erreichen, wenn der Unterschied zwischen den getesteten Varianten zu klein oder gering ist. In solchen Situationen kann das sequenzielle A/B-Testing nicht von den Vorteilen einer verminderten Stichprobengröße profitieren und möglicherweise zu einer längeren Testdauer führen. (Aho, 2020)

Insgesamt hängt die Wahl zwischen Fixed-Horizon-Tests und sequential A/B-Tests von den spezifischen Umständen und Zielen des Tests ab. Beide Ansätze haben ihre Stärken und Schwächen, die zu betrachten sind.

Im Folgenden werden daher allgemeine Empfehlungen gegeben, wann die jeweiligen Methoden üblicherweise in ihrem Anwendungsbereich verwendet werden. Das sequenzielle A/B-Testing wird häufig eingesetzt, wo eine hohe Anzahl von Datensätzen pro Variante gegeben ist. Im Kontext des A/B-Testings im E-Commerce sind Datensätze oft die Anzahl der Besucher oder Nutzer auf einer Webseite gemeint. Der Fixed-Horizon-Test wird hingegen verwendet,

wenn eine geringere Datenmenge - weniger als 500 Datensätze pro Variante vorhanden ist. Im Allgemeinen ist die sequenzielle Methode besonders nützlich, wenn das Hauptinteresse darin besteht, signifikante oder große Veränderungen in einer Metrik wie bspw. der CR oder dem ARPU zu erkennen. Das sequenzielle A/B-Testing ermöglicht eine frühere Erkennung von größeren Veränderungen, da es regelmäßige Überprüfungen der Ergebnisse ermöglicht. Demgegenüber ist der der Fixed-Horizon-Test eher für kleine Veränderungen besser geeignet. Er eignet sich auch für Fälle, in denen es nicht notwendig ist, die Ergebnisse vor ihrer Fertigstellung zu überprüfen, und in denen potenzielle falsch-positive Ergebnisse weniger Auswirkungen haben. Wenn die Auswirkungen von falsch-positiven Ergebnissen von Bedeutung sind, ist das sequenzielle Testing aufgrund seines geringeren Risikos die bevorzugte Wahl. Des Weiteren erfordert das sequenzielle Testing im Vergleich zum Fixed-Horizon-Test keine umfangreichen Vorkenntnisse. Somit ermöglicht es, den Test direkt zu starten und die Ergebnisse regelmäßig zu überprüfen, ohne dass umfangreiche statistische Analysen im Vorfeld durchgeführt werden müssen. (Pierce & Wu, 2023) (split)

Letztendlich aber hängt die Wahl der geeigneten Methode von verschiedenen Faktoren ab, wie beispielweise der Dringlichkeit der Ergebnisse oder der Auswertung bestimmter Kennzahlen. Es ist wichtig die Anforderungen und Rahmenbedingungen des Experiments mit zu berücksichtigen.

3.2.3 Bedeutung und Positionierung von A/B-Tests im E-Commerce

Der E-Commerce hat die Art und Weise, wie Unternehmen ihre Produkte und Dienstleistungen anbieten und Kunden einkaufen, revolutioniert. „E-Commerce“ – Electronic Commerce wird auch oft mit den verwandten Begriffen wie Online-Handel, E-Retailing oder Internetvertrieb beschrieben. „Im Kern geht es um den elektronischen Handel mit Waren und Dienstleistungen, deren Transaktion, d. h. die Anbahnung, der Abschluss und die Abwicklung des Kaufs oder Verkaufs, über das Internet mithilfe interaktiver Informations- und Kommunikationstechnologien durchgeführt wird.“ (DEGES, 2020, S. 2)

Die Statistik „Umsatz durch E-Commerce (B2C) in Deutschland in den Jahren 1999 bis 2022 sowie eine Prognose für 2023 (in Milliarden Euro)“ von Statista Research Department zeigt, wie rasant der Umsatz im Bereich Business-to-Customer E-Commerce gestiegen ist. So beliefen sich die Netto-Umsätze im Jahr 2022 auf rund 84,5 Milliarden Euro. Trotz eines Rückgangs von 2,5 Prozent im Vergleich zum Vorjahr bleibt der Umsatz weiterhin auf einem bemerkenswert hohen Level. Auch die Prognose für das Jahr 2023 deutet darauf hin, dass der Umsatz weiterhin auf einem vergleichsweise hohen Niveau bleiben wird. (HDE, 2023)

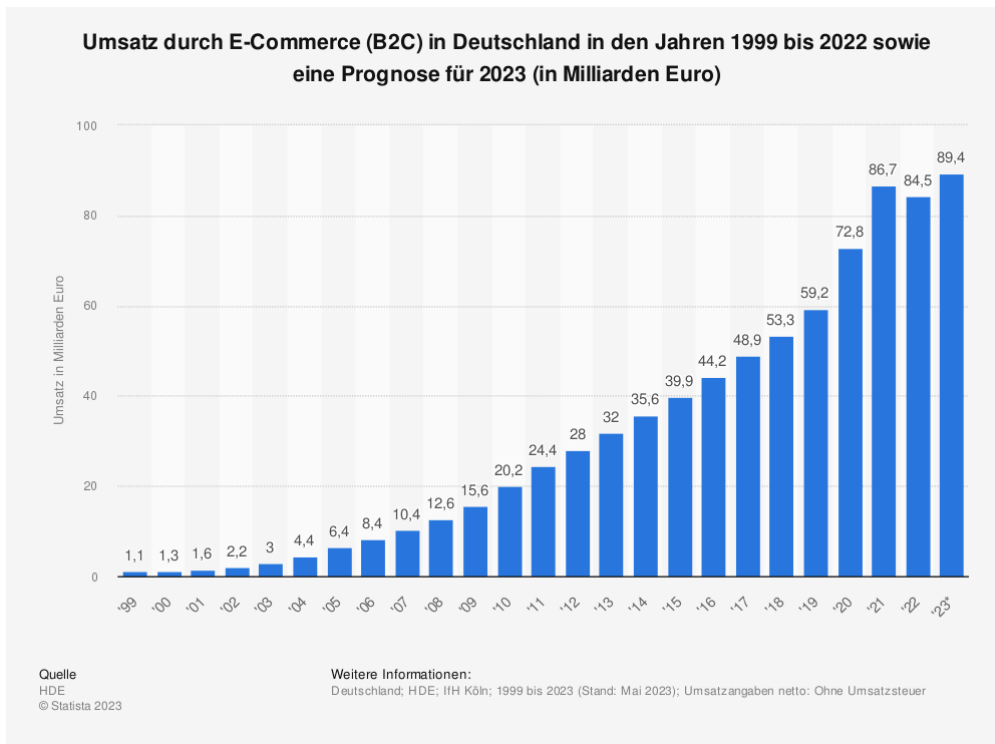


Abbildung 9: Umsatz durch E-Commerce (B2C) in Deutschland in den Jahren 1999 bis 2022 sowie eine Prognose für 2023 (in Milliarden Euro)

(HDE, 2023)

Durch das stetige Wachstum und den stetigen Wandel im E-Commerce ist es daher von großer Bedeutung seine Kunden zu kennen und stets seine Webseite oder Dienstleistung sowie den Kaufprozess zu optimieren.

Oft spielen deshalb A/B-Tests eine zentrale Rolle bei der Optimierung von Websites, Kundenerfahrungen und Marketingstrategien. In diesem Kapitel wird daher die Bedeutung und Positionierung des A/B-Testings im E-Commerce untersucht. A/B-Tests gewinnen im E-Commerce immer mehr an Relevanz, da sie die Möglichkeit bieten, datenbasierte Entscheidungen und Optimierungen aufgrund fundierter Informationen zu ermöglichen. Eine der wichtigsten Anwendungen von A/B-Tests liegt in der Optimierung der Benutzer-freundlichkeit und der damit verbundenen Customer Journey. Die Customer Journey umfasst den gesamten Prozess, den ein Kunde durchläuft, von der ersten Interaktion mit einer Marke oder einem Produkt bis hin zum Kauf und möglicherweise auch zur Weiterempfehlung. (ADVIDERA) Durch das Testen verschiedener Elemente können Unternehmen somit herausfinden, welche Kombinationen die Benutzer dazu anregen länger auf der Seite zu bleiben, mehr Produkte in den Warenkorb zu legen oder den Checkout-Prozess abzuschließen. Mit jedem A/B-Test werden umfassende Erkenntnisse geliefert, welche Aspekte für die Zielgruppe von Bedeutung sind und welche emotionalen Einflüsse letztendlich wirksam sind. Durch die Implementierung einer

langfristigen Testing-Kultur im Unternehmen wird eine schrittweise Erfassung der langfristigen Kaufmotive und Bedürfnisse von Nutzern ermöglicht. (Witzenleiter, 2021, S. 2)

Des Weiteren können A/B-Tests im E-Commerce auch bei der Steigerung der CR unterstützen und dadurch den Umsatz und gegebenenfalls den Gewinn erhöhen. Durch das Testen verschiedener Varianten können Unternehmen herausfinden welche Version die höhere CR bzw. den höheren Umsatz je Besucher erzielt. „In der Regel haben dabei diese Elemente den größten Einfluss auf die CR:

- Seitenübergreifende Elemente wie Header und Menüleisten
- Kategorie- und Produktseiten
- Check out“ (Knappe)

In einem Onlineshop kann somit bspw. herausgefunden werden:

- Welche Darstellungsformen sowie Filter- und Sortieroptionen der Produkte von der Zielgruppe auf der Kategorieseite bevorzugt werden
- Welche Abbildungen und Inhalte die Zielgruppe ansprechen
- Ob Produktbilder, Produktbeschreibungen, Kundenbewertungen oder Preis darstellungen auf der Produktseite den Benutzer zum Kauf beeinflusst
- Wie der Checkout-Prozess gestaltet ist – ist dieser benutzerfreundlich und schafft Vertrauen & Transparenz?

Darüber hinaus eröffnet ein A/B-Test im Onlineshop die Möglichkeit, zahlreiche weitere Erkenntnisse zu gewinnen.

Wichtig jedoch zu erwähnen ist, dass neben der CR auch andere KPIs eine wichtige Rolle spielen, die berücksichtigt werden sollten. Durch die Betrachtung von Metriken wie der Verweildauer oder Absprungrate können zusätzliche Optimierungspotenziale identifiziert werden. (Knappe)

Ergänzend dazu ist das A/B-Testing eine risikoarme Methode, um Änderungen an einer Webseite vorzunehmen. Anstatt umfangreiche Änderungen auf einmal vorzunehmen, haben UN die Möglichkeit, verschiedene Varianten zu testen und diejenige auszuwählen, die die besten Ergebnisse erzielt. Durch diesen schrittweisen Ansatz werden potenzielle Risiken und Auswirkungen fehlerhafter Änderungen minimiert.

Im E-Commerce hat das A/B-Testing einen maßgeblichen Stellenwert als strategisches Instrument, da es zur Qualitätssicherung beiträgt. (LinkedIn & Puscher, 2015) Das Ziel der Qualitätssicherung besteht darin, sicherzustellen, dass der gesamte Einkaufsprozess entlang der Customer Journey reibungslos und vertrauenswürdig abläuft. Durch gezieltes Testing können

Maßnahmen und Prozesse in einem Onlineshop sichergestellt werden. Die Qualitätssicherung dient somit dazu, das Einkaufserlebnis in Onlineshops zu optimieren, Fehler zu minimieren und die Benutzerfreundlichkeit zu erhöhen. (Puppe, 2022)

3.3 A/B-Teststatistik

Bereits in Kapitel 2.2.1 wurde der A/B-Testing Prozess erklärt. Dabei spielten die Analyse und Interpretation der Ergebnisse auch eine entscheidende Rolle, um fundierte Entscheidungen zu treffen und den Erfolg einer Änderung zu bewerten. In diesem Kapitel werden daher die wichtigsten Schritte erläutert, die für eine erfolgreiche Auswertung eines A/B-Tests notwendig sind. Dieser Prozess beginnt mit der Aufstellung systematischer Hypothesen, umfasst die Auswahl der Testvariable, die Festlegung des Stichprobenumfangs und der Testdauer sowie die Auswahl des statistischen Signifikanzniveaus und der sign. Power.

3.3.1 Notwendigkeit zur Aufstellung systematischer Hypothesen und Auswahl der Testvariable

Hypothesen bilden das Fundament in der Statistik und sind bei der Durchführung von A/B-Tests von großer Bedeutung. Damit aussagekräftige und valide Ergebnisse beim A/B-Testing erzielt werden können, bedarf es der Notwendigkeit zur Aufstellung von systematischen Hypothesen und der damit verbundenen Testvariable. Sie geben klare Aussagen darüber, welche Veränderungen in einem A/B-Test untersucht werden sollen, welche Auswirkungen oder Ergebnisse zu erwarten sind und welche Optimierungspotenziale sich daraus ergeben können. In der Regel entstehen Hypothesen aus einer Idee oder einem theoretischen Hintergrund. Die Idee hinter einer Hypothese kann beispielsweise sein, dass eine bestimmte Veränderung einen bestimmten Effekt auf eine abhängige Variable haben könnte. (Stotz, 2022, S. 16–17) Diese Idee wird dann in Form einer Hypothese formuliert. Bei der Formulierung von Hypothesen gibt es drei grundlegende Bestandteile, die bei der Aufstellung unterstützen können.

1. **Bedingung:** Die Bedingung – das „Wenn“ in der Hypothese beschreibt die geplante Veränderung, die durchgeführt werden soll. Das könnte z.B. ein geänderter Prozess oder eine abgewandelte Version einer Webseite sein. Ein konkretes Beispiel wäre: „Wenn das Suchfeld beschriftet wird,...“ (Witzenleiter, 2021, S. 25)
2. **Folge:** Die Folge – das „Dann“ in der Hypothese beschreibt das erwartete Ergebnis der Bedingung und die möglichen Veränderungen, die darauf zurückzuführen sind. Dabei ist es wichtig, dass das Ergebnis bzw. die Hypothese auf einer klaren und messbaren Annahme basieren muss. Klare messbare Metriken, dienen als objektive

Maßstäbe, um aussagekräftige Ergebnisse bei A/B-Tests zu erzielen. Wird das Beispiel von oben verwendet bedeutet das: „Wenn das Suchfeld beschriftet wird, dann wird die Suche mehr genutzt und mehr auf den „Suchbutton“ geklickt,...“ (Witzenleiter, 2021, S. 25) Eine klare messbare Metrik könnte dann die Anzahl der Suchanfragen oder die Anzahl der Klicks auf den Suchbutton sein.

3. Erklärung: Die Erklärung – das „Weil“ in der Hypothese dient als Begründung, um mögliche Verhaltensänderungen der Nutzer zu erklären. Sie kann dabei helfen, die Richtung des A/B-Tests zu lenken und die Erwartungen bezüglich der Ergebnisse zu formulieren. Eine gut durchdachte Erklärung sollte daher auf Informationen basieren, die aus vorherigen Datenanalysen, Marktforschungen oder Best Practices stammen. Eine klare Erklärung der Hypothese ermöglicht es unter anderem den Beteiligten im Unternehmen, wie zum Beispiel Teammitgliedern oder Stakeholdern, das Ziel und den Zweck eines A/B-Tests besser zu verstehen und nachzuvollziehen. Das folgende Beispiel zeigt: „Wenn das Suchfeld beschriftet wird, dann wird die Suche mehr genutzt und mehr auf den „Suchbutton“ geklickt, weil diese dadurch schneller auffindbar ist“ (Witzenleiter, 2021, S. 25)

Zur Aufstellung und Formulierung einer aussagekräftigen Hypothese ist auch die Auswahl der Testvariable von Bedeutung. Sie hat direkten Einfluss darauf welche Erkenntnisse aus dem A/B-Test gewonnen werden können. Die Auswahl der Testvariable steht eng mit dem Testziel in Verbindung, da sie einen direkt Bezug hat konkrete Verbesserungen zu erzielen. (Lapp) So kann eine Variable wie bspw. ein CTA-Button zu einer höheren CR führen. Im zuvor genannten Beispiel fungiert die Beschriftung des Suchfelds als unabhängige Variable und führt letztendlich zu mehr Klicks auf den Suchbutton.

Insgesamt lässt sich festhalten, dass die Aufstellung von Hypothesen und der Testvariable eine unverzichtbare Grundlage für erfolgreiche A/B-Tests darstellt. Durch eine klare Formulierung von Bedingung, Folge und Begründung wird eine gezielte Untersuchung von Veränderungen ermöglicht und die Erwartungen bezüglich der Ergebnisse präzisiert. Eine gut durchdachte Hypothese und die Auswahl der richtigen Testvariable erleichtern die Auswertung der Testergebnisse und ermöglichen es, aussagekräftige Erkenntnisse zu gewinnen.

3.3.2 Festlegung des Stichprobenumfangs und der Testdauer

Auch die Festlegung des Stichprobenumfangs und der Testdauer spielen eine entscheidende Rolle bei A/B-Tests. Eine zu kleine Stichprobengröße kann zu ungenauen Ergebnissen und zu keiner erreichbaren statistischen Signifikanz führen. Eine zu große Stichprobengröße wiederum kann zu unnötigen Ressourcen- und Zeitaufwand führen. Aus diesem Grund ist es von

Bedeutung, ein angemessenes Verhältnis bezüglich des Stichprobenumfangs zu finden. Hierfür können z.B. Online-Rechner verwendet werden, die es ermöglichen den idealen Stichprobenumfang zu berechnen.

Bei der Berechnung des Stichprobenumfangs müssen vier Faktoren berücksichtigt werden, um die Eingabewerte des Rechners zu verstehen:

1. Zu erwartende Steigerung der Primary KPI wie z.B. die CR: Diese Variable basiert auf eine geschätzte Wahrscheinlichkeit, mit der ein Nutzer in der jeweiligen Variante eine gewünschte Aktion ausführt. Die erwartete CR kann durch Schätzungen, die Analyse vergangener Daten oder den Vergleich mit Branchen-Benchmarks ermittelt werden.
2. Minimale Effektgröße: Eine angemessene minimale Effektgröße gewährleistet, dass der Test ausreichend empfindlich ist, um relevante Veränderungen zu erkennen, während unnötige Ressourcen und Zeit gespart werden, indem zu kleine vermieden werden.
3. Konfidenzlevel: In Bezug auf den Stichprobenumfang bedeutet das Konfidenzlevel, wie sicher man sein kann, dass die Schätzung eines Wertes in der gesamten Population enthalten ist. Wenn also das Konfidenzlevel bei 95% liegt, bedeutet dies, dass 95 von 100 Stichproben „ähnlich“ sind. Ähnlich in dem Sinne, dass Benutzer in beiden Varianten ähnliche Merkmale, Hintergründe oder demografische Eigenschaften haben sollten, um einen sicheren Vergleich der Varianten durchzuführen. Durch eine Randomisierung bei der Zuordnung von Besuchern zu den Varianten kann sichergestellt werden, dass die Stichproben ähnlich sind. Je höher das Konfidenzlevel ist, desto mehr braucht man eine größere Stichprobe.
4. Statistische Power (Teststärke): Eine hohe statistische Power bedeutet, dass der Test zuverlässig ist und größere Unterschiede zwischen den Varianten erkennen kann. Der typische Bereich liegt bei 80%

(Müller, 2022) (Looschelders, 2023, S. 77–78)

Im Folgenden soll ein anschauliches Beispiel (siehe Abb. 10 & 11) präsentiert werden, welches verdeutlicht, wie die Eingaben zu machen sind:

„Wir wollen unsere Betreffzeile des E-Mail-Newsletters optimieren, welche bisher eine durchschnittliche Öffnungsrate von 30% erzielte. Daher setzen wir für die erwartete CR 30% ein. Unsere neue Betreffzeile sollte mindestens um 10% besser sein, damit wir sie der derzeitigen vorziehen. Daher ist der MDE 10% groß. Bei unveränderter Teststärke und Konfidenzlevel kommen wir zu dem Schluss, dass wir den neuen Betreff an 3691 Kontakten testen müssen, damit unser Testergebnis statistisch signifikant wird.“ (Müller, 2022)

Conversion Rate [?]	<input type="text" value="30"/> %	Erforderliche Stichprobengröße pro Variante 3.691
Minimum Detectable Effect [?]	<input type="text" value="10"/> %	
Statistical Significance [?]	<input type="text" value="95"/> %	
Statistical Power [?]	<input type="text" value="80"/> %	

Abbildung 10: Stichprobenrechner mit Beispielwerten zur Berechnung der Stichprobengröße

(AB Tasty)

Auch die Testdauer kann mithilfe eines Online-Rechners berechnet werden. Dazu wird der durchschnittlicher Traffic pro Tag und die Anzahl der zu testenden Varianten angegeben.

Durchschnittlicher Traffic pro Tag [?]	<input type="text" value="50"/>	Mindestlaufzeit in Tagen 15
Anzahl der Varianten [?]	<input type="text" value="2"/>	

Abbildung 11: Stichprobenrechner mit Beispielwerten zur Berechnung der Testdauer

(AB Tasty)

Da heutzutage viele zahlreiche Online-Rechner zur Verfügung stehen, die den Stichprobenumfang oder die Testdauer von A/B-Tests kalkulieren, können diese Rechner je nach Anwendung oder Methodik variieren. UN sollten daher sorgfältig den passenden Rechner auswählen, die für ihre Kennzahlen und Ziele am relevantesten sind.

3.4 User Experience im E-Commerce

Die User Experience (UX) spielt eine entscheidende Rolle im E-Commerce und ist ein wichtiger Erfolgsfaktor. Sie beeinflusst das Verhalten der Nutzer und ihre Zufriedenheit, was sich wiederum auf den Erfolg eines Onlineshops auswirkt. Bereits in Kapitel 3.1.5 wurde der Begriff ausführlich erläutert, im folgenden Kapitel werden daher die Bedeutung und Einflussfaktoren auf die UX in Onlineshops untersucht.

