Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

Wirtschaftlichkeit von Usability-Evaluationsmethoden

MASTER-ARBEIT

2023

Autor
Honegger, Luca

Begleitperson
Prof. Dr. Sonderegger, Andreas

Praxispartner*in
sinnhaft GmbH
Kontaktperson Meyer, Martina

ZUSAMMENFASSUNG

In der vorliegenden Masterarbeit wird die Wirtschaftlichkeit von drei Usability-Evaluationsmethoden (UEM) untersucht. Im Rahmen eines Feldexperiments wurden die folgenden UEM miteinander verglichen: Heuristische Evaluation (HE), moderierter Usability-Test (UTM) sowie automoderierter Usability-Test (UTA) (N = 5 Expert*innen sowie N = 44 Testpersonen). Dabei wurden die Fragen beantwortet, inwiefern sich die gefundenen Usability-Probleme je nach Methode unterscheiden und was das Auffinden eines einzelnen Problems kostet.

Die Resultate zeigen, dass in UTM die meisten Probleme identifiziert werden. Die gefundenen Probleme der heuristischen Evaluation waren hingegen schwerwiegender. Die Kosten pro identifiziertes Problem sind beim automoderierten Ansatz am tiefsten.

Neben falschen Alarmen (False Positives) und verpassten Problemen (False Negatives) stellt auch die tiefe Reliabilität von Expert*innen und Evaluator*innen ein Risiko für Unternehmen dar, dem Beachtung geschenkt werden sollte.

Stichworte: Usability, Evaluationsmethoden, Wirtschaftlichkeit, Usability-Tests, heuristische Evaluation

ABSTRACT

This thesis investigates the economic efficiency of usability evaluation methods (UEM). The three UEMs, namely heuristic evaluation (HE), moderated usability testing (UTM), and auto-moderated usability testing (UTA), were compared in the context of a field experiment ($N = 5$ experts and $N = 44$ participants). The research questions asked whether the usability problems found differed depending on the method used and what the cost was to identify a single problem.

The results show that moderated usability tests identify the most problems. However, the problems found using heuristic evaluation were more serious. The costs per identified problem are lowest for the auto-moderated approach. In addition to false alarms ("false positives") and missed problems ("false negatives"), the low reliability of experts and evaluators also represents a risk that companies should take into account.

Keywords: Usability, evaluation methods, cost-effectiveness, usability testing, heuristic evaluation.

Für Giulia und Finca

Danke

Andreas für deine Begleitung und Inspiration

Elena, Martina, Melanie, Maximilian, Nathanel, Myvienne und Simon für eure grosse Arbeit

Stephan, Melanie und Janis von professional.ch für euer Vertrauen und die Grosszügigkeit

Martina von sinnhaft, für dein Vertrauen und unsere Vision

Jonas von Solid für deine Vision und Leidenschaft

GLOSSAR

Avatar	symbolisches Abbild eines Menschen, z. B. als Bild oder Animation
B2B	Business-to-Business bezeichnet geschäftliche Beziehungen zwischen Unternehmen
Customer-Journey	gesamter Lebenszyklus, den Kund*innen durchlaufen, wenn sie mit einem Unternehmen oder Marke interagieren
Evaluator*in	Evaluator*innen werten die Aufzeichnungen eines Usability-Tests aus
Expert*in	Expert*innen führen eine heuristische Evaluation durch
HE	heuristische Evaluation
NAP	New Application Process (Testobjekt der vorliegenden Studie)
PANAS	Positive and Negative Affect Schedule (PANAS). Fragebogen zur Messung des emotionalen Affekts
PSSUQ	Post-Study System Usability Questionnaire (PSSUQ). Fragebogen zur Messung der subjektiven Zufriedenheit
TLX	NASA Task Load Index. Fragebogen zur Messung der koginitiven Belastung
UEM	Usability Evaluationsmethoden; eine Untergruppe der User-Research-Methoden
URM	User-Research-Methoden
UTA	automoderiertes Usability-Testing
UTM	moderiertes Usability-Testing
UX-Professionals	Personen, die sich beruflich mit der Thematik UX bzw. Usability auseinandersetzen

INHALTSVERZEICHNIS

1 EINLEITUNG

In der vorliegenden Masterarbeit geht es um die Wirtschaftlichkeit von Methoden zur Überprüfung der Benutzerfreundlichkeit eines Produkts, beispielsweise einer Webseite. Dabei wurde untersucht, inwiefern sich diese Methoden hinsichtlich ihrer Wirksamkeit und Effizienz unterscheiden. In den folgenden Abschnitten werden zunächst die Ausgangslage und die Problemstellung erläutert. Es wird dann ein Bezug zur wissenschaftlichen Relevanz hergestellt und abschliessend werden die Fragestellungen vorgestellt.

1.1 Ausgangslage und Problemstellung

Um den langfristigen Erfolg einer Dienstleistung oder eines Produkts sicherzustellen, müssen neben betriebswirtschaftlichen Aspekten auch die wachsenden Anforderungen von Kund*innen berücksichtigt und befriedigt werden. Nach Gebauer, Krempl und Fleisch (2008) bedingt dies ein strukturiertes Vorgehen mit den Schritten der Ideengenerierung, der Konzeptionierung sowie der Validierung. Dabei bedarf es einer gezielten und regelmässigen Einbindung von Nutzenden, z. B. durch Befragungen oder Markt- bzw. Kundentests.

Die Norm *DIN EN ISO 9241-210* (Deutsches Institut für Normung [DIN], 2020a) definiert den *menschenzentrierten Gestaltungsansatz* (engl. *Human Centered Design Process*, HCD), wonach zunächst der Nutzungskontext verstanden und daraus die Anforderungen der Nutzer*innen abgeleitet werden müssen. Auf dieser Basis kann eine passende Lösung gestaltet werden, die dann gemeinsam mit (potenziellen) Nutzer*innen evaluiert wird. Dieser Ablauf ist in der Abbildung 1 dargestellt.