3.4.1 Bedeutung und Einflussfaktoren auf die UX in Onlineshops

In einem Onlineshop gibt es verschiedene Einflussfaktoren, die die UX beeinflussen können. Dabei liegt der Fokus darauf, sicherzustellen, dass Nutzer einen Mehrwert bei der Nutzung erfahren. Peter Morville, ein namhafter UX-Experte, entwickelte das Konzept des "User Experience Honeycomb", das die Qualitätsaspekte der UX in Form einer Honigwabe in sieben Facetten darstellt (siehe Abb. 12). (Wesolko, 2016) Im weiteren Verlauf werden diese Facetten genauer untersucht und im Kontext des E-Commerce beschrieben.

Die sieben Facetten des User Experience Honeycomb sind:

1. Useful - Nützlich

Die Nützlichkeit im Kontext des „User Experience Honeycomb“ bezieht sich auf die Bewertung, ob das Produkt, die Dienstleistung oder Anwendung einen Nutzen für den Zielkunden bietet. Wenn dies keinen sinnvollen Nutzen für den Kunden bietet, ist es auf dem Markt zwecklos. (+sitegeist) Durch die Identifizierung der Nutzerbedürfnisse können positive Nutzererfahrungen mit einem Produkt, einer Dienstleistung oder einer Anwendung geschaffen werden. In der Produktbeschreibung eines Onlineshops könnte der Nutzen und die damit verbundenen Vorteile der Produkte hervorgehoben werden.

2. Useable - Benutzbar

Hier geht darum, dass Nutzer ihre Ziele und Aufgaben durch das Produkt effektiv und effizient erreichen können. (+sitegeist) Daher sollte die Anwendung, Funktion oder das System, in dem das Produkt oder die Dienstleistung bereitgestellt wird, so konzipiert sein, dass sie einfach verständlich und benutzerfreundlich sind. Idealerweise sollte das System oder die Anwendung so intuitiv wie möglich gestaltet werden, um die Lernkurve und die Einarbeitung für Benutzer auf ein Minimum zu reduzieren. (Wesolko, 2016) Ein Beispiel ist eine klare und intuitive

Navigation im Onlineshop, die es den Kunden ermöglicht, leicht zwischen verschiedenen Kategorien und Produkten zu navigieren.

3. Findable - Auffindbar

Der Hauptaspekt hierbei ist die Navigationsstruktur: Sind die gesuchten Informationen für die Nutzer leicht auffindbar? Der Nutzer ist dazu befähigt relevante Inhalte oder Funktionen schnell und einfach finden zu können. (Wesolko, 2016) Durch eine angemessene Strukturierung und die Implementierung geeigneter Navigations- und Suchfunktionen kann dies gewährleistet werden. Ein Beispiel im Onlineshop wäre eine effektive Suchfunktion, die relevante Ergebnisse basierend auf den eingegeben Suchbegriffen liefert.

4. Credible - Glaubwürdigkeit

Die Glaubwürdigkeit spielt in der UX eine wichtige Rolle, denn es bezieht sich darauf wie vertrauenswürdig und zuverlässig ein Produkt oder eine Anwendung wahrgenommen wird. Neben der Erfüllung seiner ursprünglichen Aufgabe ist es wichtig, dass das Produkt eine angemessene Qualität aufweist und dass die bereitgestellten Informationen präzise und zweckdienlich sind. (+sitegeist) Das Produkt sollte seinen Versprechen gerecht werden. Durch Kundenbewertungen, die Verwendung von Sicherheitszertifikaten oder Shop-Gütesiegel bspw. kann ein Unternehmen Vertrauen aufbauen.

5. Desirable - Begehrtestwert

Der 5. Punkt bezieht sich auf die Attraktivität eines Produkts. Dabei liegt der Fokus auf die ästhetischen Aspekte, die durch ein gutes Branding, Image und emotionales sowie ansprechendes Design erzeugt wird. (+sitegeist) Ein Produkt oder System mit einem ansprechenden Design wird im Allgemeinen einem anderen Produkt oder System mit ähnlichen Funktionen vorgezogen. Es erzeugt Stolz beim Benutzer und erweckt Interesse bei anderen Nutzern.

6. Accessible - Zugänglich

Eine gute UX bedeutet auch, dass eine Webseite, Anwendung oder ein System in dem Produkte oder Dienstleistungen angeboten werden, für alle Nutzer einschließlich Menschen mit Behinderungen leicht zugänglich und nutzbar sind. (Wesolko, 2016) Eine Implementierung von barrierefreien Designs und Funktionen sind wichtig, um eine optimale Nutzung für alle

Kunden zu ermöglichen. Im nachfolgenden Kapitel „Barrierefreiheit im E-Commerce“ wird das Thema ausführlicher beschrieben.

7. Valuable - Wertvoll

Sowohl für den Benutzer als auch für das Unternehmen selbst muss das Produkt, die Dienstleistung oder die Anwendung einen Mehrwert bieten. (+sitegeist) Dabei geht es darum, ob Nutzer das Gefühl haben, ob sie ihre Zeit und Ressourcen gut investiert haben und sie daraus einen Mehrwert erhalten können. Unternehmen müssen herauszufinden welche Wünsche und Interessen Nutzer verfolgen. Natürlich hängt dies auch maßgeblich von den jeweiligen Nutzern ab. Ein Beispiel im E-Commerce ist das Angebot von zusätzlichen Vorteilen wie kostenloser Versand, exklusive Rabatte, Kundenservice oder Garantie, um Kunden einen Mehrwert zu bieten.

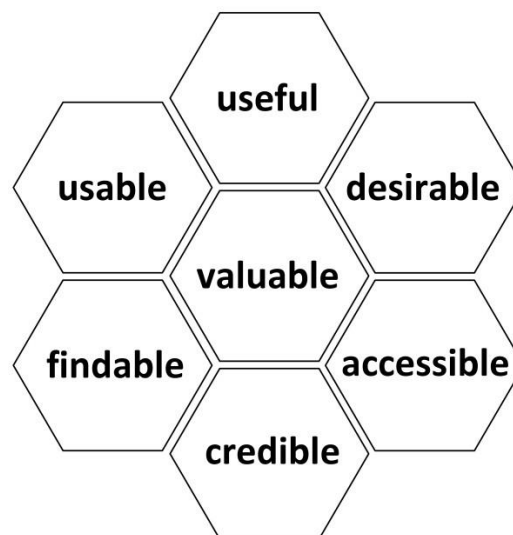


Abbildung 12: Konzept des User Experience Honeycombs von Peter Morville

(ResearchGate)

Die UX gewinnt somit zunehmend an Bedeutung und stellt einen wichtigen Erfolgsfaktor für Online-Shops im E-Commerce dar. Sie hat einen fundamentalen Einfluss auf das Verhalten der Nutzer, ihre Zufriedenheit, das Vertrauen und letztendlich ihre Bereitschaft positive Weiterempfehlungen auszusprechen. Durch eine gute UX können Onlineshops ihre Kunden langfristig binden, CR und Umsätze steigern sowie Kundenerlebnisse schaffen, die sich vom Wettbewerb abheben. Es kann dazu beitragen, dass ein Onlineshop als attraktiv, benutzerfreundlich und vertrauenswürdig wahrgenommen wird. Insgesamt spielt die UX eine maßgebliche Rolle für den Erfolg eines Onlineshops im E-Commerce.

3.4.2 Barrierefreiheit im E-Commerce

„Zum Jahresende 2021 lebten in Deutschland rund 7,8 Millionen schwerbehinderte Menschen“ (DESTATIS Statistisches Bundesamt, 2022) Demnach ist die Förderung der Gleichstellung und Teilhabe von Menschen mit Behinderungen von großer Bedeutung. Um diesen Zweck zu erreichen, wurde das "Gesetz zur Gleichstellung von Menschen mit Behinderungen (Behindertengleichstellungsgesetz - BGG)" erlassen. Dieses Gesetz hat zum Ziel, die Barrierefreiheit in verschiedenen Lebensbereichen sicherzustellen.

„Barrierefrei sind daher nach dem „Gesetz zur Gleichstellung von Menschen mit Behinderungen (Behindertengleichstellungsgesetz - BGG)“ bauliche und sonstige Anlagen, Verkehrsmittel, technische Gebrauchsgegenstände, Systeme der Informationsverarbeitung, akustische und visuelle Informationsquellen und Kommunikationseinrichtungen sowie andere gestaltete Lebensbereiche, wenn sie für Menschen mit Behinderungen in der allgemein üblichen Weise, ohne besondere Erschwernis und grundsätzlich ohne fremde Hilfe auffindbar, zugänglich und nutzbar sind. Hierbei ist die Nutzung behinderungsbedingt notwendiger Hilfsmittel zulässig.“ (Bundesministerium für Justiz)

Kurz zusammengefasst bedeutet das Prinzip der Barrierefreiheit, allen Menschen unabhängig von ihren individuellen Fähigkeiten oder Einschränkungen einen angemessenen und gleichberechtigten Zugang zu ermöglichen. Im heutigen Zeitalter der fortschreitenden Digitalisierung, die alle Bereiche des privaten und öffentlichen Lebens erfasst hat, gewinnt die Realisierung einer barrierefreien Digitalisierung zunehmend an Bedeutung. (Beauftragter der Bundesregierung für die Belange von Menschen mit Behinderungen, 2019) Daher ist es umso wichtiger digitale Angebote so zu gestalten, dass diese barrierefrei sind.

Ein Statistik (siehe Abb. 13) zufolge aus dem Jahr 2020 zeigt jedoch, dass Barrierefreiheit im Internet kaum berücksichtigt wird. „Die gemeinnützige Organisation WebAIM hat 2019 und 2020 jeweils im Februar die Majestic Million-Liste der meistreferenzierten Websites der Welt auf ihre Zugänglichkeit für Menschen mit Behinderungen untersucht und stellte dabei fest, dass nur rund zwei Prozent der darin enthaltenden Homepages keine Mängel aufweisen.“ (Boksch, 2020) Laut den WCAG-Richtlinien (Web Content Accessibility Guidelines) – ein Standard zur barrierefreien Gestaltung von digitalen Inhalten, ist das häufigste Problem von Websites der schwache Kontrast zwischen Text und Hintergrund, wodurch Menschen mit eingeschränkten Sehvermögen Schwierigkeiten haben. Auch fehlende Alternativtexte bspw. unter Bildern oder fehlerhafte Links und Schaltflächen erschweren die Zugänglichkeit und Nutzung von digitalen Inhalten für Menschen mit Behinderungen. (Boksch, 2020)

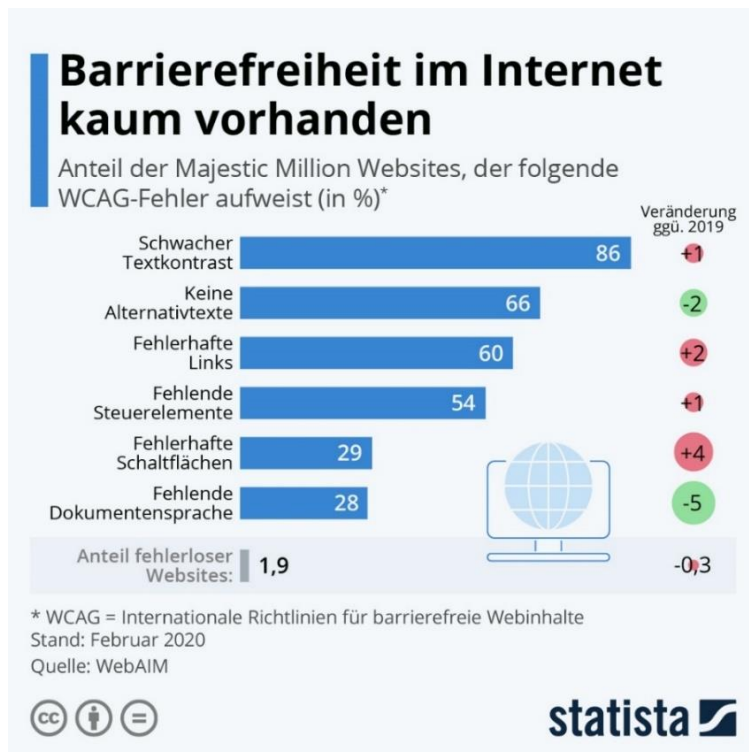


Abbildung 13: Barrierefreiheit im Internet kaum vorhanden (Boksch, 2020)

Die Abb. 14 zeigt, dass auch 3 Jahre später 96,3% der weltweit führenden Websites nachweisliche Verstöße gegen die WCAG 2 Richtlinien aufweisen. Das bedeutet, dass nur 3,7% der Websites frei von Verstößen bzw. Mängel sind. Das Diagramm „Home pages with most common WCAG failures (% of home pages)“ zeigt, dass auch die häufigsten Fehler, die bereits im Jahr 2020 gemacht wurden meist unverändert blieben. (WebAIM)

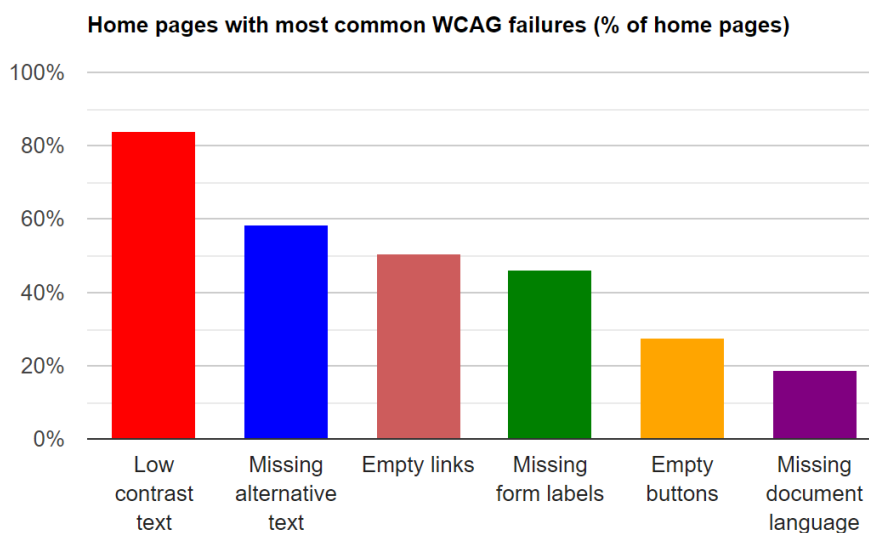


Abbildung 14: Webseiten mit den häufigsten WCAG-Fehlern 2023 (% der Webseiten) (WebAIM)

Ein Vergleich (siehe Abb. 15) zwischen den Jahren 2019 bis 2023 zeigt ebenfalls, dass es nur geringfügige Rückgänge bei der Anzahl der erkannten WCAG Fehlertypen gibt. Obwohl es in den letzten drei Jahren eine leichte Verbesserung bei bestimmten Fehlertypen wie Alternativtexten oder Dokumentensprache gegeben hat, ist der Fortschritt im Bereich der barrierefreien Digitalisierung insgesamt eher langsam oder teilweise fast unverändert. Durch eine gezielte Behebung dieser spezifischen Fehlerarten, wie von WebAIM empfohlen, könnte eine signifikante Verbesserung der Zugänglichkeit des Internets erzielt werden. (WebAIM)

Home pages with most common WCAG 2 failures

WCAG Failure Type	% of home pages in 2023	% of home pages in 2022	% of home pages in 2021	% of home pages in 2020	% of home pages in 2019
Low contrast text	83.6%	83.9%	86.4%	86.3%	85.3%
Missing alternative text for images	58.2%	55.4%	60.6%	66.0%	68.0%
Empty links	50.1%	49.7%	51.3%	59.9%	58.1%
Missing form input labels	45.9%	46.1%	54.4%	53.8%	52.8%
Empty buttons	27.5%	27.2%	26.9%	28.7%	25.0%
Missing document language	18.6%	22.3%	28.9%	28.0%	33.1%

Abbildung 15: Vergleich 2019-2023 - Webseiten mit den häufigsten WCAG-Fehlern 2023 (% der Webseiten)

(WebAIM)

Eine Behebung der verschiedenen Fehler kann sich jedoch bald ändern, denn am 28. Juni 2025 tritt das Barrierefreiheitsstärkungsgesetz (BFSG) in Kraft. Das BFSG wurde im Juli 2021 verkündet und setzt die europäische Barrierefreiheitsrichtlinie (Richtlinie (EU) 2019/882 über die Barrierefreiheitsanforderungen für Produkte und Dienstleistungen) um. Somit verpflichtet das Gesetz unter anderem Onlineshopbetreiber im E-Commerce-Sektor, ihre Produkte oder Dienstleistungen barrierefrei zu gestalten. (Bundesfachstelle Barrierefreiheit, 2021) Alle Unternehmen, mit mehr als 10 Mitarbeitern und mit einem Jahresumsatz von mehr als 2 Millionen Euro müssen die Barrierefreiheitsanforderungen ab 2025 erfüllen. (Bundesfachstelle Barrierefreiheit) Dabei stellt sich die Frage, welche wichtigen Elemente auf der grafischen Benutzeroberfläche barrierefrei sein müssen und wie diese umzusetzen sind. Die grafische Benutzeroberfläche (GUI) bezieht sich auf den visuellen Teil einer Software oder eines Systems, mit

dem Benutzer interagieren. Sie umfasst visuelle Elemente wie Schaltflächen, Menüs, Symbole und Fenster, die es Benutzern ermöglichen, Befehle zu geben, Informationen anzuzeigen und Aktionen auf einem Bildschirm auszuführen. (Juviler, 2022) Im Hinblick auf die Mensch-Computer-Interaktion (Human-Computer Interaction, HCI) ist es von großer Bedeutung, dass die grafische Benutzeroberfläche barrierefrei gestaltet wird, um eine inklusive Nutzung für alle Benutzer zu ermöglichen.

Damit die Gestaltung von Benutzeroberflächen oder Interaktionskonzepten optimal und benutzerfreundlich gewährleistet werden kann, gibt es in der Mensch-Computer-Interaktion die sogenannten Subsysteme des menschlichen Körpers, die bei der Interaktion mit Computersystemen eine Rolle spielen (siehe Abb. 16). (Vieritz, 2015, S. 16)

Perzeption -> bezeichnet den Vorgang der Wahrnehmung und Auslegung sensorischer Reize durch die Sinne, wie Sehen, Hören und Tasten. (Dr. No, Dr. Antwerpes, Frank)

1. Visuelle Subsystem: Hier geht es um die visuelle Wahrnehmung, das heißt es werden visuelle Informationen von der Benutzeroberfläche eines Systems erfasst.
2. Auditives Subsystem: Das Subsystem ermöglicht die Aufnahme von akustischen Informationen. Das sind zum Beispiel Töne, die von einem Computersystem erzeugt werden.
3. Haptisches Subsystem: Es betrifft den Tastsinn und taktilen Feedback wahrzunehmen. Durch haptisches Feedback wie bspw. einer Vibration können Benutzer eine verbesserte Interaktion des Systems erhalten.

Kognition -> bezieht sich auf Prozesse, Strukturen und Fähigkeiten des Menschen in Bezug auf Wahrnehmung, Gedanken, Erinnerungen, Lernen, Aufmerksamkeit, Problemlösung und Entscheidungsfindung (Hänsel, Baumgärtner, Kornmann & Ennigkeit, 2016, S. 23–24)

4. Kognitives Subsystem: Durch die oben genannten Aspekte wird eine effektive Informationsverarbeitung und Entscheidungsfindung ermöglicht

Operation/ Motorik -> bezieht sich auf die physischen Handlungen oder Aktionen, die ein Benutzer ausführt, um mit einem Computersystem zu interagieren

5. Motorisches Subsystem: betrifft die motorischen Fähigkeiten des Menschen, wie bspw. die Steuerung der Muskeln und Gliedmaßen, die zur Interaktion mit einem Computersystem verwendet werden. Das kann zum Beispiel die Bedienung mit einer Maus oder der Tastatur sein.

(Steinicke & Wittenburg, S. 11)

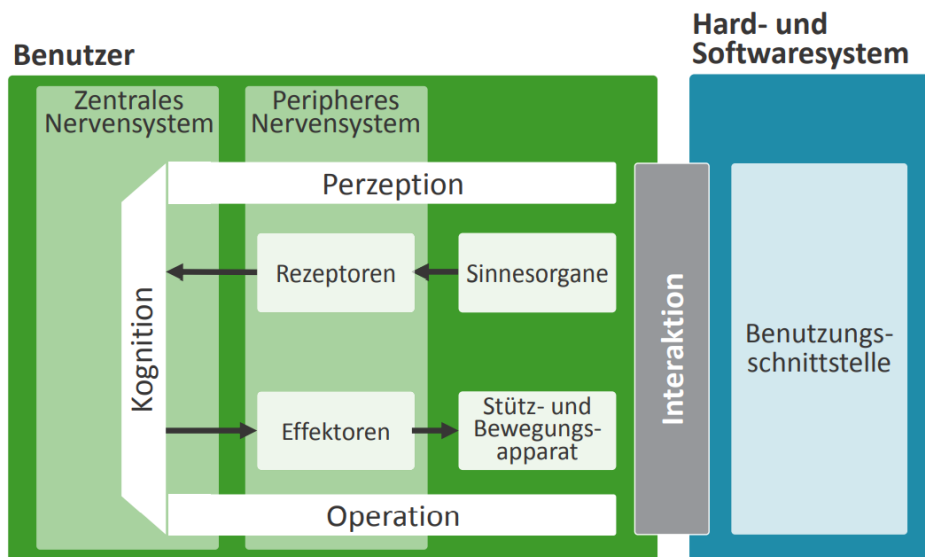


Abbildung 16: Subsysteme des menschlichen Organismus in der HCI

(Vieritz, 2015, S. 17)

Die Abb. 16 soll das Verständnis der Interaktion zwischen Benutzern und Computersystemen erleichtern. Dabei ist auch zu beachten, dass Benutzerschnittstellen so gestaltet werden sollen, dass Menschen mit beeinträchtigten oder eingeschränkten Subsystemen diese dennoch nutzen können.

Im nächsten Abschnitt werden daher unterschiedliche Elemente der grafischen Benutzeroberfläche auf ihre Barrierefreiheit hinsichtlich der verschiedenen Subsysteme erläutert. Dafür wird der BITV-Test (Barrierefreie-Informationstechnik-Verordnung-Test) in Betracht gezogen - ein Prüfprozess zur Überprüfung der Barrierefreiheit von Webseiten.

Da es aber 98 Prüfschritte gibt werden nur allgemeine Schritte aus den unterschiedlichen „Kategorien“ vorgestellt, die für Onlineshops oder Webseiten von besonderer Bedeutung sind - insbesondere die von oben aufgelisteten Fehler.

„Kontraste von Texten, Grafiken sowie grafische Bedienelemente ausreichend“ – diese Prüfschritte dienen dazu, dass Inhalte auf der Webseite für Menschen mit Sehbeeinträchtigungen gut erkennbar sind. Für die Prüfung der Kontraste werden die verwendeten Farben und Helligkeiten der Elemente mit einem Farbkontrast-Tool wie bspw. dem WCAG Color Contrast Checker analysiert.

Die gängigen Kontrastverhältnis-Vorgaben sind:

- Normaler Text ($\geq 18\text{pt}$ oder $\geq 14\text{pt}$ fettgedruckt): Ein Kontrastverhältnis von mindestens 4,5:1 zwischen Textfarbe und Hintergrund.

- Großer Text ($\geq 24\text{pt}$ oder $\geq 18\text{pt}$ fettgedruckt): Ein Kontrastverhältnis von mindestens 3:1 zwischen Textfarbe und Hintergrund. (BIK BITV Test)
- Grafiken und Bedienelemente: Ein Kontrastverhältnis von mindestens 3:1 zwischen dem Element und dem Hintergrund. (BIK BITV Test)

„Alternativtexte für Bedienelemente, Grafiken und Objekte“ – diese Prüfschritte sind wichtig für blinde Benutzer oder solche, die das Laden von Grafiken für schnellere Ladezeiten deaktivieren. Des Weiteren werden Icon Fonts als Schriftarten mit Symbolen per CSS eingebunden, jedoch werden sie von assistiven Technologien (z.B. Screenreader) möglicherweise nicht richtig dargestellt oder es fehlt der Kontext für ihre Bedeutung. Daher werden Textalternativen eingesetzt, um den Inhalt verständlich zu machen. Textalternativen werden durch das "alt"-Attribut in HTML definiert und sollten den Inhalt und die Bedeutung der Grafik vermitteln. Die Überprüfung erfolgt mit einem Tool, bspw. mithilfe des Web Developer Toolbars. Werden zudem Icon Fonts verwendet sollte man sicherstellen, dass diese mit ARIA-Labels (Accessible Rich Internet Applications) im HTML-Code angereichert werden, um ihre Bedeutung im Kontext zu vermitteln. (BIK BITV Test) (BIK BITV Test)

„Navigation schlüssig und nachvollziehbar“ - es werden verschiedene allgemeine Aspekte der Navigation untersucht, da diese mehrere Prüfschritte umfassen. Eine schlüssige Navigation ermöglicht es Benutzern, sich auf der Webseite leicht zu orientieren. Sie sollte daher konsistent, logisch und leicht nachvollziehbar sein. Das bedeutet konkret:

- Navigationselemente sollten einheitlich gestaltet sein und sich auf allen Seiten der Webseite ähnlich oder identisch verhalten. Benutzer sollten sich auf konsistente Platzierung und Funktion der Navigationselemente verlassen können, um eine vertraute und leicht verständliche Nutzererfahrung zu gewährleisten. (BIK BITV Test)
- Die Hierarchie der Navigation sollte eine klare und logische Struktur aufweisen. Hauptmenüpunkte sollten thematisch oder inhaltlich geordnet sein, um Benutzern eine intuitive Orientierung auf der Webseite zu ermöglichen. (BIK BITV Test) (BIK BITV Test)
- Tastaturnavigation: Die Navigation sollte nicht nur mit der Maus, sondern auch mit der Tastatur leicht durchführbar sein. Benutzer sollten alle Navigationspunkte problemlos mit der Tabulatortaste oder anderen Tastaturbefehlen erreichen können. Durch die Berücksichtigung wird gewährleistet, dass auch motorisch eingeschränkte Menschen oder blinde Benutzer die Webseite benutzen können. (BIK BITV Test)

„klare Struktur und Semantik“ – ein gut strukturierter und semantisch ausgezeichneter Inhalt erleichtert die Navigation und das Verständnis der Webseite erheblich. Zudem sind eine deutliche Struktur und semantische Auszeichnung besonders wichtig für Screenreader-Benutzer, da der Screenreader den Inhalt anhand seiner semantischen Struktur vorliest und

interpretiert. Eine klare Struktur ist durch eine sinnvolle und hierarchisch aufbauende HTML Struktur aufgebaut. Das bedeutet:

- Überschriften, Listen, Absätze und andere Elemente wie Listen sollten entsprechend ihrer Bedeutung und Funktion korrekt zugeordnet werden. Beispielsweise sollten Überschriften mit den HTML-Elementen h1, h2, h3, usw. ausgezeichnet werden. Durch die Einhaltung der Semantik bekommen Benutzer ein besseres Verständnis der Seite und können den Inhalt leichter erfassen. (BIK BITV Test) (BIK BITV Test)
- Tabellen sollen strukturell richtig aufgebaut sein, damit Screenreader den Inhalt einer Tabelle korrekt interpretieren und vorlesen kann. Die korrekte Verwendung von Tabellen-Markup umfasst die Verwendung der HTML-Elemente th (für Tabellen-überschriften) für die Kopfzeile der Tabelle und td (für Tabellenzellen) für den eigentlichen Inhalt, oder alternativ die entsprechende Verwendung von ARIA-Rollen. (BIK BITV Test)

Zusammenfassend lässt sich sagen, dass die digitale Barrierefreiheit in den kommenden Jahren weiter an Bedeutung gewinnen wird. Gesetzliche Vorgaben zur Barrierefreiheit werden voraussichtlich weiter ausgebaut, um eine umfassende Inklusion zu gewährleisten. Technologische Innovationen werden die Umsetzung von Barrierefreiheit erleichtern und neue Möglichkeiten für die Integration barrierefreier Funktionen bieten. Dennoch sind weiterhin Anstrengungen erforderlich, um die digitale Barrierefreiheit voranzutreiben und die Barrierefreiheitsprinzipien in allen digitalen Produkten und Dienstleistungen zu verankern.

4 Empirische Untersuchung

In Zusammenarbeit mit der Agentur DRIP Agency wurde das Thema der Bachelorarbeit formuliert und bearbeitet. DRIP Agency ist eine Agentur mit Hauptsitz in Traunstein und wurde Ende 2018 von Samuel Hess & Fabian Gmeindl gegründet. Der Fokus des UN liegt auf der datengetriebenen Conversion Optimierung von Onlineshops mittels A/B-Testing. Hauptziel ist es, UN dabei zu unterstützen, ihre Online-Präsenz zu stärken, den Website-Traffic zu erhöhen und deren Umsätze zu verbessern. (DRIP. AGENCY) Mithilfe verschiedener Testing und Statistik-Tools, u.a. dem „Analytics Toolkit“, welches auch in dieser Bachelorarbeit verwendet wurde, werden die A/B-Tests durchgeführt. Im Zusammenhang mit der Bachelorarbeit, die sich auf die 2 verschiedenen A/B-Testing Methoden Fixed-Horizon-Test und Sequential Test fokussiert, wurden die A/B-Tests in Kooperation mit der Marke SNOCKS durchgeführt. Das UN wurde im Jahr 2016 von Felix Bauer und Johannes Kliesch ins Leben gerufen. SNOCKS ist eine aufstrebende und moderne Marke, die sich auf die Produktion von hochwertigen Socken, Unterwäsche und Basic-Kleidung spezialisiert hat. Das Hauptziel von SNOCKS besteht darin, die erste Anlaufstelle für Basics und Lifestyle-Kleidung zu sein. Dabei agieren sie nicht nur als Klamottenhersteller, sondern auch als Partner für Lifestyle-Produkte. (SNOCKS)

SNOCKS und A/B-Tests

Im Jahr 2016 begann SNOCKS mit der Einführung ihres Onlineshops und der Umsetzung ihrer Multichannel-Strategie. (DRIP. AGENCY) Mit diesen neuen Ambitionen setzte das UN ehrgeizige Ziele. Durch bezahlte Werbekampagnen wurden viele Besucher auf den Online-shop aufmerksam, jedoch war der Anteil an Besuchern, welcher eine Bestellung im Shop durchführt, verhältnismäßig gering. Aufgrund dieser Situation waren dringende Optimierungsmaßnahmen im Onlineshop notwendig. In Kooperation mit der DRIP Agency sollten daher High Impact Areas (jene Bereiche oder Maßnahmen, die einen signifikanten Einfluss auf den Erfolg und die Leistung des Onlineshops haben können) identifiziert und dazu ableitend kundenzentrierte Tests entwickelt werden. (DRIP. AGENCY) Das Ziel von SNOCKS besteht daher seit heute noch für Kunden ein durchgehendes Einkaufserlebnis in allen Bereichen zu schaffen. Somit konnte mit einer ausführlichen Datenauswertung des User Flows festgestellt werden, dass 76% der User zwar auf die Produktseite gelangen aber weniger als 10% die Produkte in den Warenkorb legten. (DRIP. AGENCY) Durch Usability Tests, Umfragen oder Webanalysen über das Nutzerverhalten wurden schließlich wichtige Erkenntnisse darüber gemacht, dass Nutzer sich unsicher über die Größen der verschiedenen Produkte waren. DRIP Agency implementierte daraufhin Features wie z.B. die Integration von Größentabellen auf den Produktseiten, damit die Dimensionen der verschiedenen Produkt-kategorien besser kommuniziert und von den Website Besuchern evaluiert werden können. Zudem wurden mit

A/B-Tests Hypothesen in Bezug auf die Passgenauigkeit der Größen aufgestellt, die dazu beitragen sollen, dass die Unsicherheit der Nutzer, welche Größe gewählt werden sollte, reduziert wird. Dieser Test zeigte z.B., dass das Hinzufügen von einer Größeneinschätzung, welche auf den Bewertungen anderer Nutzer basierte, zu einem signifikant positiven Uplift in der Testvariante führte. Durch weitere A/B-Tests konnte die DRIP Agency seither über 3.1 Mio. zusätzlichen Mehrumsatz für SNOCKS generieren. (DRIP. AGENCY) Um die Nutzererfahrung stetig weiterzuentwickeln und zukunftssträchtig zu gestalten, werden auch noch heute kontinuierlich A/B-Tests im Onlineshop durchgeführt.

Im Rahmen des praktischen Teils dieser Bachelorarbeit konnten 3 A/B-Tests konzipiert, durchgeführt und analysiert werden – mit dem Ziel herauszufinden welche der 2 vorgestellten Testing-Methoden (Fixed-Horizon vs. Sequential) im direkten Vergleich unter den gegebenen Bedingungen für den Onlineshop von SNOCKS besser geeignet ist.

Im ersten Kapitel liegt der Fokus auf der Identifizierung von Potenzialen zur Optimierung des Shops. Hier werden verschiedene Aspekte der Webseite untersucht, um weitere Verbesserungsmöglichkeiten zu identifizieren und diese mit Hypothesen und A/B-Tests zu überprüfen. Im nächsten Kapitel "UI/UX Designumsetzung und Prototyping" werden die priorisierten Testideen aus dem vorherigen Kapitel in die Tat umgesetzt. Es geht darum, das User Interface und die User Experience zu gestalten und Prototypen zu erstellen, um die geplanten Optimierungen visuell zu veranschaulichen und umzusetzen. In dem darauf folgenden Kapitel "Durchführung & Auswertung der Testideen" werden die drei priorisierten Testideen in ihren einzelnen Testprozessschritten ausgewertet.

Schließlich werden im Kapitel „Interpretation der Testergebnisse/ Handlungsempfehlungen und abschließendes Fazit" die gewonnenen Erkenntnisse der Fixed-Horizon und sequenziellen Test-Ansätze zusammengeführt und mögliche Änderungen im Nutzerverhalten als Konsequenz der Optimierungen abgeleitet. Dieses Kapitel bildet den Abschluss der Arbeit und gibt wichtige Handlungsempfehlungen für die zukünftige Entwicklung des Onlineshops.

4.1 Research - Identifizierung von Optimierungspotenzialen im Shop

Die Identifizierung von Optimierungspotenzialen ist von entscheidender Bedeutung, da er als Ausgangspunkt für die Entwicklung und Durchführung von A/B-Tests dient. Im Rahmen des A/B-Testing-Prozess wurden verschiedene Seiten des Onlineshops auf Optimierungspotenziale analysiert. Dies geschah unter anderem über die Betrachtung von Heatmap Daten (z.B. wie viele Nutzer sehen bzw. scrollen noch bis zu einem bestimmten Bereich), Session Recordings (Aufzeichnung der User Journey: wie bewegen sich Nutzer durch den Shop) und Google Analytics Daten (z.B. wie viele Nutzer klicken oder hovern über ein Element, wie viele Nutzer besuchen eine bestimmte Seite).

Auf Basis der Research-Findings wurden Optimierungspotenziale festgelegt, welche das Ziel haben, die Aufmerksamkeit der Nutzer auf eine bestimmte Aktion zu lenken bzw. die Kunden zum Kauf zu motivieren. Folgende Potenziale konnten dabei festgehalten werden:

1. Platzierung einer Announcement Bar (Banner) auf der Landing Page -> Im Research Prozess konnte erkannt werden, dass ein Newsletter Pop Up nur wenige Sekunden nach dem Landen auf der Seite auf dem Bildschirm erscheint - allerdings die User Journey der Website-Besucher stört, weil die Nutzer das Pop Up sofort schließen, ohne mit diesem zu interagieren. Dies lässt vermuten, dass die im Pop-Up beschriebene Vorteilskommunikation (20% Rabatt auf Newsletteranmeldung) nicht wahrgenommen werden kann. In einem zweiten Schritt wurde mittels Heatmaps untersucht, wie viele Nutzer die Kommunikation bzgl. des bevorstehenden App Launches im Footer wahrnehmen. Die Zahl der Nutzer betrug weniger als 5%, was dementsprechend nur einen sehr kleinen Anteil ausmacht. Da die Rate der Newsletter Anmeldungen und die Aufmerksamkeit auf den bevorstehenden App Launch erhöht werden sollte, fiel die Entscheidung für das erste Optimierungspotenzial auf die Anzeige eines Announcement Bar Banners. Ein Banner ist ein grafisches Element, das auf einer Webseite oder in einer mobilen App platziert wird, um bestimmte Informationen oder Werbebotschaften zu präsentieren. (Xovi) Durch die Platzierung des Banners im Header soll die Aufmerksamkeit stärker auf die Newsletter Anmeldung und auf den App Launch gelenkt werden - mit dem Ziel, Newsletter Anmeldungen und App Downloads zu erhöhen. Die zu untersuchenden Metriken umfassen daher die Anzahl der Newsletter-Anmeldungen, die Download-Raten der neuen App sowie die allgemeine Klickrate auf das Banner. Durch die Analyse dieser Metriken wird festgestellt, ob der Banner im Header eine signifikante Auswirkung auf das Kundenverhalten hat und ob es die gewünschten

Aktionen, wie die Anmeldung für den Newsletter oder den Download der neuen App, fördert.

Die aufgestellte Hypothesen lauten daher:

WENN eine Announcement Bar mit der Möglichkeit, sich für den Newsletter anzumelden, an oberster Stelle auf der Landing Page platziert wird,

DANN werden die Klickzahlen sowie die Anzahl der Newsletter-Anmeldungen signifikant steigen,

WEIL Besucher durch ein attraktives Angebot motiviert werden, sich für den Newsletter zu registrieren.

WENN eine Announcement Bar mit einem Hinweis zur neuen SNOCKS App an oberster Stelle der Landing Page platziert wird,

DANN wird die Downloadrate der App signifikant steigen,

WEIL die sichtbare Promotion die Neugier der Besucher weckt und zudem durch einen attraktiven Anreiz motiviert werden.

2. Erstellung einer „Vorteilsbox“ oder „Registrierungsangebot“ -> Über die Backend-Daten von SNOCKS konnte herausgefunden werden, dass weniger als 10% der Kunden ein Kundenkonto besitzen. Um die Einkaufserfahrung der Nutzer in Zukunft besser zu personalisieren und dadurch eine engere Kundenbindung zu schaffen, wurde das zweite Optimierungspotenzial festgehalten: die Erstellung einer sog. Vorteilsbox oder eines Registrierungsangebots. In dieser „Box“ werden die verschiedenen Vorteile und Anreize dargestellt, die Kunden dazu ermutigen sollen, ein Kundenkonto zu erstellen, wie z.B. exklusive Rabatte, Zugang zu Sonderaktionen, schnellere Bestellabwicklung, Bestellverlauf einsehen, Newsletter-Anmeldung, Belohnungsprogramme und mehr. Durch die Registrierung der Nutzer, profitiert der Onlineshop auf verschiedene Weisen: Durch die Anmeldung der Nutzer erhält SNOCKS wertvolle Informationen wie Demografische Daten (Alter, Geschlecht, etc.) oder Präferenzen des Kunden. So kann SNOCKS eine engere Beziehung zu den Kunden aufbauen und sie gezielt ansprechen, z.B. durch personalisierte Empfehlungen oder Angebote. Das trägt wiederum zu einem verbesserten Einkaufserlebnis und stärkeren Kundenbindung bei. Die im Kundenkonto gespeicherten Daten ermöglichen zudem auch eine umfassende Datenanalyse, um das Kundenverhalten besser zu verstehen und Trends zu identifizieren. Zu den untersuchenden Metriken gehören die tatsächliche Anzahl der Registrierungen, das Nutzerengagement wie beispielsweise der Klick auf den Registrierungs Button. Durch die

Analyse dieser Metriken kann der Onlineshop feststellen, wie erfolgreich die Vorteilsbox bei der Registrierung neuer Nutzer ist und wie effektiv sie bei der Förderung der gewünschten Aktionen ist.

Die Hypothese lautet hier:

WENN ein attraktives "Registrierungsangebot" direkt neben dem Registrierungsfeld platziert wird,

DANN wird die Anzahl der Neuregistrierungen signifikant zunehmen,

WEIL potenzielle Nutzer durch unmittelbare Anreize oder Vorteile motiviert werden, sich anzumelden.

3. „schnell vergriffen“ Anzeige im Produkt -> In einem Research Gespräch mit Mitarbeitern der DRIP Agency wurde darauf hingewiesen, dass regelmäßig Beschwerden beim Kundensupport von SNOCKS eingehen, dass beliebte Farben nach einem angekündigten Re-Stock sehr schnell ausverkauft sind - aber das den Nutzern im Shop aktuell nicht kommuniziert wird. Daraufhin konnte das dritte Optimierungspotenzial ausgearbeitet werden: eine "Schnell vergriffen" Anzeige bei entsprechenden Produkten zu integrieren. Im Marketing wird die Anzeige auch oft als „Flag“ bezeichnet. Sie ist eine visuelle Markierung oder Kennzeichnung, die auf der Produktseite angezeigt wird, um den Kunden auf eine bestimmte Sache hinzuweisen, wie bspw. die begrenzte Verfügbarkeit. Ziel ist es, die Kunden auf die Knappheit des Produkts aufmerksam zu machen und sie zu einer Kaufentscheidung zu bewegen, bevor das Produkt ausverkauft ist. (Versa commerce, 2023a). Eine wichtige Metrik, die verwendet werden kann ist die CR. Sie gibt an, wie viele Besucher, die die Flag gesehen haben, tatsächlich eine Kaufentscheidung getroffen und das Produkt gekauft haben. Eine hohe CR zeigt an, dass die Flag erfolgreich dazu beiträgt, Kunden zum Kauf zu bewegen.

Die Hypothese lautet:

WENN bei einem Produkt ein Flag mit der Kennzeichnung „schnell vergriffen“ platziert wird,

DANN wird die Klickrate auf dieses Produkt sowie den durchschnittlichen Umsatz pro Besucher steigen

WEIL Kunden durch diese Kennzeichnung ein erhöhtes Interesse zeigen, das Produkt näher zu betrachten oder zu erwerben.

4. Feature „Hover-Image“ zum Vergleich der Produkte -> Beim Analysieren von Aufzeichnungen der User Sessions und Heatmaps der Produktdetailseiten konnte festgestellt

werden, dass Desktop Besucher eine sehr hohe Interaktion mit den verschiedenen Produktfarben aufweisen, sehr häufig zwischen den Farben hin und her wechseln und 2-3 Farben nebeneinander in separaten Tabs öffnen - vermutlich um diese dadurch miteinander zu vergleichen. Aus dieser Beobachtung heraus entstand das vierte Optimierungspotenzial: die Einführung eines "Hover-Image" Features zum besseren Vergleich der Produktfarben. Die Funktion „Hover-Image“ bezieht sich auf die Möglichkeit, mit dem Mauszeiger über ein Produkt zu fahren und ein Bild anzuzeigen, um das Produkt mit einer anderen Farbe zu vergleichen. Dem Kunden wird so ermöglicht, das Produkt genauer zu betrachten oder verschiedene Farboptionen zu vergleichen, ohne die aktuelle Produktseite verlassen zu müssen. Durch einfaches Überfahren des Produkts mit dem Mauszeiger kann sich der Kunde ein besseres Bild von den verschiedenen Farben oder Varianten machen und so eine fundiertere Kaufentscheidung treffen. Das Feature hilft dem Kunden das passende Produkt auszuwählen und ein einfacheres Einkaufserlebnis anzubieten. Durch sogenannte Mouseover Events kann gemessen werden, wie oft der Mauszeiger über das Produkt bewegt wurde, um das Hover-Image anzeigen zu lassen. Dies gibt Aufschluss darüber, wie oft Kunden das Feature für weitere Produktansichten nutzen.

Die Hypothese lautet:

WENN das Feature „Hover-Image“ eingeführt wird, um Produktvarianten durch Überfahren mit der Maus näher zu betrachten

DANN steigt die Interaktionsrate und der durchschnittliche Umsatz pro Besucher

WEIL Kunden durch diese intuitive Funktion eine fundiertere Kaufentscheidung treffen können, ohne die Produktseite verlassen zu müssen.