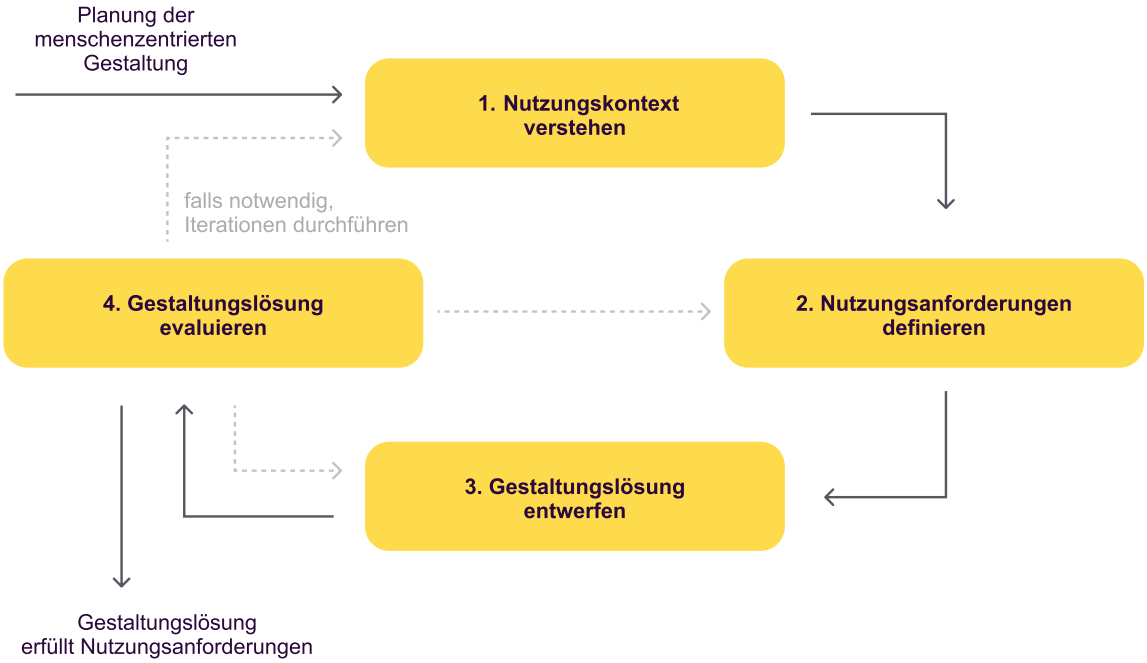


Abbildung 1. Menschenzentrierter Gestaltungsansatz nach DIN EN ISO 9241-210 (DIN, 2020a). Eigene Darstellung.

Die Praxispartnerin für die vorliegende Masterthesis ist die *sinnhaft GmbH*, ein Beratungsunternehmen im Bereich User-Experience (UX) mit nationalen Business-to-Business (B2B)-Kunden mit Sitz in Zürich. Ihre Dienstleistungen umfassen die Beratung, die Konzeption und die Validierung von digitalen Produkten und Dienstleistungen mit einem Schwerpunkt auf Usability und UX. Es sei an dieser Stelle darauf hingewiesen, dass der Autor Mitinhaber und Co-Geschäftsführer der sinnhaft GmbH ist, für die vorliegende Masterarbeit jedoch nicht in dieser Rolle fungiert hat.

Die sinnhaft GmbH bietet ihre Dienstleistungen entlang dieses menschenzentrierten Gestaltungsprozesses an. Bei Beratungen wurde festgestellt, dass insbesondere beim Schritt der Evaluation vermehrt Diskussionen bezüglich der Wirtschaftlichkeit der Usability-Methoden entstehen. Kund*innen äussern zwar das Bedürfnis nach einer Evaluation, gehen jedoch oftmals davon aus, dass die Meinung von Expert*innen ausreichend und auch kostengünstiger sei. Gestützt auf den menschenzentrierten Ansatz empfiehlt die sinnhaft GmbH das Miteinbeziehen von (potenziellen) Nutzer*innen, was jedoch oft mit höheren Kosten verbunden ist. Diese Mehraufwände und Kosten

werden seitens der Kund*innen in Frage gestellt und es besteht Unsicherheit, ob diese Aktivitäten einen Mehrwert bringen. Es wird seitens der Kund*innen argumentiert, dass die eigene Zielgruppe bereits gut bekannt sei und Untersuchungen mit Endkund*innen werden daher als unnötig betrachtet. Dieses Phänomen ist in der UX-Branche bekannt unter dem Motto «You are not the user» (Budiu, 2017). Dies kann als ein *False-Consensus-Effekt* (Ross, Greene & House, 1977) gesehen werden, wonach Unternehmen davon ausgehen, dass die eigenen Meinungen und Überzeugungen auch von anderen Personen geteilt werden.

Daraus entstand bei der sinnhaft GmbH das Bedürfnis, Argumente zu erarbeiten, die Kosten und den Nutzen von Usability-Evaluationen besser zu veranschaulichen. Die vorliegende Erhebung wurde in Zusammenarbeit mit *professional.ch* durchgeführt, ein Unternehmen der *Yousty AG*. Diese bietet Schweizer Firmen verschiedene Angebote im Bereich Rekrutierung sowie *Employer Branding* (dt. Arbeitgebermarkenbildung) an. Die Dienstleistungen beinhalten Massnahmen, um ein Unternehmen als attraktiv für Mitarbeitende darzustellen, die Effizienz der Personalrekrutierung zu steigern, die Qualität der Bewerber*innen langfristig zu erhöhen sowie die emotionale Bindung an das Unternehmen zu stärken (Tomczak, Walter & Henkel, 2011). Zu den Dienstleistungen von professional.ch gehört eine Online-Stellenvermittlungsplattform mit Inseraten. Durch diese Plattform soll die Integration von jungen Berufsleuten beschleunigt werden. Dabei verfolgt das Unternehmen die Vision, einen Beitrag zur Senkung der Jugendarbeitslosigkeit zu leisten (professional.ch, 2023). Die Yousty AG entwickelt einen neuen Online-Bewerbungsprozess für ihre Plattform professional.ch, der als *New Application Process* (NAP) bezeichnet wird. Er dient als Testobjekt für die vorliegende Arbeit.