5. Bedienungshilfe: Text- und Anzeigeeinstellungen -> Wie im obigen Kapitel bereits beschrieben, müssen alle Online Shops bis 2025 bestimmte Accessibility Features umgesetzt haben. Im diesem Zuge wurde die Website von SNOCKS auf Bedienungshilfen analysiert. Dabei ist aufgefallen, dass es bisher keine Elemente im Shop gibt, welche beeinträchtigte Menschen dabei unterstützen, durch den Shop zu navigieren. Die Integration einer Bedienungshilfe stellte daher das vierte Optimierungspotenzial auf. Die Bedienungshilfe ist dazu da, die Benutzerfreundlichkeit und Zugänglichkeit des Onlineshops für verschiedene Benutzergruppe zu verbessern. Mit dieser Funktion können Kunden die Darstellung von Texten und Anzeigen auf den SNOCKS Seiten individuell nach ihren Vorlieben anpassen. Durch das Feature können Menschen mit einer Sehbehinderung oder eingeschränktem Sehvermögen die Schriftgröße sowie

den Kontrast anpassen, um Texte besser lesen zu können. Die Möglichkeit, die Text- und Anzeigeeinstellungen anzupassen, trägt dazu bei, dass sich alle Kundinnen und Kunden unabhängig von ihren individuellen Bedürfnissen und Vorlieben willkommen und berücksichtigt fühlen. Durch das Tracken von Nutzerinteraktionen in Google Analytics kann zudem herausgefunden werden, wie häufig die Bedienungshilfe verwendet wird. Durch diese Analyse kann festgestellt werden, ob Kunden das Feature nutzen, und ob dies Auswirkungen auf ihr Verhalten auf der Website hat.

Es wird daher folgende Hypothese aufgestellt:

WENN die Bedienungshilfe mit individuellen Text- und Anzeigeeinstellungen implementiert wird, um die Benutzerfreundlichkeit für Menschen mit Seheinschränkungen zu erhöhen,

DANN wird die Verweildauer und Interaktion von dieser Zielgruppe auf der SNOCKS Webseite signifikant steigen,

WEIL sie sich durch die angepassten Darstellungsoptionen besser zurechtfinden.

Insgesamt wurden 6 Hypothesen aufgestellt, anhand derer die festgestellten Optimierungspotenziale validieren werden können. Dadurch können gezielte Maßnahmen ergriffen werden, um bestimmte Aspekte des Onlineshops zu verbessern. Diese dienen dann als Grundlage für die Durchführung der A/B-Tests. In Anbetracht des Zeitaufwands und des erwarteten Nutzens wurden schließlich drei Hypothesen aus den identifizierten Optimierungspotentialen priorisiert, da von diesen erwartet wird, dass sie den größten Einfluss auf das Nutzerverhalten und dadurch auch eine Steigerung des Umsatzes haben:

- Feature „Hover-Image“ zum Vergleich der Produkte
- Platzierung eines Announcement Bar (Banner) auf der Landing Page
- Bedienungshilfe: Text- und Anzeigeeinstellungen

Da die Implementierung eines umfassenden Bedienungshilfen-Elements eine hohe Komplexität birgt und den vorgegebenen, zeitlichen Rahmen dieser Bachelorarbeit verfehlt hätte, wurde die Entscheidung getroffen, das Bedienungshilfe-Element über einen Smoke Test durchzuführen. Ein Smoke Test ist eine Testmethode, um zu überprüfen, ob eine Funktion oder ein Feature funktioniert oder genutzt wird, noch bevor es weiter entwickelt oder technisch umgesetzt wird. Auf diese Weise können hohe Kosten und weitere Mühen gesenkt werden und gleichzeitig die Benutzerfreundlichkeit sowie Nutzererfahrung getestet werden. (10X studio) Mit dem Smoke Test kann herausgefunden werden, ob das geplante Feature

erfolgsversprechend ist und das Interesse der Kunden weckt, bevor weitere Ressourcen in die eigentliche Entwicklung investiert werden.

Die Hypothesenprüfung der anderen beiden priorisierten Optimierungspotenziale erfolgt über normale A/B-Tests.

Gründe dafür warum die anderen zwei Testideen (2 & 3) nicht bei SNOCKS umgesetzt wurden, waren folgende: Optimierungspotenzial 2 - die Erstellung einer Vorteilsbox bei der Kunden-Registrierung - lässt sich zwar schnell und einfach implementieren, allerdings wurde kurz nach Beginn dieser Bachelorarbeit ein Tool eines Drittanbieters erworben, welches mehr Funktionalitäten und Personalisierungen abdeckt, als im Optimierungspotenzial initial angedacht. Da der Einbau des Tool bereits in Gange war, stellte die Durchführung eines A/B- Tests, der selbst bei positivem Ausgang nicht hätte implementiert werden können (weil bereits ein Tool eingebaut wurde) keine gute wirtschaftliche Entscheidung dar.

Das dritte Optimierungspotenzial wurde geringer priorisiert und nicht umgesetzt, da ein bevorstehendes Sales Event mit hoher Wahrscheinlichkeit zu einem Ausverkauf vieler Produkte geführt hätte. Das Risiko, dass Nutzer eine Beschwerde einreichen, weil sie reduzierte Angebote nicht mehr erhalten, da der Abverkauf der Artikel weiter durch die Anzeige eines "Schnell Vergriffen"-Badges gefördert wird, war sowohl SNOCKS als auch der DRIP Agency zu riskant.

4.2 UI/ UX Designumsetzung und Prototyping der priorisierten Testideen

Testidee 1: Feature „Hover-Image“ zum Vergleich der Produkte

Das Design (siehe Abb. 17), das in diesem Kontext beschrieben wird, zeichnet sich durch seine Nützlichkeit sowie Einfachheit aus. Ein effektives Design sollte nicht nur ästhetisch ansprechend sein, sondern auch die Bedürfnisse des Benutzers berücksichtigen. In diesem Fall wurde das Design so gestaltet, dass es für den Nutzer intuitiv ist, was bedeutet, dass es leicht verständlich ist. Durch die Implementierung einer „Vergleichsbox“, die neben dem aktuell betrachteten Produkt erscheint, wird den Kunden durch einfachen Überfahren der Produkte („Hovern“) ermöglicht, einen direkten Vergleich anzuschauen. Die Verwendung einer Box mit einem weißen Rand und einem Pfeil trägt dazu bei, dass der Benutzer das Produkt schnell erfassen und verarbeiten kann. Aus technischer Sicht ist es wichtig, dass die Box reaktionsfähig in Echtzeit auf die Aktionen des Benutzers reagiert. Die bewusste Gestaltung der Box in einer reduzierten Größe steht zu einem ausgewogenen Verhältnis zur Gesamtgestaltung, um eine Überladung der Seite zu vermeiden und gleichzeitig die Lesbarkeit der Informationen zum Produkt zu gewährleisten. Somit kann der Benutzer schnell und effizient entscheiden,

welche Farbe ihm besser gefällt. Wenn Nutzer über eine Farbe hovern, wird zudem auch der Farbname bei „Farbe“ angezeigt – das ist auch sehr hilfreich für die Evaluierung des Produkts.

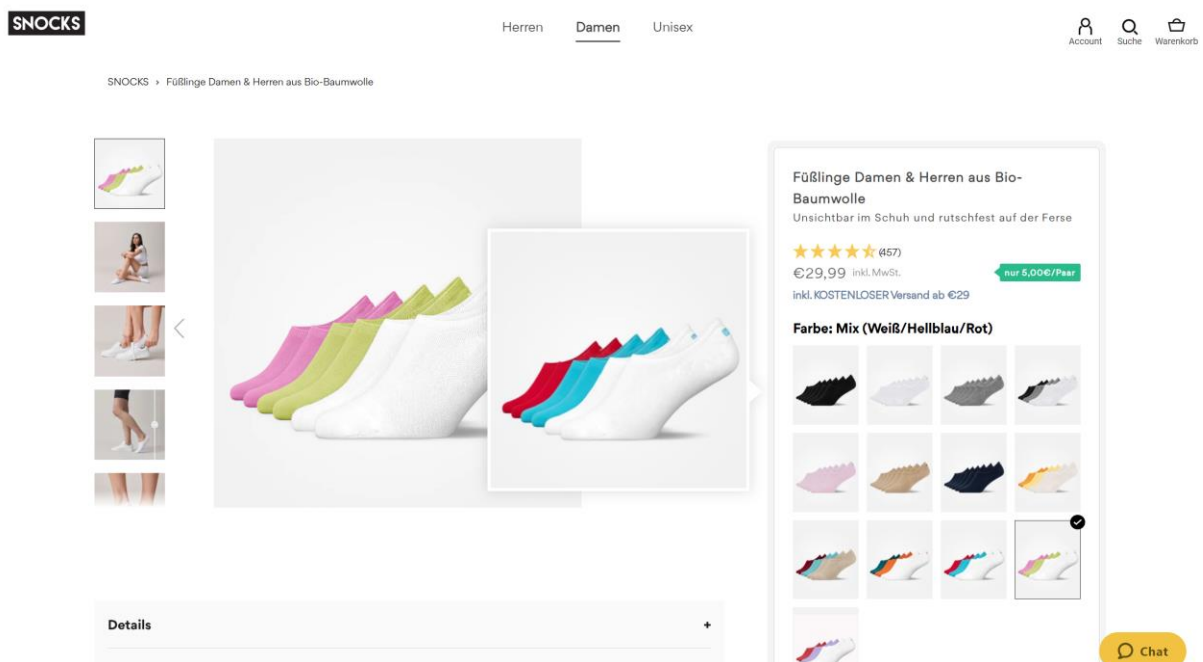


Abbildung 17: Feature "Hover-Image" zum Vergleich der SNOCKS Produkte

Testidee 2: Platzierung einer Announcement Bar (Banner) auf der Landing Page

In der Initiierungsphase des Designprozesses spielt die Schaffung einer effektiven Nutzerkommunikation eine zentrale Rolle, um den Nutzer zur Durchführung einer gewünschten Aktion oder Handlung zu animieren. Dieses Kommunikationsmittel wird als "Call-to-Action" (CTA) bezeichnet. Das Ziel besteht darin, den Nutzer dazu zu motivieren, eine spezifische Handlung auszuführen, wie z.B. in diesem Fall das Abonnieren des Newsletters oder das Herunterladen der neuen SNOCKS App. Aus diesem Grund wurde jeweils ein CTA-Button im Banner hinzugefügt, die den Nutzer zur Anmeldung für den Newsletter oder zum App Store weiterleiten. Um den Nutzer zur Durchführung einer bestimmten Aktion zu motivieren, erfordert es eine gezielte Bereitstellung von Vorteilen und Anreizen. Diese Anreize wurden in Form von CTA Sätzen übermittelt. Aus einer Auswahl von jeweils vier verschiedenen Call-to-Action-Sätzen hat sich SNOCKS für den jeweils fett hervorgehobenen Satz entschieden.

Newsletter Anmeldung:

- zum Newsletter anmelden und bis zu 30% Rabatt auf den ersten Einkauf sichern
- **Bis zu 30% auf deine erste Bestellung sichern**
- Bis zu 30% Willkommensrabatt: Melde dich zum Newsletter an und spare bei deinem ersten Einkauf!

- Newsletter-Anmeldung lohnt sich: bis zu 30% Rabatt auf den ersten Einkauf!

SNOCKS App Download:

- Entdecke unsere neue App – Jetzt kostenlos herunterladen und 15% Rabatt sichern!
- Snocks ist jetzt mobil - Hole dir unsere neue App und sichere dir 15% Rabatt!
- Jetzt verfügbar: Unsere neue & innovative Snocks App – 15% Rabatt sichern
- **SNOCKS APP 15% Rabatt auf die erste Bestellung**

SNOCKS entschied sich für die beiden kurzen Varianten (fett markiert), da sie leicht zu verstehen sind und die gewünschte Handlung klar auf den Punkt bringt. Mit wenigen Worten kann somit eine starke und wirkungsvolle Botschaft vermittelt werden. In der heutigen schnelllebigen digitalen Welt stehen Nutzern nur wenige Augenblicke zur Verfügung, um ihre Entscheidungen zu treffen. Daher ist es wichtig, den CTA-Satz kurz und prägnant zu halten, um ihre Aufmerksamkeitsspanne hoch zu halten und sie zur gewünschten Handlung zu motivieren. Da SNOCKS über eine definierte Auswahl an Farben verfügt, die in der visuellen Kommunikation verwendet werden sollen, wurden die Banner (siehe Abb. 18-21) mit einem Grauton (#f1f1f1) gestaltet, der im vorgesehenen Styleguide enthalten ist. Durch die Verwendung dieses spezifischen Grautons wird sichergestellt, dass die Banner das gewünschte professionelle und einheitliche Erscheinungsbild der Marke widerspiegeln und eine klare Verbindung zur SNOCKS-Identität hergestellt wird. Zudem sollen die Banner nicht aufdringlich wirken und vom Nutzer weggeklickt werden können, wenn er nicht daran interessiert ist. Für jeden Banner wurden 2 Versionen erstellt – jeweils einen für den Desktop und einen für mobile Geräte.

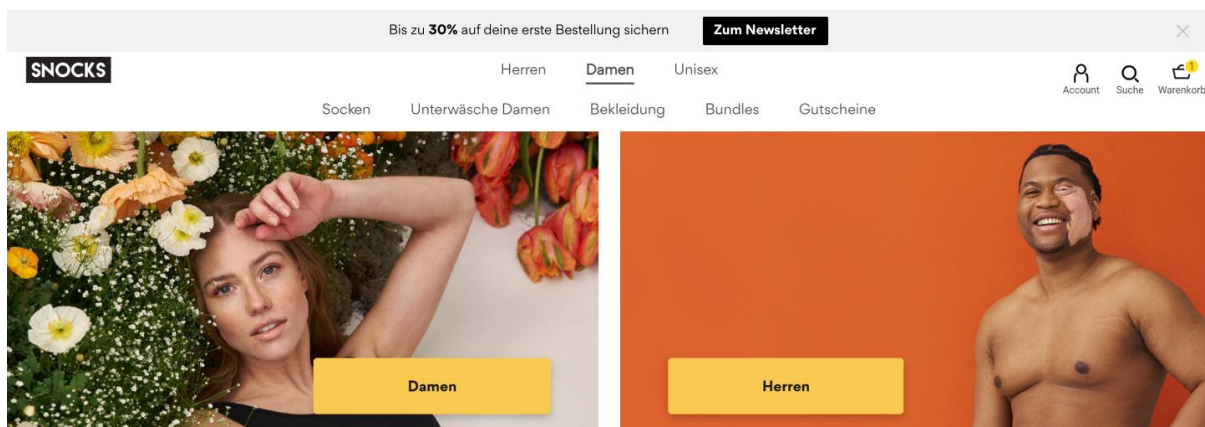


Abbildung 18: Platzierung des Newsletter Banners auf der Startseite (Desktop Version)

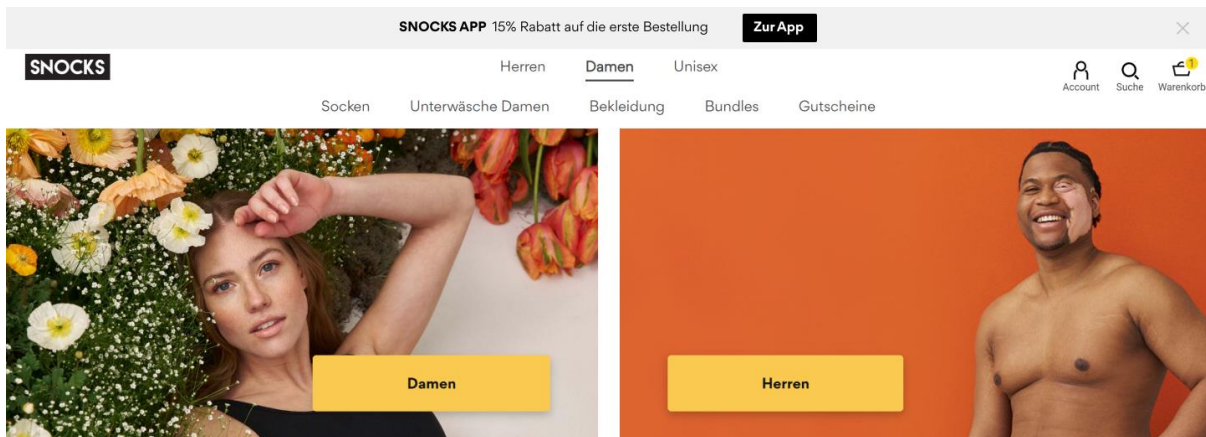


Abbildung 19: Platzierung des SNOCKS App Banners auf der Startseite (Desktop Version)

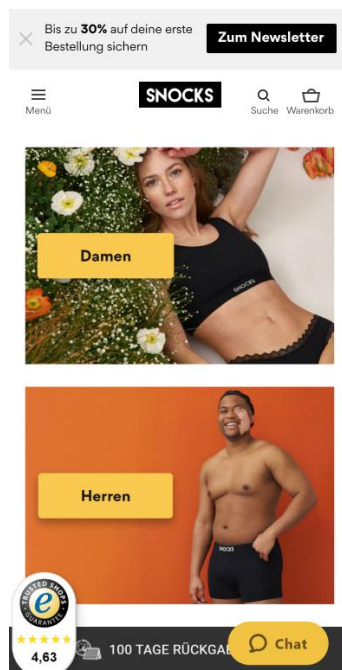


Abbildung 20: Platzierung des Newsletter Banners auf der Startseite (Mobile Version)

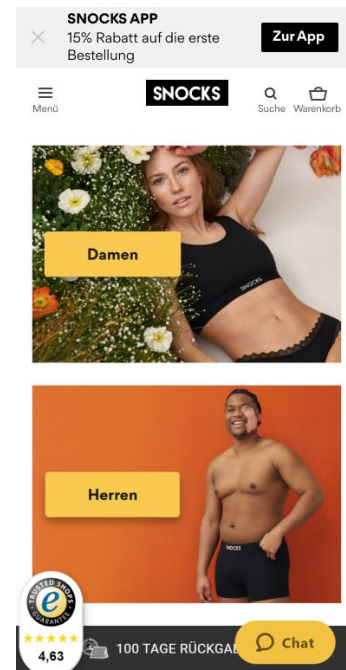


Abbildung 21: Platzierung des SNOCKS App Banners auf der Startseite (Mobile Version)

Testidee 3: Smoke Test Bedienungshilfe: Text- und Anzeigeeinstellungen

Beim Design (siehe Abb. 22-24) der Bedienungshilfe gab es mehrere wichtige Faktoren, die zu beachten waren. Dabei ist es von Bedeutung eine zugängliche und benutzerfreundliche User Experience zu gewährleisten. Durch eine intuitive Bedienungshilfe soll der Nutzer die Text- und Anzeigeeinstellungen leicht verstehen können. Da die Größen S, M, L und XL in Onlineshops im Zusammenhang mit dem Kauf von Kleidung verwendet werden, könnten sie vielen Nutzern vertraut sein. Daher soll die Verwendung von diesen Kurzbezeichnungen die Benutzeroberfläche vereinfachen und dem Benutzer die Auswahl erleichtern. Jedoch sollte

man im Hintergrund behalten, dass diese Bezeichnungen auch relativ vage sein könnten und je nach Benutzer unterschiedlich interpretiert werden können. Für den Smoke Test war es jedoch erst einmal wichtig, insbesondere die unterschiedlichen Textgrößen ohne Verlust an Funktionalität oder Design zu ändern. Für die Anzeigeeinstellungen wurden 3 verschiedene Kontrastmodi gewählt – der Hell-Dunkel-Kontrast, der Dunkel-Hell-Kontrast und der „Farbwähler“. Durch die Bereitstellung der Modi trägt es zur Barrierefreiheit und Benutzer-freundlichkeit bei, indem es ein breiteres Spektrum von Benutzerbedürfnissen und -präferenzen berücksichtigt. Die Farben wurden weitgehend nach dem SNOCKS Styleguide umgesetzt, um eine konsistente und einheitliche Markenidentität widerzuspiegeln. Auch die Schriftart *Gordita* ist die üblich verwendende Typographie bei SNOCKS.



Abbildung 22: Smoke Test Bedienungshilfe: Text- und Anzeigeeinstellungen - Icon



Abbildung 23: Smoke Test Bedienungshilfe: Text- und Anzeigeeinstellungen - Icon aufgeklappt Light Version



Abbildung 24: Smoke Test Bedienungshilfe: Text- und Anzeigeeinstellungen - Icon aufgeklappt Dark Version

Da sich die drei Kontrastmodi nur in ihren spezifischen Kontrastverhältnissen unterscheiden, wird der Hell-Dunkel-Kontrast im Hauptteil der Analyse ausführlich behandelt. Die beiden anderen Kontrastmodi, die ebenfalls relevant, aber im Wesentlichen identisch sind, werden im Anhang beigefügt, um eine vollständige und umfassende Darstellung zu gewährleisten.

Der Hell-Dunkel-Kontrastmodus zeichnet sich durch einen weißen Hintergrund oder eine andere helle Farbe aus, kombiniert mit einer dunklen Farbe wie Schwarz oder Dunkelgrau für Texte und Symbole, um einen hohen Kontrast zum hellen Hintergrund zu bieten. Um einen optimalen Kontrast zu erzielen wurden weiße Schriftzüge oder Bedienelemente wie Buttons mit einer schwarzen Umrandung versehen.

S (Small / Klein): Dies ist die Standard Textgröße und kann für Benutzer mit gutem Sehvermögen geeignet sein, aber möglicherweise schwierig für Personen mit Sehbehinderungen zu lesen sein. Die Schriftgröße sowohl in der Navigation als auch in den Bedienelementen wie Buttons wurden auf 16px. festgelegt. (siehe Abb. 25)



Abbildung 25: Hell-Dunkel Kontrastmodus mit Textgröße S

M (Medium / Mittel): Eine mittlere Textgröße für Benutzer, die ein komfortableres Leseerlebnis bevorzugen oder leichte Sehschwächen haben. Die Schriftgröße sowohl in der Navigation als auch in den Buttons wurden um 4px. auf 20px. vergrößert. (siehe Abb. 26)



Abbildung 26: Hell-Dunkel Kontrastmodus mit Textgröße M

L (Large / Groß): Eine größere Textgröße, die für Benutzer mit Seheinschränkungen hilfreich sein können. Durch die Anpassung der Schriftgröße auf 24px. wird darauf abgezielt, die Lesbarkeit deutlich zu erhöhen. (siehe Abb. 27)



Abbildung 27: Hell-Dunkel Kontrastmodus mit Textgröße L

XL (Extra Large / Extra Groß): Die größte verfügbare Textgröße, die speziell für Benutzer mit deutlichen Seheinschränkungen entwickelt wurde, um die Lesbarkeit zu maximieren. In der Benutzeroberfläche wurde eine Schriftgröße von 28 Pixeln sowohl für die Navigationsleiste als auch für die Buttons festgelegt. (siehe Abb. 28)



Abbildung 28: Hell-Dunkel Kontrastmodus mit Textgröße XL

Nach dem Aufstellen, Priorisieren und Designen der priorisierten 3 Optimierungspotenziale konnte im nächsten Schritt mit der Durchführung der A/B-Tests gestartet werden.

4.3 Durchführung & Auswertung der Testidee 1: Feature „Hover-Image“ zum Vergleich der Produkte“

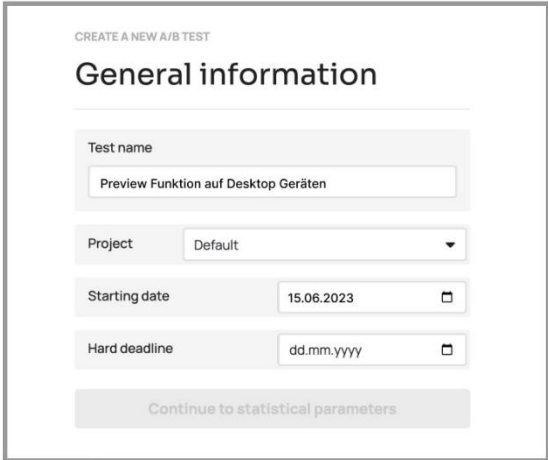
Die folgende Analyse stellt die Durchführung und Auswertung der Testidee 2 dar, die sich auf die Einführung der Preview Funktion „Hover-Image“ zur besseren Vergleichbarkeit der Produkte konzentriert. Schritte wie die Auswahl der Art des Monitorings, die Festlegung der statistischen Parameter und die detaillierte Planung der Tests sind Teil der Evaluierung. Im Rahmen dieser Arbeit erfolgt daher eine strukturierte Aufbereitung der einzelnen Testprozessschritten mit den jeweiligen Test-Methoden Fixed-Horizon und Sequential A/B-Testing.

Es werden folgende Schritte beim Sequential A/B-Testing betrachtet:

Zur Testplanung und Anlegen eines neuen A/B-Tests bedarf es zuerst an allgemeine Informationen und Parametern, die im Vorfeld definiert werden müssen.

Schritt 1: Test Basics

Der Test trägt den Namen "Preview Funktion auf Desktop Geräten" und wurde am 15.06.2023 gestartet. Beim Sequential Ansatz wird für diesen Test kein fester Endzeitpunkt oder eine sogenannte "Hard Deadline" festgelegt. (sich Abb. 29)



The image shows a web form titled "CREATE A NEW A/B TEST" with a sub-heading "General information". The form contains the following fields: "Test name" with the value "Preview Funktion auf Desktop Geräten"; "Project" with a dropdown menu set to "Default"; "Starting date" with the value "15.06.2023" and a calendar icon; "Hard deadline" with the value "dd.mm.yyyy" and a calendar icon. At the bottom of the form is a button labeled "Continue to statistical parameters".

Abbildung 29: Sequential Testing - Testidee 1: Allgemeine Informationen

Schritt 2: statistische Parameter

Beruhend auf diesen Test wurde der "Superiority"-Ansatz als Testtyp gewählt. Dies bedeutet, dass die getestete Variante einen signifikant positiven Effekt im Vergleich zur Kontrollgruppe aufweisen muss. So hat es das Ziel bspw. die Konversionsraten, Klickraten oder die Verweildauer zu verbessern. Daneben gibt es auch den Non-Inferiority-Ansatz, der darauf abzielt, dass die neue Variante bis zu einem bestimmten Prozentansatz schlechter sein darf, als die

Kontrollvariante. Superiority ist jedoch der gebräuchlichste Ansatz, insbesondere wenn das Ziel darin besteht, eine neue Funktion oder Änderung zu validieren, die voraussichtlich zu besseren Ergebnissen führen wird. Im Kapitel 3.3.2 wurde die initiale Fragestellung zur Stichprobenplanung in den Vordergrund gestellt, nämlich welchen Effekt man zu erzielen anstrebt. Allerdings verfolgt die DRIP Agency mit dem Testingtool Analytics Toolkit einen alternativen Ansatz. Hierbei wird primär evaluiert, auf welcher Seite der Test durchgeführt wird und wie hoch das Nutzeraufkommen auf dieser Seite ist. Anschließend wird der durchschnittliche Umsatz pro Nutzer ermittelt und die zugehörige Standardabweichung berechnet. Basierend auf diesen Daten wird in der Regel ein Signifikanz- und Powerlevel von 80% angewendet. Die zentrale Fragestellung in diesem Kontext lautet daher: Welche Testlaufzeit oder wie viele Nutzer sind erforderlich, um einen spezifizierten Effekt zuverlässig messen zu können? Als primäre Metrik zur Bewertung des Tests wird deshalb der durchschnittliche Umsatz pro Besucher, auch bekannt als ARPU, herangezogen. Die DRIP Agency präferiert die ARPU, da diese Metrik die monetäre Leistung jedes einzelnen Nutzers unabhängig von der Anzahl seiner Sessions

erfasst und somit aussagekräftiger ist. Aus den Daten/Reports von Google Analytics wird erwartet, dass wöchentlich etwa 164.000 Nutzer die Produktdetailseite für Desktop-Geräte besuchen. Der Baseline Average, also der durchschnittliche Umsatz pro Besucher, liegt bei 3.30 €. Zudem beträgt die Standardabweichung für diesen Wert 16.5 (siehe Abb. 30).

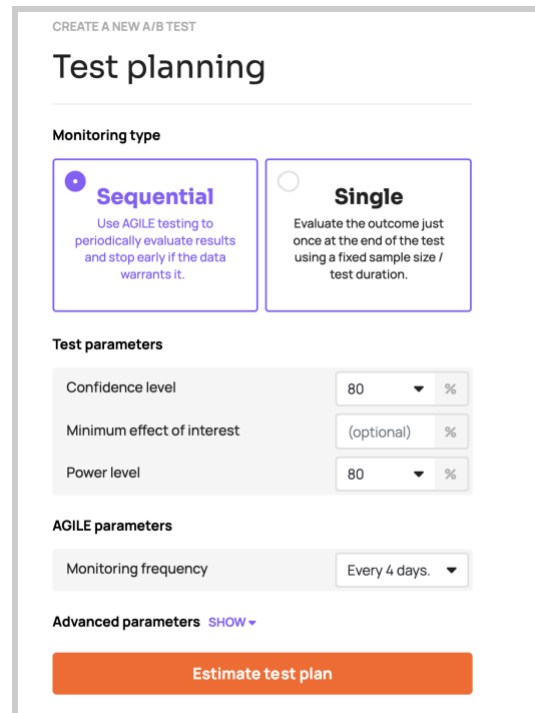
The screenshot shows the 'CREATE A NEW A/B TEST' interface with the following details:

- Statistical parameters:**
 - Test type:** A winning variant has to demonstrate. Superiority, Non-inferiority.
 - Effect size as relative difference:** A horizontal bar chart showing the null hypothesis (red) and alternative hypothesis (green) regions. The x-axis ranges from -25% to 25%.
- Test variants:** Number of variants: 1 (selected), 2, 3, 4, 5, 6.
- Primary metric:**
 - The primary metric is: an average per user (dropdown).
 - Eligible users per week: 164000 (input field).
- Summary data:**
 - Baseline average: 3.30 (input field).
 - Baseline standard deviation: 16.5 (input field).
- Buttons:** 'Continue to business parameters' (orange button).

Abbildung 30: Sequential Testing - Testidee 1: Statistische Parameter

Schritt 3: Test Planning

Für diese Testidee (siehe Abb. 31) wurde ein Power Level von 80% festgelegt, was bedeutet, dass der Test eine 80%ige Wahrscheinlichkeit hat, einen tatsächlichen Effekt zu erkennen, wenn er vorhanden ist. Die Frequenz, mit der die Test-daten überwacht und abgerufen werden, ist auf alle 4 Tage festgelegt. (Bei DRIP Agency werden die Daten im Schnitt alle 3-6 Tage ausgewertet) Ein weiterer wichtiger Parameter ist der MDE. Bei einem Konfidenzniveau von 80% und einer Power von 80% muss der erkannte Effekt mind. 1,81% betragen, um als signifikant betrachtet zu werden. Obwohl der sequenzielle Ansatz betrachtet wird, ist die maximale Dauer dieses Tests auf 6 Wochen begrenzt, da in der Praxis oft Kunden bzw. Unternehmen nicht die Zeit und Ressourcen haben einen längeren Test auszuführen. Daher wird mit den angegebenen Parametern auch eine Zeitangabe angegeben, bis wann der Test den MDE erreichen sollte.



○ 6w 2d 1.81%

Abbildung 31: Sequential Testing - Testidee 1: Testplanung

Schritt 4: Test Monitoring

Im Rahmen des sequenziellen Testansatzes werden in regelmäßigen Abständen, konkret alle vier Tage, die relevanten Kennzahlen wie Nutzerdaten, der durchschnittliche Umsatz pro Nutzer und die Standardabweichung sowohl für die Referenz als auch für die Testvariante erhoben und einer tiefgehenden statistischen Analyse unterzogen. Jeder dieser Datenabrufe manifestiert sich als ein

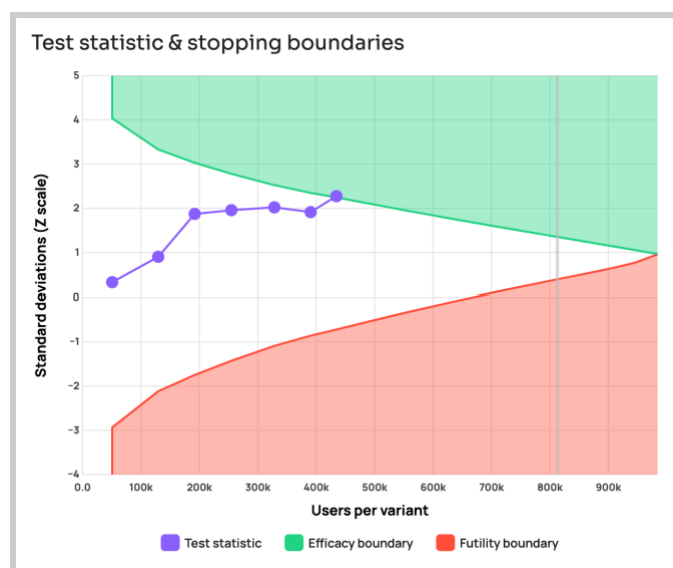


Abbildung 32: Sequential Testing - Testidee 1: Darstellung des Testverlaufs

einzelner Punkt in der grafischen Darstellung des Testverlaufs (siehe Abb. 32). Wenn der Graph die sogenannte "Efficacy Boundary", repräsentiert durch eine grüne Linie, überschreitet, weist dies auf eine signifikante positive Abweichung hin. In einem solchen Fall kann der Test vorzeitig beendet und die Änderung in der Praxis umgesetzt werden. Im Gegensatz dazu signalisiert das Überschreiten der "Futility Boundary", dargestellt durch eine rote Linie, dass der Test keine signifikante positive Wirkung zeigt. Dies impliziert, dass der erzielte Effekt entweder zu marginal ist oder keinen zusätzlichen Nutzen für den Shop bietet, weshalb der Test abgebrochen und die Änderung nicht implementiert wird. Im konkreten Fall konnte der Test durch den sequentiellen Ansatz und aufgrund des Überschreitens der Efficacy Boundary bereits nach 3 Wochen und 6 Tagen erfolgreich abgeschlossen werden.

Schritt 5: Zusammenfassung & Testauswertung

Zusammenfassend lässt sich sagen:

- Im geplanten Test wird eine Variante gegen eine Kontrollversion verglichen. Dabei wird eine Konfidenz von 80% angestrebt, um zu bestimmen, ob die Variante besser als die Kontrolle ist. Das Experiment sieht insgesamt 21 Datenauswertungen vor, wobei maximal 984.342 Nutzer pro Variante erfasst werden sollen. Mit dieser Stichprobengröße besteht eine 80%ige Chance, einen tatsächlichen Unterschied von 1,62% zu erkennen.
- Es gibt auch eine flexible Grenze, die "Futility-Grenze", die besagt, dass der Test nach Belieben gestoppt oder fortgesetzt werden kann, je nachdem, wie die Daten aussehen.
- Die zugrunde liegenden Hypothesen sind einfach: Entweder zeigt die Variante einen signifikanten positiven Effekt im Vergleich zur Kontrolle (Alternativhypothese) oder sie tut dies nicht (Nullhypothese).

Ergebnis: Die bis zu diesem Zeitpunkt gesammelte statistische Evidenz (44,2932% der geplanten Höchstzahl von 984.342 Nutzern) zeigt, dass es genügend Daten gibt, um den Test zu stoppen und die Variante 1 als Gewinner zu erklären. Bei Analyse Nr. 7 (Datenpunkt 7) mit etwa 435.997 Nutzern pro Variante liegt die Efficacy Boundary (Wirksamkeitsgrenze) bei 2,246, während der beobachtete Wert der Statistik 2,275 beträgt, was darauf hindeutet, dass Variante 1 mit einem Konfidenzniveau von 98,44 % zum Gewinner erklärt wird (der bereinigte p-Wert beträgt 0,015644). Die beobachtete Verbesserung der Best Performing Variante (Variante 1) beträgt 3,77 % mit einem 80 %-Konfidenzintervall von 1,62 % bis 5,95 % (siehe Abb. 33). Die Anzahl der Nutzer, die erforderlich ist, um zu der aktuellen Schlussfolgerung zu gelangen, beträgt 53,64 % derjenigen eines entsprechenden Plans mit festem

Stichprobenumfang, was bedeutet, dass der Test mit 46,36 % weniger Nutzern durchgeführt wurde, als dies bei einem einzelnen Evaluierungsplan der Fall gewesen wäre.

Cumulative data at the final observation					
Key metrics and statistical estimates for all test variants.					
	USERS	MEAN	SIGMA	% CHANGE	CONFIDENCE
Control	435,997	3.687	28.01		
Variant A	433,228	3.826	29.01	3.77%	98.44%

Abbildung 33: Sequential Testing - Testidee 1: Finale Testauswertung

Für die Testidee 2, die auf dem gleichen Konzept basiert, jedoch nun mit einem festgelegten Zeitraum (Fixed Horizon) durchgeführt wird, werden im Folgenden die relevanten Daten und Parameter dargelegt.

Schritt 1: Test Basics

Analog zum sequenziellen Verfahren wurde auch im Fixed-Horizon Ansatz der Test unter der Bezeichnung "Preview Funktion auf Desktop Geräten" initiiert. Der Startzeitpunkt beider Testverfahren war identisch und datiert auf den 15.06.2023. Ein charakteristisches Merkmal des Fixed-Horizon Ansatzes ist die Festlegung eines Endzeitpunkts, welcher in diesem Fall auf den 27.07.2023 terminiert ist.

Schritt 2: Statistische Parameter

Dieser Testschritt ist beim Fixed-Horizon Ansatz identisch mit dem des sequenziellen Ansatzes.

- **Test Type:** Superiority wird gewählt, da die Variante einen signifikant positiven Effekt nachweisen muss.
- **Anzahl der Test Varianten:** Es gibt nur eine Testvariante in diesem Test.
- **Primäre Metrik:** Durchschnittlicher Umsatz pro Besucher (ARPU).
- **Erwartete Nutzer pro Woche:** 164,000 Nutzer auf der Produktdetailseite für Desktop Geräte.
- **Baseline Average:** Durchschnittlicher Umsatz pro Besucher beträgt 3.30 €.
- **Standard Deviation:** 16.5.

Schritt 3: Test Planning

Analog zum sequenziellen Ansatz wurde auch hier ein Power Level von 80% festgelegt. Da beide Methoden miteinander verglichen werden, erfordert dies die gleichen statistische Niveau & Power Settings. Bei einem Konfidenzniveau von 80% und einer Power von 80% muss der erkannte Effekt beim Fixed-Horizon Ansatz mindestens 1,70% betragen, um als signifikant betrachtet zu werden. Die maximale Testdauer beträgt 6 Wochen (siehe Abb. 34)

CREATE A NEW A/B TEST

Test planning

Monitoring type

Sequential
Use AGILE testing to periodically evaluate results and stop early if the data warrants it

Single
Evaluate the outcome just once at the end of the test using a fixed sample size/ test duration

Test parameters

Confidence level: 80 %

Minimum effect of interest: (optional) %

Power level: 80 %

Fixed parameters

Monitoring frequency: none

Advanced parameters [SHOW](#)

[Estimate test plan](#)

6w 2d 1.70%

Abbildung 34: Fixed-Horizon-Test - Testidee 1: Testplanung

Schritt 4: Test Monitoring

Nach einer festgelegten Laufzeit von 6 Wochen wurde der Test einmalig ausgewertet, ohne Zwischenstände oder "Peeking", wobei die Daten zu Nutzern, ARPU und STDEV erfasst und analysiert wurden. Da es beim Fixed-Horizon-Test nur eine einmalige Analyse gibt, hat dieser keinen Test-Graphen mit Futility- und Efficacy-Boundary.

Schritt 5: Testauswertung

Die Abb. 35 zeigt, dass in der Referenzgruppe insgesamt 628,257 Nutzer erfasst wurden, während die Testvariante 629,200 Nutzer verzeichnete. Der durchschnittliche Umsatz pro Nutzer (ARPU) in der Referenz betrug 3,98 €, wohingegen die Test-Variante einen ARPU von 4,09 € aufwies. Dies resultierte in einem festgestellten Uplift von +2,76%. Somit konnte auch im Fixed Test ein Gewinner ermittelt werden.

	USERS	MEAN	SIGMA	% CHANGE	CONFIDENCE
Control	628,257	3.980	35.64		
Variant A	629,200	4.090	36.23	2.76%	95.60%

Abbildung 35: Fixed-Horizon-Test - Testidee 1: Finale Testauswertung

Zusammenfassend zeigt die Auswertung, dass der sequenzielle Ansatz im Vergleich zum Fixed-Horizon Ansatz überlegen ist, da durch diese Methode eine Testlaufzeit von etwas mehr als zwei Wochen eingespart wurde. Diese Effizienzsteigerung ermöglicht eine zeitnahe Implementierung der signifikant positiven Testvariante in den Shop, wodurch potenzielle Umsatzsteigerungen früher realisiert werden können.

4.4 Durchführung & Auswertung der Testidee 2 „Platzierung einer Announcement Bar (Banner) auf der Landing Page“

Der folgende Abschnitt befasst sich mit der Durchführung & Auswertung der Testidee 2 und beginnt mit dem Sequentiellen Ansatz:

Schritt 1: Test Basics

Am 05.07.2023 wurde der Test zur "Implementierung einer Announcement Bar mit Hinweis auf Newsletter Sign Ups und App Downloads" gestartet. Gemäß dem sequenziellen Ansatz ist kein festes Enddatum bestimmt.

Schritt 2: Statistische Parameter

Für den Test, der den Typ "Superiority" hat, wird eine Testvariante genutzt. Als primäre Metrik wurde der durchschnittliche Umsatz pro Besucher ausgewählt. Auf der Testseite, der Produktdetailseite für Desktop-Geräte, werden wöchentlich etwa 84.000 Besucher erwartet. Der durchschnittliche Umsatz je Besucher auf dieser Seite beträgt €4.51. Die Standardabweichung dieses Durchschnittswerts liegt bei 18.9.

Schritt 3: Test Planning

CREATE A NEW A/B TEST

Test planning

Monitoring type

Sequential
Use AGILE testing to periodically evaluate results and stop early if the data warrants it.

Single
Evaluate the outcome just once at the end of the test using a fixed sample size / test duration.

Test parameters

Confidence level: 80 %

Minimum effect of interest: (optional) %

Power level: 80 %

AGILE parameters

Monitoring frequency: Every 5 days

Advanced parameters SHOW

Estimate test plan

6w 2d **2.35%**

Abbildung 37: Sequential Testing - Testidee 2: Testplanung

POSSIBLE STATISTICAL PLANS

Select test duration

Select how long the test will run based on its sensitivity expressed below as the minimum relative percentage effect detectable with probabilities as shown in the column headers. For sequential tests the *expected duration* is significantly shorter than the maximum displayed below.

Maximum duration / Power	50%	80%	90%	95%	99%
<input type="radio"/> 1w 3d	2.09%	4.18%	5.27%	6.17%	7.86%
<input type="radio"/> 2w 1d	1.74%	3.48%	4.39%	5.14%	6.55%
<input type="radio"/> 2w 6d	1.53%	3.05%	3.85%	4.51%	5.74%
<input type="radio"/> 3w 4d	1.38%	2.75%	3.47%	4.07%	5.18%
<input type="radio"/> 4w 2d	1.27%	2.53%	3.19%	3.74%	4.76%
<input type="radio"/> 5 weeks	1.18%	2.35%	2.97%	3.48%	4.43%
<input type="radio"/> 5w 5d	1.11%	2.21%	2.79%	3.27%	4.16%
<input type="radio"/> 6w 3d	1.05%	2.09%	2.64%	3.09%	3.94%
<input type="radio"/> 7w 1d	0.99%	1.99%	2.51%	2.94%	3.74%
<input type="radio"/> 7w 6d	0.95%	1.90%	2.40%	2.81%	3.58%
<input type="radio"/> 8w 4d	0.91%	1.82%	2.30%	2.69%	3.43%
<input type="radio"/> 9w 2d	0.88%	1.75%	2.21%	2.59%	3.30%

Abbildung 36: Sequential Testing - Testidee 2: Möglicher Testplan zur Erreichung des MDE

Für die Testidee 2 wurden folgende Daten ermittelt:

- Konfidenzlevel: 80%
- Power Level: 80%
- Sequential Parameter - Monitoring Frequency: alle 5 Tage (siehe Abb. 37)
- MDE: Bei 80% Konfidenz und 80% Power liegt dieser in diesem Test bei 2.35%
- Maximale Testlaufzeit, in welcher dieser Effekt mit den oben genannten Testplanungs-Daten nachgewiesen werden kann: 5 Wochen (siehe Abb. 36)

Schritt 4: Test Monitoring

Alle fünf Tage wurden die Daten zu Nutzern, ARPU und STDEV sowohl für die Referenz als auch für die Variante erhoben und statistisch ausgewertet. Durch den sequenziellen Ansatz und die damit überschrittene Futility-Boundary, konnte der Test schon nach 1 Woche und 3 Tagen frühzeitig abgeschlossen werden (siehe Abb. 38).