Die sinnhaft GmbH ist weiter eine Partnerschaft mit der Schweizer Softwareherstellerin *Solid Technologies AG* eingegangen. Diese entwickelt die Plattform *Solid User Tests*, mit der Usability-Tests automatisiert durchgeführt werden können. Neuartig dabei ist, dass ein Avatar die Testsitzungen moderiert.

1.2 Wissenschaftliche und praktische Relevanz

Es existieren wenig aktuelle wissenschaftliche Arbeiten in Bezug auf die Wirtschaftlichkeit (Kosten-Nutzen-Analyse) sowie die Vergleichbarkeit der Aussagekraft von Usability-Methoden, wie in einer Übersichtsarbeit von Rajanen (2020) gezeigt wird. Das letzte publizierte Buch, das sich diesem Thema widmet, stammt von Bias und Mayhew (2005) und trägt den Titel «Cost-Justifying Usability: An Update for the Internet-Age (Interactive Technologies)». Arbeiten von Rajanen und Iivari (2007) und Rajanen (2007) weisen auf den Mangel der Auseinandersetzung mit der Thematik der Wirtschaftlichkeit in Zusammenhang mit Usability hin.

Aus Sicht der praktischen Relevanz sei nochmals darauf hingewiesen, dass Unternehmen die wachsenden Kundenbedürfnisse abdecken müssen, um den langfristigen Erfolg von Produkten und Dienstleistungen zu sichern (Gebauer et al., 2008). Bezugnehmend auf den menschenzentrierten Gestaltungsansatz müssen somit wirtschaftliche sowie psychologische Aspekte betrachtet und miteinander verknüpft werden.

Gemäss Rajanen (2007) haben nur wenige Organisationen im Bereich der Softwareentwicklung Usability-Aktivitäten als einen integralen Bestandteil ihres Entwicklungsprozesses etabliert. Als mögliche Ursache dafür nennt Rajanen, dass die Kosten und der Nutzen für das Management nicht ausreichend transparent sind.

1.3 Zielsetzung

Aus der beschriebenen Ausgangslage, der Problemstellung und den identifizierten Forschungslücken gehen folgende zwei übergeordnete Ziele für diese Arbeit hervor:

1. Für die sinnhaft GmbH sollen Erkenntnisse in Bezug auf die Kosten und Nutzen von Usability-Evaluationsmethoden gewonnen werden.
2. Für die Yousty AG soll eine Entscheidungsgrundlage geschaffen werden, welche Teile und Funktionen von professional.ch überarbeitet sowie weiterentwickelt werden sollen.

1.4 Fragestellungen

Basierend auf dem geschilderten Forschungsinteresse und in Absprache mit der Praxispartnerin wurden zwei Fragestellungen festgelegt, durch die der inhaltliche sowie der finanzielle Unterschied der Usability-Evaluationsmethoden näher beleuchtet werden soll.

Die Praxispartnerin hat diejenigen Methoden bestimmt, die für sie von grösstem Interesse sind und die am besten ihren praktischen Alltag widerspiegeln: *Heuristische Evaluation* (HE) sowie *moderiertes Usability-Testing* (UTM). Diese gehören zugleich allgemein zu den am häufigsten angewendeten Methoden (Fernandez, Insfran & Abrahão, 2011; Law & Hvannberg, 2004). Aufgrund der strategischen Zusammenarbeit mit der Solid Technologies AG möchte die sinnhaft GmbH zudem den neuen Ansatz für *automoderierte Usability-Tests* (UTA) von Solid genauer analysieren und Wissen in Bezug auf die Effektivität sowie die Effizienz im praktischen Einsatz aufbauen. Daher wurden diese drei Usability-Evaluationsmethoden für die vorliegende Studie von der Praxispartnerin explizit so gewünscht. Die Methoden werden im Kapitel 2.2 näher beschrieben.

In der ersten Fragestellung dieser Masterarbeit geht es um den inhaltlichen Unterschied der Usability-Evaluationsmethoden. Laut Koutsabasis, Spyrou und Darzentas (2007) sind relevante Ergebnisse einer Usability-Evaluationsmethode die tatsächlichen Probleme, die bei Nutzer*innen auftreten und sich durch einen höheren Schweregrad auszeichnen. Der Schweregrad drückt aus, wie gross der negative Einfluss eines Problems auf die Usability ist. Aus diesem Grund lautet die erste Fragestellung:

Fragestellung 1: Inwiefern unterscheiden sich die gefundenen Usability-Probleme, die durch die verschiedenen Usability-Methoden erhoben wurden, in Bezug auf ihre Relevanz?

Nachdem die inhaltlichen Unterschiede der verschiedenen Usability-Evaluationsmethoden beleuchtet wurden, soll in der zweiten Fragestellung die finanzielle Seite näher betrachtet werden. Aus ökonomischen und praktischen Gründen werden heute Usability-Tests vermehrt in Form eines sogenannten Remote-Testing (z. B. via Videokonferenztools) durchgeführt (Chynał & Sobecki, 2015). Dadurch ergeben sich verschiedene Vorteile, wie das Wegfallen von Anfahrtswegen und der

aufwändigen Infrastruktur eines Usability-Labors oder eine flexiblere Einbettung in den Arbeitsalltag der Testpersonen. Daraus stellt sich die Frage, wie sich solche Veränderungen auf die Kosten von Usability-Evaluationen auswirken.

Fragestellung 2: Was kostet das Auffinden eines einzelnen Usability-Problems je nach Methode?

1.5 Aufbau der Arbeit

Zunächst werden in Kapitel 2 die theoretischen Grundlagen erläutert, die für die vorliegenden Fragestellungen relevant sind. Dabei werden die Begriffe Usability und User Experience definiert, die Evaluationsmethoden vorgestellt, was Usability-Probleme sind und wie diese klassifiziert werden, wie typischerweise die Anzahl von Testpersonen in Usability-Tests anhand der branchenüblichen Regel *Magic Number 5* bestimmt wird sowie was Wirtschaftlichkeit im Kontext von Usability und UX bedeutet. In Kapitel 3 wird das methodische Vorgehen beschrieben. Hierzu werden die Hypothesen hergeleitet und das Forschungsdesign erläutert. Zudem wird das Testobjekt New Application Process (NAP) vorgestellt, die Operationalisierung und die Stichprobenplanung sowie -auswahl dargestellt. Weiter werden die Vorbereitungen, Pretests, die Durchführung und die Datenaufbereitung und -auswertung sowie die verwendeten statistischen Methoden ebenfalls erläutert.

Die Ergebnisse werden schliesslich in Kapitel 4 entlang der Fragestellungen bzw. Hypothesen dargestellt. In Kapitel 5 wird die Diskussion und der Ausblick präsentiert. Hier werden die Ergebnisse zusammengefasst und interpretiert. Es werden Handlungsempfehlungen sowie eine kritische Reflexion und Limitierungen gegeben. Die Masterarbeit wird mit dem Kapitel 6 abgeschlossen, in dem ein Fazit und Ausblick gegeben wird.

2 THEORETISCHE GRUNDLAGEN

Im Folgenden werden die relevanten theoretischen Konzepte erläutert, die zur Beantwortung der Fragestellungen dienen. Zunächst werden die Begriffe *Usability* (dt. Gebrauchstauglichkeit) und *User Experience* (UX) (dt. Nutzererlebnis) erläutert. Es wird diskutiert, wie Usability und UX gemessen werden und wie ein Usability-Problem definiert wird. Abschliessend wird der Aspekt der Wirtschaftlichkeit dargelegt und in den vorliegenden Kontext eingeordnet.

2.1 Usability und User Experience (UX)

Es existiert eine Vielzahl an verschiedenen Definitionen der Begriffe Usability und UX. Die Usability betrifft die Gebrauchstauglichkeit von Produkten und Dienstleistungen, während es bei der UX um das Gesamterlebnis der Nutzung eines Systems geht. Die beiden Begriffe werden oft synonym verwendet, weshalb es zentral ist, sie zu unterscheiden (Richter & Flückiger, 2016; Sauer, Sonderegger & Schmutz, 2020). Aufgrund der weiten Verbreitung werden in dieser Arbeit die englischen Begriffe verwendet.

Die *International Organization for Standardization* (ISO) definiert in der Norm 9241-11 den Begriff Usability mit den folgenden drei Hauptmerkmalen (2020b):

1. *Effektivität*: Wie gut kann ein Ziel erreicht werden?
2. *Effizienz*: Wie viele Ressourcen sind nötig, dieses Ziel zu erreichen?
3. *Zufriedenheit*: Wie zufrieden sind Nutzende bei der Bearbeitung der Aufgabe?

Richter und Flückiger (2016) betonen, dass eine gute Usability mehr als nur die Eigenschaft eines Produkts ist. Nutzer*innen haben eine Aufgabe, die sie bewältigen müssen. Dazu verwenden sie ein entsprechendes Werkzeug (beispielsweise ein Produkt oder ein System). Der jeweilige Kontext beeinflusst diese Nutzung. Folgendes Beispiel veranschaulicht diese Zusammenhänge: Ein Hammer ist ein geeignetes Werkzeug, um einen Nagel in die Wand zu schlagen, sofern dies durch eine fähige Person getan wird. Hingegen eignet sich der Hammer nicht, um Papier zu heften, oder ist ineffektiv, wenn er einem Kleinkind in die Hände gegeben wird. Abschliessend ist die Betrachtung des Kontexts

ausschlaggebend. So kann es sein, dass die Wand zu dünn ist, um einen Nagel einzuschlagen. Damit wäre ein selbstklebender Haken zielführender und das eigentliche Vorhaben obsolet.

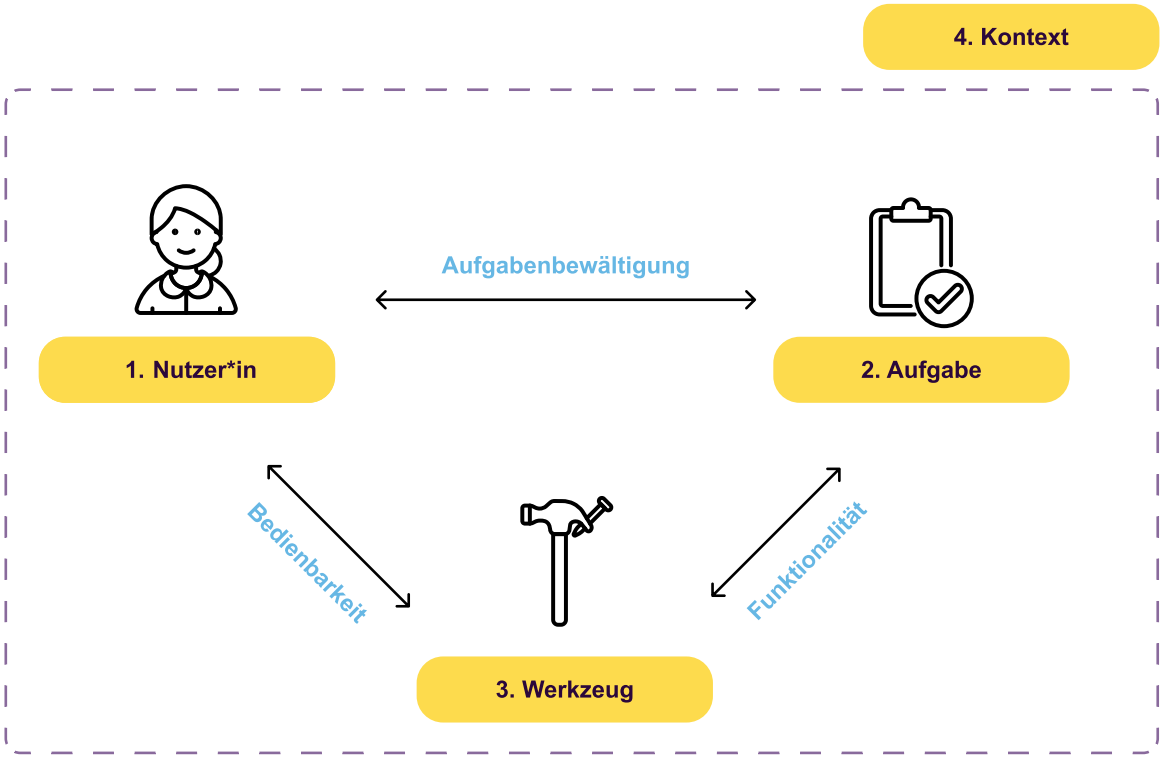


*Abbildung 2. Usability als Zusammenhang zwischen Nutzer*in, Aufgabe und Werkzeug im Kontext nach Sarodnick und Brau (2011). Eigene Darstellung.*