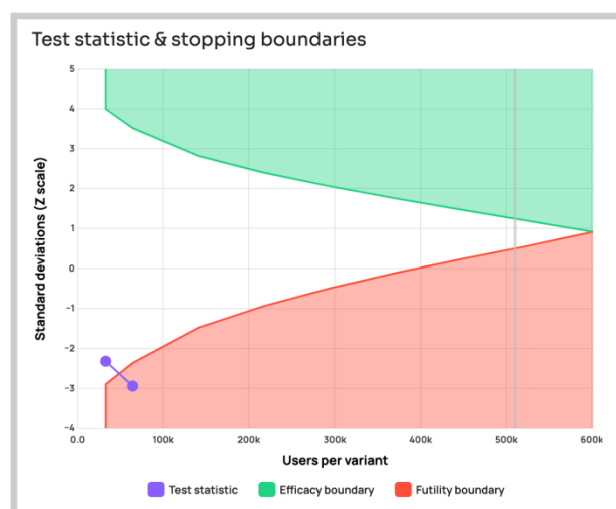


Abbildung 38: Sequential Testing - Testidee 2: Darstellung des Testverlaufs

Schritt 5: Zusammenfassung & Testauswertung

- Das Experiment ist für maximal 9 Datenauswertungen geplant, was einer maximalen Anzahl von Nutzern pro Variante von 210.000 entspricht. Diese maximale Stichprobengröße ergibt eine 80%ige Wahrscheinlichkeit, einen wahren, relativen Unterschied von 2,35% bei einer Konfidenzschwelle von 80% zu entdecken.
- Der Test ist mit einer unverbindlichen Futility-Grenze geplant, d. h. er kann nach eigenem Ermessen abgebrochen oder fortgesetzt werden, wenn die Teststatistik die Futility-Grenze überschreitet.

Ergebnis: Die bis zu diesem Punkt gesammelte statistische Evidenz (25,460 % der geplanten Höchstzahl von 210.000 pro Variante) zeigt, dass es genügend Daten gibt, um vorzuschlagen, den Test auf Nutzlosigkeit abubrechen und alle Versionen, einschließlich der leistungsstärksten Version A, als schlechter als die Kontrollversion zu erklären.

Die Abb. 39 zeigt, dass die Kontrollversion mit insgesamt 12.755 User 1.354 Transaktionen registrierte und eine CR von 10,62% sowie einen RPU von 6,71€ aufwies. Im Gegensatz dazu verzeichnete die Variante mit 12.705 Nutzern 1.300 Transaktionen mit einer CR von 10,23% und einem RPU von 6,59€. Betrachtet man nun die beiden Varianten näher sieht man, dass lediglich nur 1.661 Nutzern auf den Newsletter Slider geklickt haben und 1.119 auf den App Slider. Dies bedeutet wiederum, dass von den 1.300 Transaktionen, die die Variation in Betracht gezogen haben, lediglich 226 Nutzer über den Newsletter und 99 über die App eine Transaktion durchführten. Diese profitierten dabei von einem gewährten Rabatt. Bei einem direkten Vergleich beider Versionen ergab sich daher eine Verringerung der Conversion Rate um 3,61% und der RPU um 1,79% in der Variante im Vergleich zur Kontrollversion. Dies deutet darauf hin, dass Nutzer, die die Variante zu Gesicht bekamen in beiden Metriken weniger effektiv war als die Kontrollversion.

Segment	Device Category	ExperimentID	Users	Transactions	Ecommerce Conversion Rate	Improvement Rate	RPU	Improvement Rate	Average Order Value	Improvement Rate
All Users	all	Control	12,755	1,354	10,62%		6,71€		€63,21	
All Users	all	Variation 1	12,705	1,300	10,23%	-3,61%	6,59€	-1,79%	€64,40	1,88%
All Users	all	Variation 1 - Click button in newsletter slide	1661	226	13,61%		9,56€		€70,26	
All Users	all	Variation 1 - Click button in app slide	1119	99	8,85%		6,71€		€75,84	

grau hinterlegte Felder (Zeile 11-12) dienen dem Vergleich. Zeile 11 + 12 bilden die Nutzer der Variante 1 ab, welche aktiv auf den Newsletter Button bzw. den App Button geklickt haben. Verbesserungsraten (Improvement Rates) können sowohl im Vergleich zur Control Group, als auch zu den Gesamt-Nutzern der Variante verglichen / gebildet werden.

Abbildung 39: Sequential Testing - Testidee 2: Finale Testauswertung

Zudem lag mit etwa 12.705 Nutzern pro Variante die Futility-Grenze bei -2,361, während der beobachtete Wert der Statistik bei -2,940 lag, was darauf hindeutete, dass es unwahrscheinlich ist, dass alle Varianten den gewünschten Mindesteffekt mit dem gewünschten

Konfidenzniveau zeigen. Die beobachtete Verbesserung in der leistungs-stärksten Version A beträgt -1,79 % mit einem vorläufigen 80 %-Konfidenzintervall, das von -5,30 % bis 1,87 %.

Es folgt der Fixed-Horizon Ansatz:

Schritt 1: Test Basics

Basierend auf dem sequenziellen Verfahren fand der Test "Implementierung einer Announcement Bar mit Hinweis auf Newsletter Sign Ups und App Downloads" auch im Kontext des Fixed-Horizon Ansatzes statt. Beide Methoden wurden am 05.07.2023 gleichzeitig gestartet, wobei das Enddatum für den Test auf den 09.08.2023 festgesetzt wurde.

Schritt 2: Statistische Parameter

➔ Siehe Testidee 2 sequenzieller Ansatz

Schritt 3: Test Planning

CREATE A NEW A/B TEST

Test planning

Monitoring type

Sequential
Use AGILE testing to periodically evaluate results and stop early if the data warrants it

Single
Evaluate the outcome just once at the end of the test using a fixed sample size/ test duration

Test parameters

Confidence level: 80 %

Minimum effect of interest: (optional) %

Power level: 80 %

Fixed parameters

Monitoring frequency: none

Advanced parameters [SHOW](#)

[Estimate test plan](#)

6w 2d **2.18%**

Abbildung 40: Fixed-Horizon-Test - Testidee 2: Testplanung

POSSIBLE STATISTICAL PLANS

Select test duration

Select how long the test will run based on its sensitivity expressed below as the minimum relative percentage effect detectable with probabilities as shown in the column headers. For sequential tests the *expected duration* is significantly shorter than the maximum displayed below.

Duration / Power	50%	80%	90%	95%	99%
<input type="radio"/> 1 week	2.43%	4.87%	6.14%	7.19%	9.16%
<input type="radio"/> 2 weeks	1.72%	3.44%	4.34%	5.08%	6.48%
<input type="radio"/> 3 weeks	1.41%	2.81%	3.54%	4.15%	5.29%
<input type="radio"/> 4 weeks	1.22%	2.43%	3.07%	3.60%	4.58%
<input checked="" type="radio"/> 5 weeks	1.09%	2.18%	2.75%	3.22%	4.10%
<input type="radio"/> 6 weeks	0.99%	1.99%	2.51%	2.94%	3.74%

[Show power graph](#)

Abbildung 41: Fixed-Horizon-Test - Testidee 2: Testplanung zur Erreichung des MDE

Bei einem Konfidenzniveau von 80% und einer Power von 80% sollte der festgestellte Effekt im Fixed-Horizon Ansatz mindestens 2,18% erreichen, um als signifikant eingestuft zu werden. Um sicherzustellen, dass die Ergebnisse rechtzeitig vorliegen und analysiert werden können, wurde die Laufzeit auf höchstens 5 Wochen begrenzt (siehe Abb. 40 & 41)

Schritt 4: Test Monitoring

Nach exakt 5 Wochen wurde der Test einmalig ausgewertet. Peeking bzw. Zwischenstände erfolgten auch hier während der Testlaufzeit nicht.

Schritt 5: Testauswertung

Die in diesem Fixed-Horizon-Test gesammelten statistischen Daten zeigen in Abb. 42, dass die Daten es rechtfertigen, alle Versionen, einschließlich der leistungsstärksten Version A, als schlechter als die Kontrollversion zu erklären. Bei 87.299 Nutzern erzielte die Kontrollversion 8.793 Transaktionen, eine CR von 10,07% und einen RPU von 6,51€. Die Variante, betrachtet mit 88.294 Nutzern, verzeichnete 8,263 Transaktionen, eine CR von 9,36% und einen RPU von 6,09€. Trotz der Interaktion von 3.247 Nutzern mit dem Newsletter Slider und 2.246 mit dem App Slider führten nur 362 und 107 Nutzer Transaktionen durch und erhielten einen Rabatt. Im direkten Vergleich wies die Variante eine um 7,09% niedrigere Conversion Rate und einen um 6,45% reduzierten RPU im Vergleich zur Kontrollversion auf. Dies legt nahe, dass die Variante in beiden Metriken weniger performant war als die Kontrollversion.

Segment	Device Category	ExperimentID	Users	Transactions	Ecommerce Conversion Rate	Improvement Rate	RPU	Improvement Rate	Average Order Value
All Users	all	Control	87.299	8.793	10.07%		6.51€		64.63€
All Users	all	Variation 1	88.294	8.263	9.36%	-7.09%	6.09€	-6.45%	65.07€
All Users	all	Variation 1 - Click button in newsletter slide	3247	362	11.15%		8.46€		75.89€
All Users	all	Variation 1 - Click button in app slide	2246	107	4.76%		4.30€		90.26€

grau hinterlegte Felder (Zeile 11-12) dienen dem Vergleich. Zeile 11 + 12 bilden die Nutzer der Variante 1 ab, welche aktiv auf den Newsletter Button bzw. den App Button geklickt haben. Verbesserungsdaten (Improvement Rates) können sowohl im Vergleich zur Control Group, als auch zu den Gesamt-Nutzern der Variante verglichen / gebildet werden.

Abbildung 42: Fixed-Horizon-Test - Testidee 2: Finale Testauswertung

Obwohl in beiden Ansätzen die Variante A suboptimal performte, zeigt sich der sequenzielle Ansatz erneut als überlegen. Innerhalb von nur 1 Woche und 3 Tagen konnte festgestellt werden, dass der Test nicht signifikant positive Ergebnisse lieferte. Dies ermöglicht eine frühzeitige Anpassung und verhindert somit potenzielle Verluste. In einem dynamischen Geschäftsumfeld kann diese zeitnahe Erkenntnisgewinnung entscheidend sein, um strategische Entscheidungen zu treffen und Ressourcen effizienter zu allokkieren.

4.5 Durchführung & Auswertung der Testidee 3 „Bedienungshilfe: Text- und Anzeigeeinstellungen“

Analog zur Testidee 1 und 2 richtet sich der Fokus dieser Analyse auf die Durchführung und anschließende Bewertung eines Accessibility-Features auf der Homepage (Desktop Geräte). Für die Durchführung des Smoke Tests zur Vorabüberprüfung grundlegender Funktionalitäten oder Informationen wurde bei der Analyse sowohl der sequenzielle Ansatz als auch der Fixed-Horizon Ansatz verwendet. Zunächst wird der sequenzielle Ansatz vorgestellt.

Schritt 1: Test Basics

Unter der Bezeichnung "Implementierung eines Accessibility-Features auf der Homepage – Desktop Geräte" wurde am 28.06.2023 der Test initiiert. Wie auch bei Test 1 und 2 wurde im Rahmen des Sequential Ansatzes keine feste Endzeit festgelegt.

Schritt 2: Statistische Parameter

Gleich wie bei Test 1 und 2 wurde der "Superiority"-Typ als Testkriterium gewählt. Für diese spezifische Untersuchung gibt es nur eine Testvariante. Als Hauptkennzahl zur Auswertung dient der durchschnittliche Umsatz je Besucher (ARPU). Für diesen Test wird erwartet, dass wöchentlich rund 49.000 Nutzer die Produktdetailseite für Desktop-Geräte besuchen. Der Baseline Average, welcher den durchschnittlichen Umsatz pro Besucher repräsentiert, beträgt €11.14. Die Standardabweichung dieses Wertes liegt bei 31.3.

Schritt 3: Test Planning

Für die Testidee 3 wurden folgende Daten ermittelt:

- Konfidenzlevel: 80%
- Power Level: 80%
- Sequential Parameter - Monitoring Frequency: alle 4 Tage
- MDE: Bei 80% Konfidenz und 80% Power liegt dieser in diesem Test bei 2.05% (siehe Abb. 43)
- Maximale Testlaufzeit, in welcher dieser Effekt mit den oben genannten Testplanungs-Daten nachgewiesen werden kann: 5 Wochen 1 Tag

CREATE A NEW A/B TEST

Test planning

Monitoring type

Sequential
Use AGILE testing to periodically evaluate results and stop early if the data warrants it.

Single
Evaluate the outcome just once at the end of the test using a fixed sample size / test duration.

Test parameters

Confidence level: 80 %

Minimum effect of interest: (optional) %

Power level: 80 %

AGILE parameters

Monitoring frequency: Every 4 days

Advanced parameters: SHOW

Estimate test plan

6w 2d 2.05%

Abbildung 43: Sequential Testing - Testidee 3: Testplanung

Schritt 4: Test Monitoring

In regelmäßigen Intervallen von vier Tagen wurden die Nutzer-, ARPU- und STDEV-Zahlen der Referenz und der Variante gezogen und statistisch analysiert. Unter Anwendung des sequenziellen Ansatzes und in Anbetracht des Überschreitens der Futility-Boundary konnte der Test bereits nach 2 Wochen und 1 Tag vorzeitig beendet werden (siehe Abb. 44)

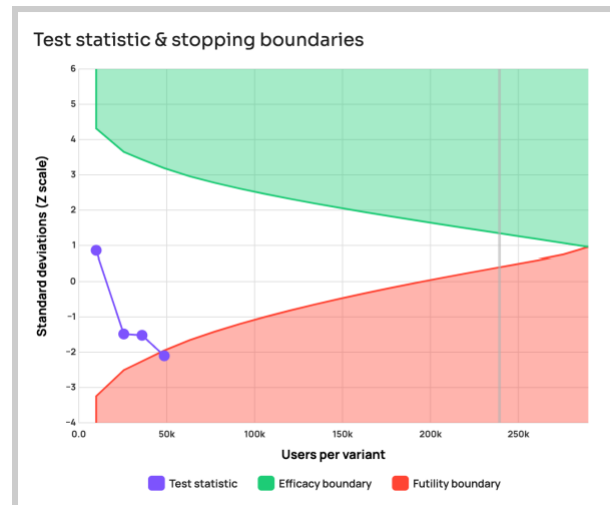


Abbildung 44: Sequential Testing - Testidee 3: Darstellung des Testverlaufs

Schritt 5: Zusammenfassung & Testauswertung

Zusammenfassend lässt sich sagen:

- Das Experiment ist für maximal 21 Datenauswertungen geplant, was einer maximalen Anzahl von Nutzern pro Variante von 252.000 entspricht. Diese maximale Stichprobengröße ergibt eine 80%ige Wahrscheinlichkeit, einen wahren, relativen Unterschied von 2,05% bei einer Konfidenzschwelle von 80% zu entdecken.
- Der Test ist mit einer unverbindlichen Futility-Grenze geplant, d. h. er kann nach eigenem Ermessen abgebrochen oder fortgesetzt werden, wenn die Teststatistik die Futility-Grenze überschreitet.

Ergebnis: Die bis zu diesem Punkt gesammelte statistische Evidenz (16,9051% der geplanten Höchstzahl von 252.000 pro Variante) zeigt, dass es genügend Daten gibt, um vorzuschlagen, den Test abubrechen und die Test-Variante als unwahrscheinlich zu erklären, unter den gegebenen Testparametern ein Gewinner sein zu können. Bei Analyse Nr. 4 mit ca. 48.285 Nutzern pro Variante liegt die Futility-Grenze bei -1,950, während der beobachtete Wert der Statistik bei -2,102 liegt, was darauf hindeutet, dass es unwahrscheinlich ist, dass alle Varianten den gewünschten Mindesteffekt mit dem gewünschten Konfidenzniveau zeigen. Die beobachtete Verbesserung in der leistungsstärksten Version A beträgt -9,38% mit einem 80%igen Konfidenzintervall von -2,22% bis 1,46%. Die Anzahl der Nutzer, die benötigt wird, um zu der

aktuellen Schlussfolgerung zu gelangen, beträgt 20,47% derjenigen eines entsprechenden Plans mit fester Stichprobengröße, was bedeutet, dass der Test mit 79,53% weniger Nutzern durchgeführt wurde, als dies bei einem einzelnen Evaluierungsplan der Fall gewesen wäre.

Cumulative data at the final observation				
Key metrics and statistical estimates for all test variants.				
	USERS	MEAN	SIGMA	% CHANGE
Control	48,285	15.913	117.40	
Variant A	48,985	14.420	103.82	-9.38%

Abbildung 45: Sequential Testing - Testidee 3: Finale Testauswertung

Es folgt der Fixed-Horizon Ansatz:

Schritt 1: Test Basics

In Anlehnung an das sequenzielle Verfahren wurde der Test "Implementierung eines Accessibility Features auf der Homepage – Desktop Geräte“ ebenfalls im Rahmen des Fixed-Horizon Ansatzes durchgeführt. Beide Ansätze starteten zeitgleich am 28.06.2023. Das Enddatum wurde hier auf den 27.07.2023 festgelegt.

Schritt 2: Statistische Parameter

→ Siehe Testidee 3 sequenzieller Ansatz

Schritt 3: Test Planning

Für die Testidee 3 wurden folgende Daten ermittelt:

- Konfidenzlevel: 80%
- Power Level: 80%
- MDE: Bei 80% Konfidenz und 80% Power liegt dieser in diesem Test bei 1.91% (siehe Abb. 46)
- Maximale Testlaufzeit, in welcher dieser Effekt mit den oben genannten Testplanungs-Daten nachgewiesen werden kann: 5 Wochen

CREATE A NEW A/B TEST

Test planning

Monitoring type

Sequential
Use AGILE testing to periodically evaluate results and stop early if the data warrants it

Single
Evaluate the outcome just once at the end of the test using a fixed sample size/ test duration

Test parameters

Confidence level: 80 %

Minimum effect of interest: (optional) %

Power level: 80 %

Fixed parameters

Monitoring frequency: none

Advanced parameters [SHOW](#)

[Estimate test plan](#)

6w 2d 1.91%

Abbildung 46: Fixed-Horizon-Test - Testidee 3: Testplanung

Schritt 4: Test Monitoring

Nach exakt 5 Wochen wurde der Test einmalig ausgewertet. Peeking bzw. Zwischenstände erfolgten auch hier während der Testlaufzeit nicht.

Schritt 5: Testauswertung

Nach exakt 5 Wochen liefert der Test eindeutige Ergebnisse: Im Vergleich zur Kontrollversion hat sich bei der Variante die Verbesserungsrate (Improvement Rate) in Bezug auf die Kennzahlen Conversion Rate und Revenue per User (Erlös pro Nutzer) verschlechtert. Bei einer detaillierten Analyse zeigt die Variation 1 in Abb. 47 eine Gesamtbenutzerzahl von 117.818, die zu 15.174 Transaktionen führte, was einer Konversionsrate von 12,88% entspricht. Im Gegensatz dazu weist die Kontrollversion eine Benutzerbasis von 117.702 auf, die 15.247 Transaktionen generierte, was in einer Konversionsrate von 12,95% resultiert. Obwohl die Unterschiede in der Anzahl der Nutzer und Transaktionen zwischen den beiden Versionen marginal sind, ist die Kontrollversion leicht überlegen und führt zu einer besseren Konversionsrate und RPU.

Last Run On	8/2/2023								
View Name	1 - Main Production View (Hauptansicht)								
Total Results Found	2								
Total Results Returned	2								
Contains Sampled Data	No								
Results Breakdown - Fixed									
Segment	Device Category	ExperimentID	Users	Transactions	Ecommerce Conversion Rate	Improvement Rate	RPU	Improvement Rate	
All Users	all	Control	117,702	15,248	12.95%		16.89€		
All Users	all	Variation 1	117,818	15,174	12.88%	-0.58%	16.79€		-0.65%

Abbildung 47: Fixed-Horizon-Test - Testidee 3: Finale Testauswertung

Abschließend lässt sich daher sagen, dass auch hier mit dem sequenziellen Ansatz fast 3 Wochen eingespart werden konnten, da dieser eine frühzeitige Erkennung der negativen Testergebnisse ermöglichte. Im Kontext des Smoke Tests bedeutet dies, dass potenzielle Schwachstellen oder Unstimmigkeiten in den Text- und Anzeigeeinstellungen rasch erkannt werden. Dies erlaubt es, Optimierungen und Anpassungen zügig vorzunehmen, bevor das Feature im Shop eingeführt wird, wodurch das Risiko einer nicht optimalen Nutzererfahrung minimiert werden kann.