Der Begriff UX fasst eine breitere Perspektive und bezieht die Gedanken, Gefühle und Wahrnehmungen von Nutzenden mit ein. Nach Albert und Tullis (2022) sind die wesentlichen Merkmale von UX, dass erstens Nutzer*innen involviert sind, dass diese zweitens mit einem Produkt bzw. System interagieren, und dass drittens das Erlebnis der Nutzer*innen von Interesse ist sowie dass dieses beobachtbar bzw. messbar ist. Albert und Tullis (2022) definieren den Unterschied zwischen Usability und UX wie folgt:

> *Usability* is usually considered the ability of the user to use the product to carry out a task successfully, whereas *user experience* takes a broader view, looking at the individual's

> entire interaction with the product, as well as the thoughts, feelings, and perceptions that result from that interaction. (S.5)

Kund*innen interagieren mit einem Unternehmen heute oftmals über eine Benutzerschnittstelle (engl. *Interface*), z. B. über Webseiten, Onlineshops, Apps oder Ähnliches. Diese Interaktion beeinflusst wiederum wesentlich die Wahrnehmung des Unternehmens. Raskin (2000) sagt dazu Folgendes: «As far as the customer is concerned, the interface is the product.» (S. 5) Aus Sicht der Kund*innen spielt es keine Rolle, warum und welche Prozesse im Hintergrund ablaufen, solange sie ihre Aufgabe effizient, effektiv und zufriedenstellend erledigen können.

Es ist darauf hinzuweisen, dass Usability je nach Anwendungsbereich anders verstanden und gemessen werden kann. So ist z. B. in einem Onlineshop das rasche Auffinden von Produkten relevant, hingegen im Bereich des öffentlichen Verkehrs, etwa bei Bussen die Verhinderung von Sturzunfällen bei der An- und Abfahrt (Mator et al., 2021).

Sauer et al. (2020) fassen die unterschiedlichen Perspektiven und Definitionen von Usability bzw. UX zusammen und zeigen auf, welche messbaren Dimensionen für diese Konstrukte bestehen. Für die vorliegende Studie wurden daraus die Dimensionen (1) *Effectiveness,* (2) *Efficiency*, (3) *Satisfaction,* (4) *Workload* und (5) *Affect* als relevant erachtet. Bei der (1) Effectiveness (dt. Effektivität) ist die Frage, wie gut eine Aufgabe gelöst werden kann. Dies kann z. B. mittels einer Erfolgsquote gemessen werden. Bei der (2) Efficiency (dt. Effizienz) geht es darum, mit möglichst wenig Ressourcen die Aufgaben erledigen zu können. Eine Möglichkeit, dies zu messen, ist die Bearbeitungszeit. Zur Messung der (3) Satisfaction (dt. Zufriedenheit) steht eine Vielzahl an validierten Fragebögen zur Auswahl. Eines der beliebtesten Instrumente ist aufgrund ihrer Einfachheit und Vergleichbarkeit der Fragebogen *System Usability Scale* (SUS). Weitere bekannte Fragebögen nach Sauro (2016) sind *Questionnaire for User Interface Satisfaction* (QUIS), *Software Usability Measurement Inventory* (SUMI), *Usability Metric for User Experience* (UMUX) oder der *Post-Study System Usability Questionnaire* (PSSUQ), welcher in der vorliegenden Studie zum Einsatz kommt. Die Begründung der Wahl des PPSUQ sowie die detaillierte Erläuterung erfolgen im Methodenteil in Kapitel 3. Die (4) Workload (dt. kognitive Belastung) ist insofern relevant, da diese einen mindernden

Einfluss auf die Leistung bzw. auf die Aufgabenbewältigung haben kann (Law & Hvannberg, 2004). Im Kontext von Usability wird zur Messung der Workload oft der Fragebogen *NASA Task Load Index* (NASA-TXL) verwendet (Baumgartner, Sonderegger & Sauer, 2017; Georgsson, 2019). Zur Messung der (5) emotionalen Zustände (Affect) kommt häufig der *Positive and Negative Affect Schedule* (PANAS) zum Einsatz. Der PANAS eignet sich für unterschiedliche Anwendungsbereiche, in denen Emotionen von Interesse sind, wie für die klinische Anwendung (Breyer & Bluemke, 2016), aber auch für Usability-Studien (Sauer et al., 2019).

Zusammenfassend lässt sich sagen, dass Usability die Gebrauchstauglichkeit eines interaktiven Systems beschreibt und UX einen breiteren Blickwinkel auf das Gesamterlebnis bedeutet. Beide Konstrukte können über verschiedenen Dimensionen gemessen werden, für das es spezifische Methoden gibt, die im Folgenden beschrieben werden.

2.2 Usability-Evaluationsmethoden

Es existiert eine grosse Anzahl an Methoden zur Evaluation der Usability. So identifizieren Vermeeren et al. (2010) insgesamt 96 verschiedene Evaluationsmethoden, die je nach Fragestellung, Phase in der Produktentwicklung und anderen Kriterien eingesetzt werden können. Pauschal kann in zwei Arten von Methoden unterschieden werden. Zunächst gibt es die sogenannte *Usability-Inspektion*, bei der Expert*innen das System evaluieren. Auf der anderen Seite stehen die *Usability-Tests*, bei denen Nutzer*innen eingebunden werden (Madan & Kumar, 2012; UXQB e.V., 2020).

Die Wahl einer passenden Usability-Methode ist jedoch nicht nur vom Aufwand und dem Anspruch auf Zuverlässigkeit abhängig. Weitere Einflussfaktoren sind unter anderem die Projektphase, die zeitlichen Anforderungen an die Verfügbarkeit der Resultate, die Einsatzmöglichkeiten (z. B. im Nutzungskontext oder unter Laborbedingungen), die Anforderung an den Stand der Entwicklung (z. B. Prototyp oder bereits funktionsfähiges Produkt), der Art der Datenerhebung (qualitativ oder quantitativ) oder die technische Anwendungsmöglichkeit (mobile Geräte, Desktopsysteme etc.) (UXQB e.V., 2020; Vermeeren et al., 2010).

Zur Präzisierung wird zwischen den Begriffen *User-Research-Methoden* (URM) und *Usability-Evaluationsmethoden* (UEM) unterschieden. Der Begriff URM bezieht sich auf alle Methoden, die im Rahmen des menschenzentrierten Gestaltungsansatzes dazu dienen, Erkenntnisse zu gewinnen, die ein tieferes Verständnis für die Bedürfnisse, für die Motivation und für die Verhaltensweisen von Nutzenden ermöglichen. Dazu gehören z. B. Methoden wie *Contextual Inquiry*, Fokusgruppen, Tagebuchstudien usw. UEM werden als spezifische Methoden verstanden, die bei der Evaluation der Gestaltungslösung zum Einsatz kommen. Sie sind in diesem Sinne eine Unterkategorie der URM. Beispiele hierzu sind die HE, die Usability-Tests, das Eye-Tracking usw.

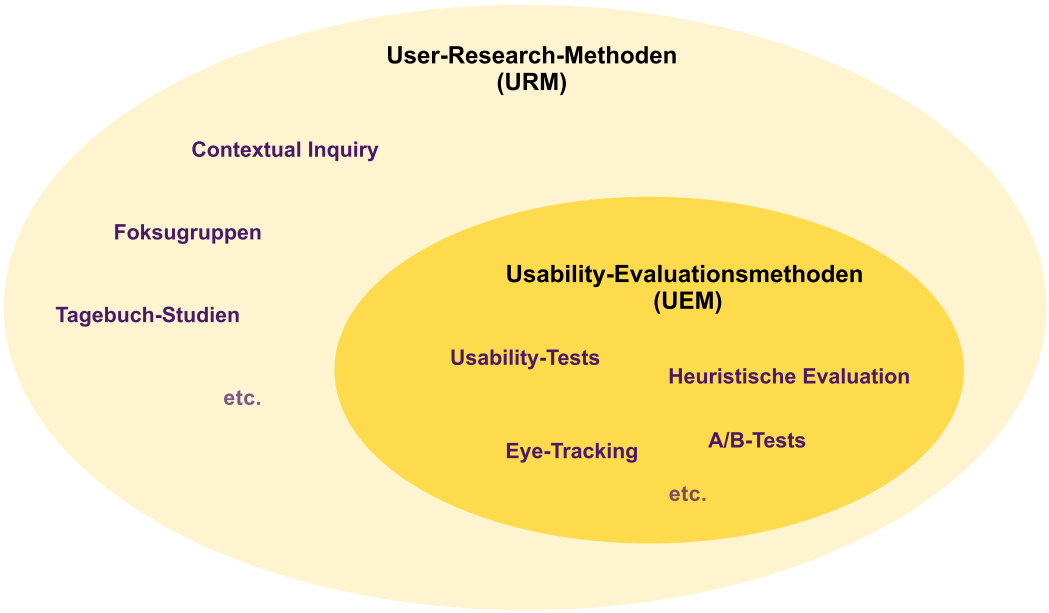


Abbildung 3. Beispielhafte, nicht abschliessende Darstellung des Unterschieds zwischen URM und UEM. Eigene Darstellung.

Nachfolgend werden jene drei UEM ausführlicher beschrieben, die in der vorliegenden Studie zum Einsatz kommen.

2.2.1 Heuristische Evaluation (HE)

Die Methode HE wurde von Nielsen und Molich (1990) vorgestellt, um die Benutzerfreundlichkeit von Systemen zu bewerten. Sie gilt als flexibler, schneller, günstiger und

intuitiver Ansatz für die Praxis. Dabei werden vorab festgelegte Regeln verwendet, sogenannte Heuristiken, um rasch Urteile über das System zu fällen. Heuristiken sind Orientierungshilfen, die wesentliche Informationen schneller erkennbar machen, jedoch sind sie keine Beweise oder Lösungen an sich (Wirtz, 2021).

Im Jahr 1994 wurden die Heuristiken der HE mittels einer Faktorenanalyse validiert und im Jahr 2020 wurde die letzte Überarbeitung vorgenommen, bei der die Definitionen der Heuristiken sprachlich leicht angepasst wurden. Jedoch blieben die Heuristiken an sich unverändert und gelten auch heute noch als aktuell (Nielsen, 1994a, 2020). Zum besseren Verständnis zeigt Tabelle 1 die besagten Heuristiken mit einer kurzen Erklärung. Über die Zeit haben sich noch andere Heuristiken entwickelt. Bader, Schön und Thomaschewski (2017) zeigen in einer Übersichtsarbeit die relevantesten Heuristiken und Auswahlkriterien für eine Evaluation auf. In Tabelle 2 findet sich eine Übersicht.

Tabelle 1

Zehn Heuristiken nach Nielsen (2020). Eigene Übersetzung und Ergänzungen.

Nr.	Heuristik	Erklärung
1	Sichtbarkeit des Systemstatus	Das System gibt Nutzer*innen innerhalb einer angemessenen Zeit eine Rückmeldung über den aktuellen Stand der Dinge (z. B. Fortschrittsanzeige).
2	Übereinstimmung zwischen System und realer Welt	Es werden Konzepte und eine Sprache verwendet, die den Nutzer*innen bekannt und vertraut sind. Informationen sind in einer natürlichen und logischen Reihenfolge dargestellt.
3	Benutzerkontrolle und Freiheit	Falls versehentlich Aktionen ausgeführt werden, gibt es Möglichkeiten, dies ohne grössere Mühe zu korrigieren, ohne den ganzen Prozess nochmals zu durchlaufen.
4	Kohärenz und Standards	Wörter, Situationen oder Handlungen sollten immer das gleiche bedeuten. Es werden sprachliche und konzeptionelle Konventionen aus der Branche eingehalten (z. B. Einkaufswagen-Symbol in einem Shop).