5 Interpretation der Testergebnisse/ Handlungsempfehlungen und abschließendes Fazit

Das letzte Kapitel widmet sich der Interpretation der Testergebnisse, die mit Hilfe der beiden unterschiedlichen Methoden des A/B-Testing gewonnen wurden, und leitet daraus im Anschluss konkrete Handlungsempfehlungen ab. Abschließend wird ein Fazit zur Effektivität von sequenziellen A/B-Tests im Vergleich zu Fixed-Horizon Tests im Kontext der Optimierung der User Experience in Onlineshops gezogen.

Testidee 1

Wenn man die Resultate der Testidee 1 betrachtet, wird ein unmittelbarer positiver Effekt ersichtlich. Das Hover-Image und der damit assoziierte Produktvergleich bieten den Nutzern eine bessere Entscheidungsgrundlage, um ihre präferierte Farboption auszuwählen.

Durch das Feature konnten Nutzer Produkte detaillierter betrachten, was nicht nur das Verständnis des Produkts fördert, sondern auch dessen Visualisierung. Dies trägt direkt zu einer erhöhten Kaufwahrscheinlichkeit bei. Das Hauptmerkmal dieses Hover-Images ist der direkte Produktvergleich, ohne die aktuelle Produktseite verlassen zu müssen. Die intuitive Darstellung und Interaktion des Hover-Images verbessert die Benutzeroberfläche und bietet den Nutzern eine effizientere Shopping-Erfahrung. Insgesamt hat dieses Feature die Entscheidungsfindung beschleunigt und somit auch die CR erhöht. Durch den positiven A/B-Test können folgende Handlungsempfehlungen für den Onlineshop SNOCKS abgeleitet werden:

- Integration von Alt-Tags: Alt-Tags auch als alternative Texte bekannt verbessern die Webseite hinsichtlich ihrer Zugänglichkeit und Benutzererfahrung. Im Kontext der digitalen Barrierefreiheit spielen Alt-Tags eine essenzielle Rolle. Sie ermöglichen es Menschen mit Sehbehinderungen, die auf Screenreader angewiesen sind, visuelle Inhalte in Textform zu erfassen und zu interpretieren. Eine Handlungsempfehlung wäre daher das Integrieren von Alt-Tags, um die Zugänglichkeit und Beschreibung der Hover-Images für alle Nutzer zu optimieren.
- Optimierung der Bildqualität: Um die bestmögliche Produktansicht anzubieten, sollte sichergestellt werden, dass die Produktbilder in hoher Auflösung und Qualität bereitgestellt werden. Hochauflösende Bilder erlauben dem Nutzer, Produktmerkmale wie Material und Muster detailliert zu erkennen und minimieren so das Risiko von Fehlkäufen durch Missverständnisse oder falschen Erwartungen.
- Erweiterte Produktvergleiche: Um weitere detaillierte Vergleiche anzubieten, kann man überlegen ob zusätzliche Informationen in das Hover-Image integriert werden sollen. Es können bspw. Informationen über die Herkunft und Nachhaltigkeit des Produkts

angegeben werden oder besondere Merkmale hervorgehoben werden. Dennoch ist es wichtig, das Hover-Image nicht zu überladen, um die Nutzer nicht zu überfordern. Ein ausgewogenes Verhältnis von nützlichen Informationen und klarem Design ist hierbei entscheidend.

Um ein signifikant positives Ergebnis sicherzustellen, ist es daher wichtig das A/B-Testing kontinuierlich fortzuführen. Dadurch kann die Effektivität des Features regelmäßig überprüft werden und gegebenenfalls weitere Anpassungen vorgenommen werden.

Testidee 2

Trotz sorgfältiger Vorbereitung und Durchführung fehlte beim zweiten Test die erwartete signifikante positive Abweichung in den Resultaten. Die durchgeführte Analyse ergab, dass die allgemein getestete Variante den durchschnittlichen Umsatz pro Besucher reduziert hat.

Betrachtet man jedoch nicht die RPU sondern den durchschnittlichen Bestellwert (AOV) zeigt sich, dass im Gegensatz zur Kontrollversion die Variante um 1,88% besser abschnitt. (siehe Abbildung 39).

Da Rabattaktionen, wie beispielsweise hier der 15%-Rabatt bei Newsletter-Anmeldung, den RPU potenziell beeinflussen können, erscheint es sinnvoll, den AOV als alternative Kennzahl zu berücksichtigen. Der AOV ermöglicht eine differenzierte Betrachtung darüber, ob Kunden trotz gewährter Rabatte im Durchschnitt mehr oder weniger pro Transaktion ausgeben. Dies liefert wertvolle Erkenntnisse über die tatsächliche Wirkung von Rabattstrategien auf das Kaufverhalten.

Basierend auf den Ergebnissen war dennoch eine Implementierung der Variante nicht gerechtfertigt, da die Anzahl der Käufe, die über die App oder den Newsletter getätigt wurden, nicht ausreichend waren.

Interessant in diesem Test war jedoch die Betrachtung der Heatmaps. Heatmaps zeigen, welche Bereiche einer Webseite am häufigsten betrachtet oder angeklickt werden. Hellere oder "wärmere" Farben (z.B. Rot) zeigen Bereiche mit hoher Aktivität, während "kühlere" Farben (z.B. Blau/Grün) Bereiche mit geringerer Aktivität anzeigen. Wenn man die Abbildungen 48 und 49 anschaut, wird deutlich, dass die Announcement Bar in der mobilen Ansicht wesentlich stärker in den Fokus der Nutzer rückte. Dies zeigte sich insbesondere durch die intensiveren "warmen" Farben in diesem Bereich der Heatmap. Dies legt nahe, dass mobile Nutzer der Announcement Bar besondere Aufmerksamkeit schenken und sie als zentrales Informations- oder Interaktionselement auf der Seite wahrnehmen. Im Gegensatz dazu zeigt die Heatmap für die Desktop-Version, dass die Announcement Bar deutlich weniger Beachtung fand. Die "kühleren" Farben in diesem Bereich deuten darauf hin, dass die Nutzer sie weniger häufig betrachtet oder angeklickt haben. Dies legt nahe, dass bei Desktop-Nutzern der Fokus auf

anderen Bereichen der Webseite lag und die Announcement Bar nicht als zentrales Element der Interaktion oder Information wahrgenommen wurde.

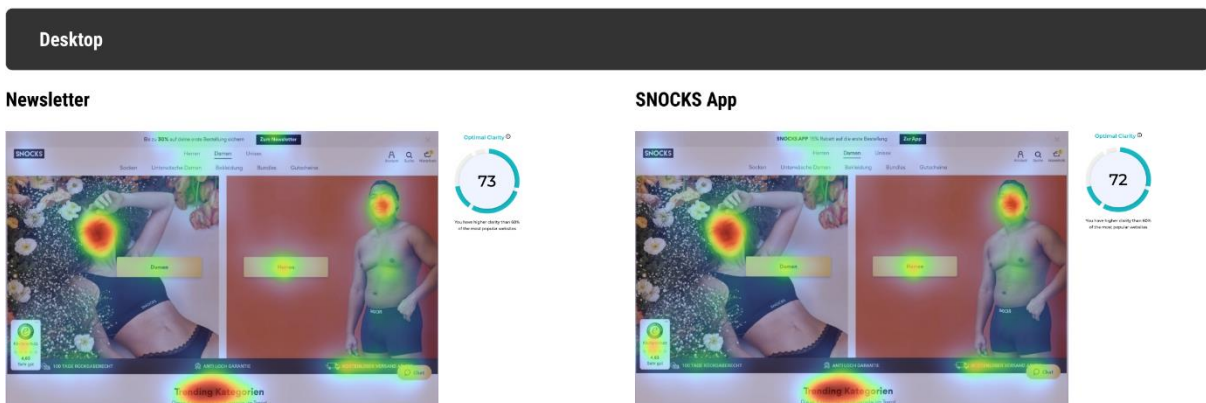


Abbildung 48: Heatmap-Analyse der Announcement Bar (Newsletter & SNOCKS App) für die Desktop-Version



Abbildung 49: Heatmap-Analyse der Announcement Bar (Newsletter & SNOCKS App) für die Mobile Version

Aus der unterschiedlichen Wahrnehmung & Interaktion mit der Announcement Bar zwischen mobilen und Desktop-Nutzern lassen sich daher folgende Handlungsempfehlung ableiten:

- Testen von verschiedenen Farboptionen: Die Farbgestaltung kann erheblich dazu beitragen, wie auffällig oder sichtbar ein bestimmtes Webseitenelement für die Nutzer ist. Daher kann eine gezielte Farbänderung die Sichtbarkeit und Interaktion mit der

Announcement Bar erhöhen. Mit weiteren A/B-Tests kann herausgefunden werden, ob eine Farbänderung zu mehr Sichtbarkeit und letztendlich Transaktionen führt.

Testidee 3

Betrachtet man die Ergebnisse der Testidee 3, insbesondere im Kontext eines Smoke Tests, wird deutlich, dass Nutzerinteraktionen und -verhalten von zentraler Bedeutung sind. Obwohl die Differenz in den Transaktionen geringfügig ist und die Kontrollversion leicht überlegen erscheint, sollte auch die Nutzungshäufigkeit des neuen Features in die finale Analyse miteinbezogen werden, um eine fundierte Beurteilung seiner Effektivität vorzunehmen. Ein möglicher Grund für die reduzierte Anzahl an Transaktionen könnte die Ablenkung durch das neu implementierte Feature sein. Jedes Element auf einer Webseite, insbesondere in einem Onlineshop, hat das Potenzial, die Kaufentscheidung eines Nutzers zu beeinflussen, sei es bewusst oder unbewusst. Dieser psychologische Aspekt sollte nicht unterschätzt werden, da er direkte Auswirkungen auf den Umsatz haben kann. Daher ist es immer auch wichtig den Kosten-Nutzen-Aufwand zu betrachten. Selbst wenn ein Feature häufig genutzt wird, aber zu einem Umsatzrückgang von beispielsweise 10% führt, könnte es kontraproduktiv für den Shop sein. Da jedoch in diesem Test ein Interaktionsanteil von etwa 40% im gegebenen Zeitraum verzeichnet wurde, lässt dies darauf schließen, dass ein erheblicher Teil der Gesamtbesucher aktiv mit dem neuen Element interagiert hat, was als ein positives Indiz gewertet werden kann.

Die Abb. 50 gibt einen Überblick darüber, wie häufig auf die unterschiedlichen Text- und Anzeigeneinstellungen geklickt wurde und stellt die Gesamtanzahl dieser Klicks dar. Die Datenanalyse veranschaulicht, dass der neue Accessibility-Button die Neugier der Nutzer geweckt hat, da sie insgesamt 20.374 Mal angeklickt wurde. Jedoch gab es nur rund $\frac{1}{4}$ der Klicks auf die unterschiedlichen Textgrößen, dies entsprachen lediglich 5.811 Gesamtklicks. Im Kontext der Textgrößen verzeichnete die Einstellung "Größe M" mit insgesamt 2.979 Interaktionen den höchsten Anteil an Klicks. In Bezug auf die Anzeigekonfigurationen wurde eine Gesamtinteraktion von lediglich 1.518 Klicks verzeichnet. Basierend darauf kann ausgegangen werden, dass der Dunkel-Hell-Kontrast mit 654 Klicks eine bevorzugte Anzeigekonfiguration von Benutzern ist. Betrachtet man die Kombinationsmöglichkeiten, so zeigt sich insbesondere die Kombination von Textgröße M und Dunkel-Hell-Kontrast mit 301 Interaktionen als dominierende Präferenz innerhalb der untersuchten Konfigurationen.

Die Präferenz der Nutzer für bestimmte Einstellungen, insbesondere die Kombination von Textgröße "M" und Dunkel-Hell-Kontrast, kann auf verschiedene Faktoren oder Gründe zurückgeführt werden:

- Tageszeitabhängigkeit: Die Uhrzeit kann die Nutzerpräferenzen beeinflussen, insbesondere in Bezug auf visuelle Einstellungen. Abends neigen viele dazu, den Dunkel-

Kontrast-Modus zu bevorzugen, um Augenbelastung zu reduzieren. Eine größere Schriftgröße kann zudem die Lesbarkeit verbessern und die Fokussierung auf Inhalte erleichtern, besonders wenn die Augen bereits ermüdet sind.

- Neugier und generelles Interesse: Wenn der Standard- oder Default-Modus bereits auf Hell-Dunkel eingestellt war, könnten viele Nutzer den Dunkel-Hell-Modus einfach aus Neugier auswählen. Menschen sind oft interessiert, Abwechslung zu suchen und Neues zu entdecken, besonders wenn es um digitale Interfaces geht.
- Ästhetik und Vermeidung von Überforderung: Menschen neigen dazu, Designs zu bevorzugen, die ästhetisch ansprechend sind. Ein harmonischer Kontrast und eine angemessene Textgröße können ein Gefühl der Zufriedenheit und des Wohlbefindens hervorrufen. Klare und gut lesbare Texte können zudem ein Gefühl der Sicherheit und Vertrautheit vermitteln, da der Benutzer das Gefühl hat, die Informationen vollständig zu verstehen.

Last Run On	8/2/2023	
View Name	1 - Main Production View (Hauptansicht)	
Total Results Found	23	
Total Results Returned	23	
Contains Sampled Data	No	
Results Breakdown		
Event label	Event action	Total Events
Clicked accessibility button	click	20374
Clicked any Textgröße	click	5811
Changed default Textgröße	click	4087
Clicked S	click	1724
Clicked M	click	2979
Clicked L	click	617
Clicked XL	click	491
Clicked any Anzeige	click	1518
Clicked Hell-Dunkel-Kontrast	click	376
Clicked Dunkel-Hell-Kontrast	click	654
Clicked Farbwähler	click	488
Clicked S and Hell-Dunkel-Kontrast	click	123
Clicked M and Hell-Dunkel-Kontrast	click	77
Clicked L and Hell-Dunkel-Kontrast	click	131
Clicked XL and Hell-Dunkel-Kontrast	click	45
Clicked S and Dunkel-Hell-Kontrast	click	169
Clicked M and Dunkel-Hell-Kontrast	click	301
Clicked L and Dunkel-Hell-Kontrast	click	121
Clicked XL and Dunkel-Hell-Kontrast	click	63
Clicked S and Farbwähler	click	217
Clicked M and Farbwähler	click	73
Clicked L and Farbwähler	click	93
Clicked XL and Farbwähler	click	105

Abbildung 50: Klickhäufigkeiten auf die unterschiedlichen Text- und Anzeigeneinstellungen (Testidee 3)

Da die Aussagen oder Gründe auf Annahmen basieren, ist es wichtig weitere Untersuchungen durchzuführen, um die Gründe für diese bestimmten Interaktionen zu ergründen. Es könnte sein, dass trotz der offensichtlichen Neugierde eine mangelnde Aufklärung über die

Funktionen des Accessibility-Buttons vorliegt. Daher ist es wichtig in Anbetracht der Ergebnisse und der Bedeutung von Nutzererfahrung, eine fundierte Handlungsempfehlung zu formulieren:

Smoke Tests haben das Risiko, nicht so präzise wie normale A/B-Tests zu sein, was die Ergebnisse beeinflussen bzw. auch evtl. verzerren kann. Es wäre daher sinnvoll, das getestete Feature in einem ausführlicheren A/B-Test zu prüfen. Ziel dabei ist es Schritt für Schritt die einzelnen TextEinstellungen zu implementieren und stärker auf die Barrierefreiheit einzugehen. In einer initialen Phase könnten beispielsweise zuerst die zwei bevorzugten Einstellungsoptionen, nämlich die Textgröße 'M' und der Modus für Dunkelheit-Helligkeit, implementiert werden. Das übergeordnete Ziel dieses Vorgehens sollte dann sein, den derzeit marginal schlechteren Effekt in Bezug auf die Transaktionen zu verbessern, sodass er letztlich zu einem positiv signifikanten Ergebnis führt. Dies würde nicht nur die UX optimieren, sondern auch den wirtschaftlichen Nutzen des Features maximieren. Mit Hilfe von Feedback-Optionen und Nutzerbefragungen kann herausgefunden werden, warum bestimmte Einstellungen bevorzugt werden und welche optimale Kombination von Einstellungen für die Nutzer am besten ist.

Der Vergleich zwischen sequenziellen Tests und Fixed Horizon Tests hat in der Theorie gezeigt, dass beide Ansätze ihre eigenen Vor- und Nachteile mitbringen. Insbesondere in Bezug auf die Geschwindigkeit der Ergebnisfindung und die Flexibilität der Testdurchführung unterscheiden sie sich deutlich. Nach eingehender Untersuchung und Analyse der drei Testideen wurde festgestellt, dass die sequenziellen Tests eine um 20-30% höhere Wahrscheinlichkeit für einen früheren Testabbruch aufwiesen und somit schneller zu Ergebnissen führten. Dies bedeutet, dass UN, die sequenzielle Tests verwenden, in der Lage sind, effizienter zu agieren, Ressourcen zu sparen und schneller auf Marktveränderungen oder Kundenfeedback zu reagieren. Zudem haben UN oft nicht die Ressourcen, um lang-wierige Tests durchzuführen. Sie wollen Zeit und Kosten sparen und gleichzeitig aber eine schnelle Entscheidungsfindung treffen.

In Anbetracht des sequenziellen Ansatz werden daher nun die zentralen Erkenntnisse in Bezug auf die Testideen zusammengefasst.

- User Experience: Durch den sequenziellen Ansatz können schneller Rückmeldungen zu neuen Features oder Änderungen gesammelt werden. Es ermöglicht UN schneller auf Nutzerverhalten zu reagieren. Insbesondere kann durch sequenzielles Testing schneller und effizienter festgestellt werden, wie gut beispielsweise verbesserte

Kontrastverhältnisse, von den Nutzern angenommen werden und ob sie die Zugänglichkeit der Website tatsächlich erhöhen.

- Frühzeitige Erkennung und Vermeidung von nachteiligen Funktionen/ Elementen: Der sequenzielle Ansatz stellt eine wertvolle Absicherungsstrategie dar, die das Risiko von Fehlinvestitionen minimiert.
- Erhöhte Rendite aus sequenziellen Tests: Eine positiv performende Variante bietet die Möglichkeit, sie zügig für die gesamte Nutzerschaft freizuschalten. Dadurch können UN ihren Umsatz zügiger und gezielter steigern.
- Follow-Up Tests mit dem sequenziellen Ansatz: Mit dem sequenziellen Ansatz bieten sich Follow-Up Tests als sinnvolle Strategie an, um die ursprünglichen Testideen, weiter zu optimieren. Es ist empfehlenswert, auf Basis der gewonnenen Erkenntnisse gezielte Anpassungen vorzunehmen und diese in nachfolgenden Tests erneut zu evaluieren. Dies ermöglicht eine schnellere und flexiblere Verbesserung und Anpassung an die Bedürfnisse der Nutzer.

Abschließend lässt sich sagen, dass das sequenzielle Verfahren im Vergleich zum Fixed-Horizon-Testing eine Reihe signifikanter Vorteile aufweist, die es zu einer bevorzugten Methode für viele Unternehmen machen. Während die Flexibilität und Effizienz des sequenziellen Ansatzes unbestreitbar sind, ist es dennoch von entscheidender Bedeutung, eine sorgfältige Analyse durchzuführen, um sicherzustellen, dass Tests nicht aufgrund kurzfristiger Schwankungen vorschnell beendet werden. In diesem Kontext, und in einem Zeitalter, in dem Technologien wie KI immer dominanter werden, bietet das sequenzielle A/B-Testing somit eine wertvolle Möglichkeit, agil auf Veränderungen im Nutzerverhalten zu reagieren. Es ermöglicht eine fortlaufende Optimierung der Online-Präsenz und hebt sich durch seine Schnelligkeit und Anpassungsfähigkeit von traditionellen Testmethoden ab. Es ist ein Werkzeug, das, wenn es richtig eingesetzt wird, den Weg für kontinuierliche Verbesserungen und Wachstum ebnen kann.