Fortsetzung der Tabelle 1

Zehn Heuristiken nach Nielsen (2020). Eigene Übersetzung und Ergänzungen.

Nr.	Heuristik	Erklärung
5	Fehlervermeidung	Fehler werden vermieden (z. B. Bestätigungsfrage, bevor Daten tatsächlich gelöscht werden)
6	Wiedererkennen statt Erinnern	Informationen, die für die Nutzung des Systems erforderlich sind, sind sichtbar oder bei Bedarf leicht abrufbar. Erinnern benötigt mehr kognitive Ressourcen als Wiedererkennen (z. B. automatische Vervollständigung von Eingaben).
7	Flexibilität und Effizienz der Nutzung	Das System ist auf die Bedürfnisse anpassbar (z. B. können wenig erfahrene Nutzer*innen Einstellungen mittels digitaler Assistenten durchführen).
8	Ästhetisches und minimalistisches Design	Das System enthält keine bedeutungslosen Informationen oder selten benötigt werden. Die Darstellung ist nach aktuellen Designstandards aufgebaut.
9	Unterstützung der Benutzer*innen bei der Erkennung, Diagnose und Behebung von Fehlern	Fehlermeldungen sind in einer verständlichen Sprache formuliert sein (ohne Fehlercodes), das Problem genau benennen und eine konkrete Lösung vorschlagen.
10	Hilfe und Dokumentation	Das System sollte ohne zusätzliche Erklärungen auskommen. Falls eine Dokumentation dennoch nötig ist, sollte diese einfach und kontextuell abrufbar sein (z. B. Fragezeichen-Symbol neben einem Feld).

Tabelle 2

Relevante publizierte Heuristiken nach Bader et al. (2017).

Name	Autor*innen	Jahr
Designing the user interface: Strategies for effective human-computer interaction	Shneiderman et al.	1987
Heuristic evaluation of user interfaces	Nielsen und Molich	1990
Developing an Expert Evaluation Method for User eXperience of Cross-Platform Web Services	Väänänen-Vainio-Mattila und Wäljas	2009
Heuristic Evaluation of Persuasive Health Technologies	Kientz et al.	2011
Usability Heuristics Validation through Empirical Evidences: A Touchscreen-Based Mobile Devices Proposal	Inostroza et al.	2012
Ten User Experience Heuristics	Arhippainen	2013

Zur Vorbereitung einer HE werden für das System passende Heuristiken ausgewählt und die Anzahl der Expert*innen wird festgelegt. Nach Bader et al. (2017) sind die Auswahlkriterien für die Heuristiken im Wesentlichen die Passung mit dem Testobjekt und die Bekanntheit bei den Expert*innen. Nach der Durchführung der Evaluation bilden die Expert*innen einen Konsens und führen die gefundenen Probleme zusammen. Abschliessend werden Lösungsvorschläge erarbeitet und die Ergebnisse werden in einem Bericht dokumentiert, präsentiert und übergeben. Nach der Durchführung der Evaluation bilden die Expert*innen einen Konsens und führen die gefundenen Probleme zusammen. Abschliessend werden Lösungsvorschläge erarbeitet und die Ergebnisse in einem Bericht dokumentiert, präsentiert und übergeben. Abbildung 4 zeigt einen exemplarischen Ablauf einer heuristischen Evaluation (Hartson & Pyla, 2012; Moser, 2012).

Abbildung 4. Ablauf einer heuristischen Evaluation nach Hartson und Pyla (2012) und Moser (2012). Eigene Darstellung.

2.2.2 Usability-Testing

UXQB e.V. (2020) definiert einen Usability-Test als eine Aktivität, bei der Nutzer*innen beim Interagieren mit einem System beobachtet werden, während diese eine bestimmte Aufgabe lösen. Usability-Tests werden laut Barnum (2020) in zwei Typen unterschieden. Beim *formativen Usability-Test* geht es darum, möglichst viele Usability-Probleme zu entdecken, um ein interaktives System zu verbessern. Dieser Ansatz wird in der Regel bei Systemen angewendet, die sich noch in der Entwicklung befinden und vor der Markteinführung auf Verständlichkeit sowie Fehler überprüft und optimiert werden sollen. Bei einem *summativen Usability-Test* werden Systeme anhand von Kennzahlen (z.B. Bearbeitungszeit, Fehlerquote, o.ä.) untersucht, um die Qualität zu beurteilen. Dieser Ansatz kommt gewöhnlich bei Systemen zum Einsatz, die bereits am Markt sind oder kurz davorstehen.

Die Entscheidung, ob ein formativer oder summativer Usability-Test durchgeführt wird, hängt von der Zielsetzung ab. Entweder geht es darum, Usability-Probleme zu entdecken und das System zu verbessern, oder die Qualität des Systems anhand von Kennzahlen zu beurteilen. Diese Entscheidung beeinflusst letztlich die Wahl der Methoden sowie die Bestimmung der Stichprobengrösse (Lewis, 2012, zitiert nach Sauro & Lewis, 2016). Diese Aspekte haben wiederum einen Einfluss auf die Kosten der Usability-Evaluation.

Weiter wird unterschieden in *moderierte* und *nichtmoderierte Usability-Tests*. Bei einem moderierten Usability-Test führt ein*e Moderator*in die Testperson durch das Verfahren und stellt bei Bedarf Rückfragen zum Verhalten und zu den Äusserungen der Testperson. Optional wird dies durch zusätzliche Beobachter*innen unterstützt, die sich Notizen zu den Beobachtungen machen. Bei der nichtmoderierten Variante werden die Nutzer*innen nicht in Echtzeit beobachtet. Die Analyse erfolgt typischerweise via Videoaufzeichnung des Bildschirms, auf der das Interaktionsverhalten mit dem System sichtbar ist (Barnum, 2020).