6 Anhang

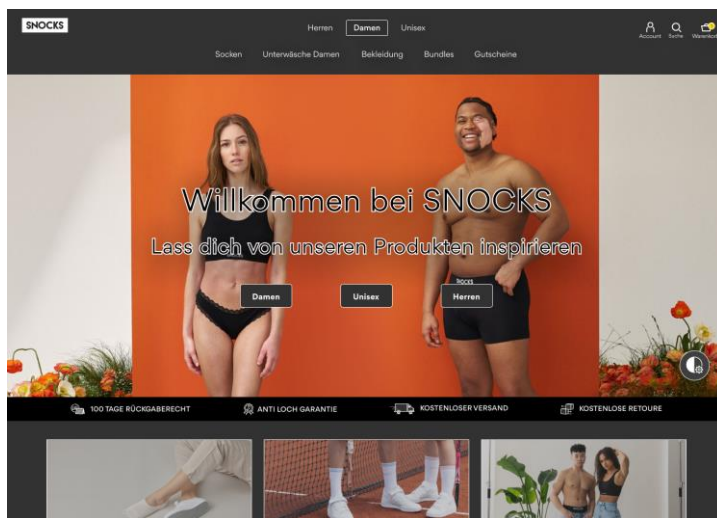


Abbildung 51: Dunkel-Hell Kontrastmodus mit Textgröße S

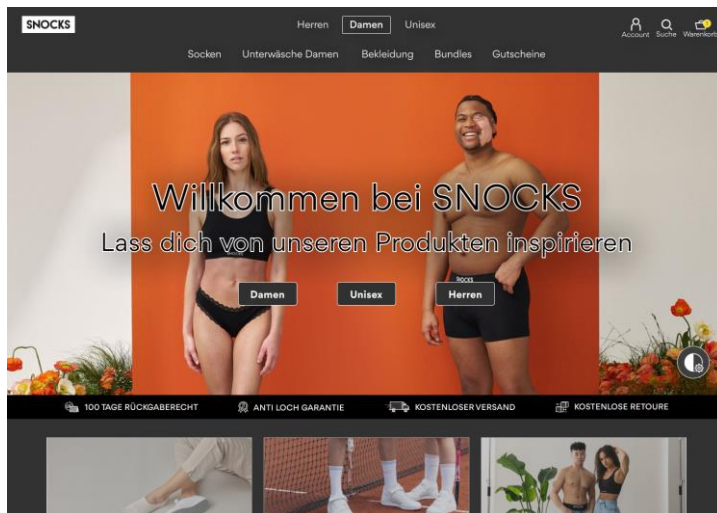


Abbildung 52: Dunkel-Hell Kontrastmodus mit Textgröße M

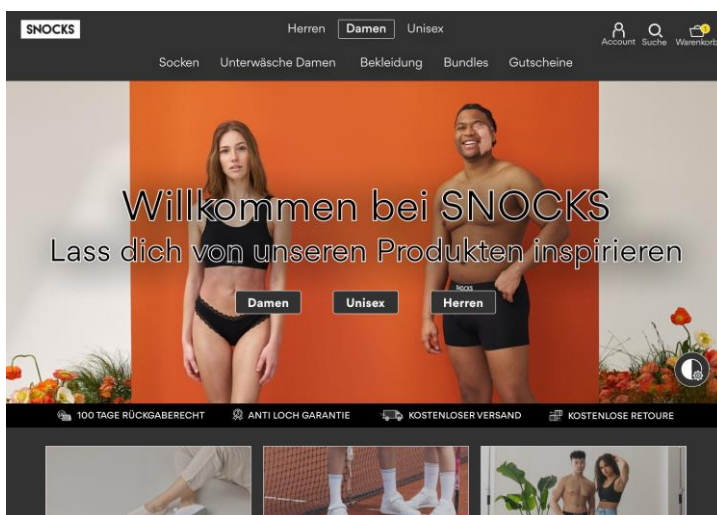


Abbildung 53: Dunkel-Hell Kontrastmodus mit Textgröße L



Abbildung 54: Dunkel-Hell Kontrastmodus mit Textgröße XL



Abbildung 55: Farben Kontrastmodus mit Textgröße S



Abbildung 56: Farben Kontrastmodus mit Textgröße M



Abbildung 57: Farben Kontrastmodus mit Textgröße L



Abbildung 58: Farben Kontrastmodus mit Textgröße XL

7 Literaturverzeichnis

- +sitegeist (Hrsg.). *DIE 7 FAKTOREN, DIE DIE BENUTZERERFAHRUNG BEEINFLUSSEN*. Verfügbar unter: <https://sitegeist.de/blog/usability-user-experience-design/die-7-faktoren-die-die-benutzererfahrung-beeinflussen.html>
- 10X studio (Hrsg.). *Mit Hilfe von Smoke Tests vor der Entwicklungsphase den Produkterfolg testen. Testen Sie das Potenzial Ihrer Geschäftsidee an echten Kunden*. Verfügbar unter: <https://10xstudio.co/smoke-testing/>
- AB Tasty (Hrsg.). *Stichprobenrechner: Sie wollen schnell statistische Zuverlässigkeit erreichen? Finden Sie heraus, wie groß Ihr Testpublikum sein muss. Kein Dokortitel in Mathematik erforderlich*. Verfügbar unter: <https://www.abtasty.com/de/stichproben-rechner/>
- ADVIDERA (Hrsg.). *Customer Journey*. Verfügbar unter: <https://www.advidera.com/glossar/customer-journey/>
- ADVIDERA (Hrsg.). *Multivariate Testing*. Verfügbar unter: <https://www.advidera.com/glossar/multivariate-testing/>
- Aho, M. (CXL, Hrsg.). (2020, 12. Mai). *Tempted to Peek? Why Sequential Testing May Help*. Verfügbar unter: <https://cxl.com/blog/peeking-sequential-testing/>
- Analytics Toolkit (Hrsg.). *State of the art statistics for A/B tests*. Verfügbar unter: <https://www.analytics-toolkit.com/>
- Analytics Toolkit (Hrsg.). *What does "Minimum Detectable Effect" mean? Definition of Minimum Detectable Effect in the context of A/B testing (online controlled experiments)*. Verfügbar unter: <https://www.analytics-toolkit.com/glossary/minimum-detectable-effect/>
- Analytics Toolkit (Hrsg.). *What does "Peeking" mean? Definition of Peeking in the context of A/B testing (online controlled experiments)*. Verfügbar unter: <https://www.analytics-toolkit.com/glossary/peeking/>
- Beauftragter der Bundesregierung für die Belange von Menschen mit Behinderungen (Hrsg.). (2019). *Teilhabempfehlungen Mehr Inklusion wagen!* Verfügbar unter: https://www.behindertenbeauftragter.de/SharedDocs/Downloads/DE/AS/PublikationenErklaerungen/Teilhabempfehlungen.pdf?__blob=publicationFile&v=9
- Becker, N. (Johner Institut, Hrsg.). (2015, 17. Juli). *User Experience ungleich Usability*. Verfügbar unter: <https://www.johner-institut.de/blog/iec-62366-usability/user-experience/>
- Bhandari, P. (Scribbr, Hrsg.). (2020, 17. September). *How to Calculate Standard Deviation (Guide) | Calculator & Examples*. Verfügbar unter: <https://www.scribbr.com/statistics/standard-deviation/#useful>
- BIK BITV Test (Hrsg.). *Prüfschritt 9.1.1.1a. Alternativtexte für Bedienelemente*. Verfügbar unter: <https://ergebnis.bitvtest.de/pruefschritt/bitv-20-web/bitv-20-web-9-1-1-1a-alternativtexte-fuer-bedienelemente>
- BIK BITV Test (Hrsg.). *Prüfschritt 9.1.1.1b. Alternativtexte für Grafiken und Objekte*. Verfügbar unter: <https://ergebnis.bitvtest.de/pruefschritt/bitv-20-web/bitv-20-web-9-1-1-1b-alternativtexte-fuer-grafiken-und-objekte>
- BIK BITV Test (Hrsg.). *Prüfschritt 9.1.3.1a. HTML-Strukturelemente für Überschriften*. Verfügbar unter: <https://ergebnis.bitvtest.de/pruefschritt/bitv-20-web/bitv-20-web-9-1-3-1a-htmlstrukturelemente-fuer-ueberschriften>
- BIK BITV Test (Hrsg.). *Prüfschritt 9.1.3.1b. HTML-Strukturelemente für Listen*. Verfügbar unter: <https://ergebnis.bitvtest.de/pruefschritt/bitv-20-web/bitv-20-web-9-1-3-1b-html-strukturelemente-fuer-listen>
- BIK BITV Test (Hrsg.). *Prüfschritt 9.1.3.1d. Inhalt gegliedert*. Verfügbar unter: <https://ergebnis.bitvtest.de/pruefschritt/bitv-20-web/bitv-20-web-9-1-3-1d-inhalt-gegliedert>

- BIK BITV Test (Hrsg.).. *Prüfschritt 9.1.3.1r. Datentabellen richtig aufgebaut*. Verfügbar unter: <https://ergebnis.bitvtest.de/pruefschritt/bitv-20-web/bitv-20-web-9-1-3-1e-datentabellen-richtig-aufgebaut>
- BIK BITV Test (Hrsg.).. *Prüfschritt 9.1.4.11. Kontraste von Grafiken und grafischen Bedienelementen ausreichend*. Verfügbar unter: <https://ergebnis.bitvtest.de/pruefschritt/bitv-20-web/bitv-20-web-9-1-4-11-kontraste-von-grafiken-und-grafischen-bedienelementen-ausreichend>
- BIK BITV Test (Hrsg.).. *Prüfschritt 9.1.4.3. Kontraste von Texten ausreichend*. Verfügbar unter: <https://ergebnis.bitvtest.de/pruefschritt/bitv-20-web/bitv-20-web-9-1-4-3-kontraste-von-texten-ausreichend>
- BIK BITV Test.. *Prüfschritt 9.2.1.1. Ohne Maus nutzbar*. Verfügbar unter: <https://ergebnis.bitvtest.de/pruefschritt/bitv-20-web/bitv-20-web-9-2-1-1-ohne-maus-nutzbar>
- BIK BITV Test (Hrsg.).. *Prüfschritt 9.2.4.2. Sinnvolle Dokumenttitel*. Verfügbar unter: <https://ergebnis.bitvtest.de/pruefschritt/bitv-20-web/bitv-20-web-9-2-4-2-sinnvolle-dokumenttitel>
- BIK BITV Test (Hrsg.).. *Prüfschritt 9.3.2.3. Konsistente Navigation*. Verfügbar unter: <https://ergebnis.bitvtest.de/pruefschritt/bitv-20-web/bitv-20-web-9-3-2-3-konsistente-navigation>
- Birkett, A. (CXL, Hrsg.). (2022, 13. Dezember). *What is A/B Testing? The Complete Guide: From Beginner to Pro*. Verfügbar unter: <https://cxl.com/blog/ab-testing-guide/>
- Boksch, R. (Statista, Hrsg.). (2020, 2. Dezember). *Barrierefreiheit im Internet kaum vorhanden*. Verfügbar unter: <https://de.statista.com/infografik/23675/anteil-der-majestic-million-websites-der-wcag-fehler-aufweist/>
- Boßow-Thies, S., Hofmann-Stöling, C. & Jochims, H. (Hrsg.). (2020). *Data-driven Marketing. Insights aus Wissenschaft und Praxis* (1. Auflage 2020). Wiesbaden: Springer Fachmedien Wiesbaden.
- Bundesfachstelle Barrierefreiheit (Hrsg.).. *Leitlinien für die Anwendung des Barrierefreiheitsstärkungsgesetzes*. Verfügbar unter: https://www.bundesfachstelle-barrierefreiheit.de/SharedDocs/Downloads/DE/Externe-Veroeffentlichungen/bmas-leitlinien-bfsg.pdf?__blob=publicationFile&v=4
- Bundesfachstelle Barrierefreiheit (Hrsg.). (2021, 22. Juli). *Mehr Barrierefreiheit für Produkte und Dienstleistungen: Das Barrierefreiheitsstärkungsgesetz tritt in Kraft*. Verfügbar unter: <https://www.bundesfachstelle-barrierefreiheit.de/SharedDocs/Kurzmeldungen/DE/barrierefreiheitsstaerkungsgesetz-verkuendet.html>
- Bundesministerium für Justiz (Hrsg.).. *Gesetz zur Gleichstellung von Menschen mit Behinderungen (Behindertengleichstellungsgesetz - BGG) § 4 Barrierefreiheit*. Verfügbar unter: https://www.gesetze-im-internet.de/bgg/_4.html
- DATAtab Team (Hrsg.). (2023a). *Hypothesen*. Verfügbar unter: <https://datatab.de/tutorial/hypothesen>
- DATAtab Team (Hrsg.). (2023b). *Online Statistics Calculator*. Verfügbar unter: <https://datatab.de/tutorial/standardabweichung-varianz-spannweite>
- DEGES, F. (2020). *GRUNDLAGEN DES E-COMMERCE*. [Place of publication not identified]: Springer Gabler.
- DESTATIS Statistisches Bundesamt (Hrsg.). (2022, 22. Juni). *Pressemitteilungen: 7,8 Millionen schwerbehinderte Menschen leben in Deutschland*. Verfügbar unter: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Behinderte-Menschen/_inhalt.html#119290
- Dodt, D. (DM EXCO, Hrsg.). (2020, 9. März). *Erfolgreiche A/B Tests im E-Commerce: Warum es auf präzise Hypothesen ankommt*. Verfügbar unter: <https://dmexco.com/de/stories/erfolgreiche-a-b-tests-im-e-commerce-warum-es-auf-precise-hypothesen-ankommt/>

- Dr. No, Dr. Antwerpes, Frank (DocCheck Flexikon, Hrsg.). *Wahrnehmung*. Verfügbar unter: <https://flexikon.doccheck.com/de/Wahrnehmung>
- DRIP. AGENCY (Hrsg.). *Case Study: 3.1 Mio € zusätzliches Wachstum pro Jahr durch strategisches Testing*. Verfügbar unter: <https://dripagency.de/case-study/snocks/>
- DRIP. AGENCY (Hrsg.). *Homepage*. Verfügbar unter: <https://dripagency.de/>
- Ertel, S. & Venzke-Caprarese, S. (2014). Google Universal Analytics. *Datenschutz und Datensicherheit - DuD*, 38(3), 181–185. <https://doi.org/10.1007/s11623-014-0072-2>
- EUPATI (Hrsg.). *Signifikanzniveau*. Verfügbar unter: <https://toolbox.eupati.eu/glossary/signifikanzniveau/?lang=de>
- Fantaye, E. (Medium, Hrsg.). (2021, 14. August). *A/B testing with Machine Learning*. Verfügbar unter: <https://euelfantaye.medium.com/a-b-testing-with-machine-learning-277da2750123>
- FUSEON (Hrsg.). *CHECKLISTE: Durchführung eines A/B Tests*. Verfügbar unter: <https://fuseon-media.com/checkliste-durchfuehrung-eines-ab-tests/>
- Gast, O. (2018). *User Experience im E-Commerce. Messung von Emotionen bei der Nutzung interaktiver Anwendungen*. Wiesbaden: Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-22484-4>
- Hänsel, F., Baumgärtner, S. D. [Sören D.], Kornmann, J. M. & Ennigkeit, F. (2016). Kognition. In F. Hänsel, S. D. Baumgärtner, J. Kornmann & F. Ennigkeit (Hrsg.), *Sportpsychologie* (Springer-Lehrbuch, S. 23–52). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-50389-8_2
- HDE (Statista, Hrsg.). (2023, 23. Mai). *Umsatz durch E-Commerce (B2C) in Deutschland in den Jahren 1999 bis 2022 sowie eine Prognose für 2023 (in Milliarden Euro)*. Verfügbar unter: <https://de.statista.com/statistik/daten/studie/3979/umfrage/e-commerce-umsatz-in-deutschland-seit-1999/>
- Hemmerich, W. (StatistikGuru, Hrsg.). (2016). *Statistische Power (Teststärke)*. Verfügbar unter: <https://statistikguru.de/lexikon/statistische-power.html>
- HPI ACADEMY (Hrsg.). *Was ist Design Thinking?* Verfügbar unter: <https://hpi-academy.de/design-thinking/was-ist-design-thinking/>
- Juviler, J. (HubSpot, Hrsg.). (2022, 25. August). *What Is GUI? Graphical User Interfaces, Explained*. Verfügbar unter: <https://blog.hubspot.com/website/what-is-gui>
- Knuppe, N. (mollie, Hrsg.). *Mit A/B Testing zur Conversion-Rate-Optimierung: ein Guide für E-Commerce-Unternehmen*. Verfügbar unter: <https://www.mollie.com/de/growth/ab-testing-im-e-commerce>
- Lapp, J. (HubSpot, Hrsg.). *Checkliste: Wie Sie A/B-Tests vor, während und nach der Datenerfassung durchführen*. Verfügbar unter: <https://blog.hubspot.de/marketing/a-b-test-checkliste>
- Lecturio (Hrsg.). *Statistische Power: Stärke eines Tests*. Verfügbar unter: <https://www.lecturio.de/artikel/medizin/statistische-power-starke-eines-tests/#:~:text=Sie%20hei%C3%9Ft%20auch%20Fehler%20weiter%20Art.&text=%CE%B2%20steht%20in%20direktem%20Zusammenhang,%20%20%25%20%3D%200%2C2.>
- Lemodo (Hrsg.). *ONLINE MARKETING GLOSSAR*. Verfügbar unter: <https://www.lemodo.de/online-marketing-glossar/#:~:text=Die%20Abk%C3%BCrzung%20%E2%80%99ECR%E2%80%99C%20steht%20f%C3%BCr,oder%20in%20einer%20Marketingkampagne%20misst.>

- LinkedIn & Puscher, F. (Autor). (2015). *A/B-Testing - Grundlagen*. Verfügbar unter: <https://www.linkedin.com/learning/a-b-testing-grundlagen/testing-ist-die-qualitatssicherung-in-e-commerce-und-online-marketing?resume=false&u=82266642>
- Looschelders, T. (2023). *Conversion-Optimierung. Über 150 Praxistipps Zu Datengetriebenem Marketing, Analytics and Webseitenoptimierung*. Wiesbaden: Springer Fachmedien Wiesbaden GmbH.
- Me&company (Hrsg.). *UX Research: Leitfaden für User Research 2023*. Verfügbar unter: <https://www.me-company.de/magazin/ux-research/>
- Müller, E. (webit!, Hrsg.). (2022, 9. November). *Statistischer Rückenwind - Stichprobengrößen im A/B-Testing*. Verfügbar unter: <https://www.webit.de/blog/2022/11/09/stichproben-ab>
- Oestreich, M. & Romberg, O. (2009). *Keine Panik vor Statistik! Erfolg und Spaß im Horrorfach nichttechnischer Studiengänge*. Vieweg+Teubner.
- Ollmann, M. (HubSpot, Hrsg.). (2020, 20. Mai). *ARPU: So berechnen Sie den Average Revenue per User*. Verfügbar unter: <https://blog.hubspot.de/service/arpur#:~:text=Was%20ist%20der%20ARPU%3F,und%20%2Dgewinn%20ih-rer%20Organisation%20beitragen>
- ONLINEMARKETING.DE (Hrsg.). *Monitoring*. Verfügbar unter: <https://onlinemarketing.de/lexikon/definition-monitoring>
- Optimizely (Hrsg.). *A/B/N-Testing*. Verfügbar unter: <https://www.optimizely.com/de/optimization-glossary/abn-testing/>
- Pekelis, L. (Optimizely, Hrsg.). (2015, 20. Januar). *Statistics for the internet age: The story behind Optimizely's new stats engine*. Verfügbar unter: <https://www.optimizely.com/de/insights/blog/statistics-for-the-internet-age-the-story-behind-optimizelys-new-stats-engine/>
- Pierce, R. & Wu, C. (split, Hrsg.). (2023, 22. Juni). *Sequential Vs Fixed Horizon*. Verfügbar unter: <https://www.split.io/blog/sequential-vs-fixed-horizon/>
- Prakash, A. (Amplitude, Hrsg.). (2022, 9. Juni). *Sequential Test vs. Fixed Horizon T-Test: When to Use Each? Learn whether a fixed horizon T-test or sequential testing is right for your next experiment*. Verfügbar unter: <https://amplitude.com/blog/sequential-test-vs-t-test-when-to-use-each?learn-whether-a-fixed-horizon-t-test-or-sequential-testing-is-right-for-your-next-experiment>
- Puppe, M. (alkima, Hrsg.). (2022, 25. August). *Ein Muss für Ihren Onlineshop – Qualitätssicherung (QA)*. Verfügbar unter: <https://www.alkima.de/blog/onlineshop-qualitaetssicherung>
- ResearchGate (Hrsg.). *Morvilles User Experience Honeycomb*. Verfügbar unter: https://www.researchgate.net/figure/Morvilles-User-Experience-Honeycomb-35-Useful-fit-for-practical-use-in-the-clinical_fig4_50347626/download
- SNOCKS (Hrsg.). *Homepage*. Verfügbar unter: <https://snocks.com/>
- Split (Hrsg.). *Monitor and experiment settings*. Verfügbar unter: <https://help.split.io/hc/en-us/articles/360020640752-Monitor-and-experiment-settings#using-fixed-horizon-testing>
- Springer Gabler. *Key Performance Indicator (KPI). Ausführliche Definition im Online-Lexikon*. Verfügbar unter: <https://wirtschaftslexikon.gabler.de/definition/key-performance-indicator-kpi-52670>
- Steidle, R. & Pordesch, U. (2008). Im Netz von Google. Web-Tracking und Datenschutz. *Datenschutz und Datensicherheit - DuD*, 32(5), 324–329. <https://doi.org/10.1007/s11623-008-0078-8>
- Steinicke, F. & Wittenburg, K.. *Informatik im Kontext 1. Grundlagen der Mensch-Computer-Interaktion*. Verfügbar unter: <https://www2.informatik.uni-hamburg.de/fachschaft/wiki/images/f/fb/IKON1-Skript.pdf>
- Still, H. (Scribbr, Hrsg.). (2021, 25. November). *Das Signifikanzniveau einfach erklärt + Beispiel*. Verfügbar unter: <https://www.scribbr.de/statistik/signifikanzniveau/>

- Stotz, N. (2022). *Experimentelle Produktentwicklung. Wie Unternehmen ihre Strategien systematisch validieren können*. Berlin, Heidelberg: Springer Gabler.
- Studyflix (Hrsg.). *Fehler 1. Art (Alphafehler)*. Verfügbar unter: <https://studyflix.de/statistik/fehler-1-art-alphafehler-1797>
- SurveyMonkey (Hrsg.). *A/B-Test-Rechner. Sind Ihre Ergebnisse statistisch signifikant?* Verfügbar unter: <https://www.surveymonkey.de/mp/ab-testing-significance-calculator/>
- Versa commerce (Hrsg.). (2023, 2. Maia). *Angebotsknappheit*. Verfügbar unter: <https://www.versacommerce.de/glossar/angebotsknappheit>
- Versa commerce (Hrsg.). (2023, 2. Maib). *KPI-Monitoring*. Verfügbar unter: <https://www.versacommerce.de/glossar/kpi-monitoring>
- Vieritz, H. (2015). *Barrierefreiheit im virtuellen Raum. Benutzungszentrierte und modellgetriebene Entwicklung von Weboberflächen*. Wiesbaden: Springer Fachmedien.
- Voxco (Hrsg.). *Einführung in unabhängige Variablen und abhängige Variablen*. Verfügbar unter: <https://www.voxco.com/de/blog/einfuehrung-in-unabhaengige-variablen-und-abhaengige-variablen/>
- WebAIM (Hrsg.). *The WebAIM Million. The 2023 report on the accessibility of the top 1,000,000 home pages*. Verfügbar unter: <https://webaim.org/projects/million/>
- Wenz, C. & Hauser, T. (2015). *Websites optimieren - Das Handbuch*. Wiesbaden: Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-07262-9>
- Wesolko, D. (Medium, Hrsg.). (2016, 15. Juni). *Peter Morville's User Experience Honeycomb*. Verfügbar unter: <https://danewesolko.medium.com/peter-morvilles-user-experience-honeycomb-904c383b6886>
- Witzenleiter, M. (2021). *Quick Guide A/B Testing*. Wiesbaden: Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-34649-2>
- Xovi (Hrsg.). *xovi Glossar. Was ist Display Marketing?* Verfügbar unter: <https://www.xovi.de/was-ist-display-marketing/>