Abschliessend kann noch bezüglich des Durchführungsorts differenziert werden. Wie in der Einleitung bereits erläutert wurde, nimmt der Anteil von Remote-Tests aufgrund diverser Vorteile in

den letzten Jahren zu (Chynał & Sobecki, 2015). Wie Sauer et al. (2019) zeigen konnten, gibt es zwischen den Tests vor Ort und Remote keine Unterschiede, solange gute Bedingungen vorhanden sind, wie eine klare Aufgabenstellung etc. Der Ablauf ist bei allen Varianten im Grundsatz ähnlich. Abbildung 5 zeigt ein typisches Vorgehen zur Durchführung eines Usability-Tests.

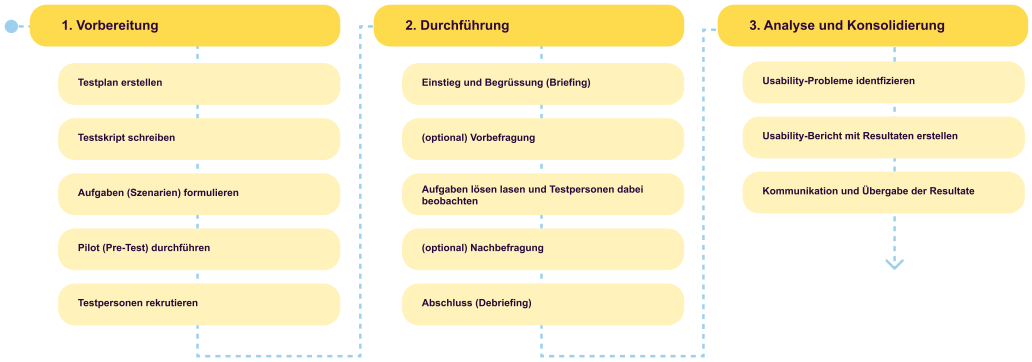


Abbildung 5. Ablauf eines Usability-Test nach UXQB e.V. (2020) mit eigenen Ergänzungen. Eigene Darstellung.

Ein zentrales Element bei formativen Usability-Tests ist das *laute Denken* (engl. *Thinking-Aloud*). Dabei werden Testpersonen gebeten, ihre Gedanken und Überlegungen laut auszusprechen, während sie die Aufgabe bearbeiten. Diese Verbalisierungen dienen neben der Beobachtung der Testperson zur Identifizierung von Usability-Problemen (Hertzum, Borlund & Kristoffersen, 2015). Für summative Usability-Tests wird davon abgeraten, da es die Messungen der relevanten Kennzahlen beeinflussen kann, wie die Bearbeitungszeit einer Aufgabe (UXQB e.V., 2020). Das laute Denken kann bei ungeübten Testpersonen vergessen gehen. Dies kann im Falle eines nichtmoderierten Usability-Tests problematisch sein, wenn dieser zu formativen Zwecken eingesetzt wird. Die Firma Solid hat für dieses Problem die Softwarelösung Solid User Tests entwickelt, bei der ein Avatar die Rolle der Moderator*in übernimmt.

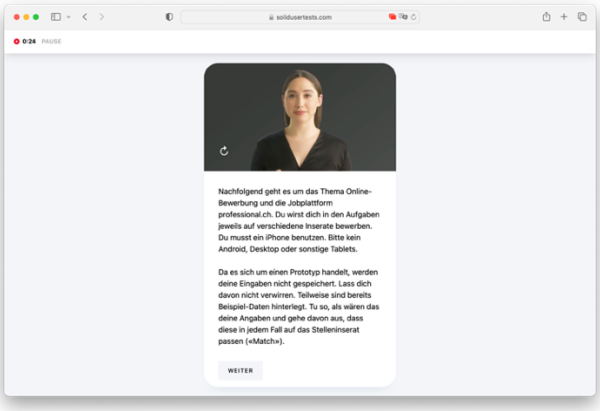


Abbildung 6. Solid User Tests für automoderierte Usability-Tests.

So werden Testpersonen automatisiert durch die Testsitzung geführt und an das laute Denken erinnert, falls sie dieses vergessen. Solid verwendet den selbstdefinierten Begriff automoderierter Usability-Test (UTA). Solid erhebt den Anspruch, dass dank dieser Lösung Usability-Tests schneller und damit wirtschaftlicher durchgeführt werden können.

2.3 Definition und Klassifizierung von Usability-Problemen

Die folgenden Abschnitte beschäftigen sich mit der Definition von Usability-Problemen, ihrer Relevanz und ihrem Schweregrad. Weiter wird die Problematik der Kategorisierung von Schwergraden und die Fehlertypen erläutert.

2.3.1 Definition

Nach UXQB e.V. (2020) ist ein Usability-Problem ein Vorkommnis bei der Nutzung eines interaktiven Systems, das sich auf die Effektivität, Effizienz und Zufriedenheit auswirkt.

2.3.2 Hits, falscher Alarm und verpasste Probleme

Im Rahmen der Signaldetektionstheorie (Green & Swets, 1966) wird die Fähigkeit betrachtet, das Vorhandensein und die Abwesenheit eines Reizes zu unterscheiden. Zum Beispiel wäre dies die Fähigkeit von einem PCR-Test, SARS-CoV-2 korrekt zu identifizieren. Im Kontext von UX wird

diese Theorie angewendet, um zu beschreiben, ob ein durch eine Usability-Evaluationsmethode angenommenes Usability-Problem auch wirklich eine Schwierigkeit ist. Begegnen Nutzer*innen bei der Interaktion mit dem System dem vorhergesagten Problem, wird von einem *Hit* (dt. Treffer) oder auch *True Positive* gesprochen (Sauro, 2016). In diesem Sinne werden Probleme, die mittels einer der beiden Usability-Test-Methoden aufgedeckt werden, in dieser Arbeit als *echte Probleme* bezeichnet, da hier direkt Nutzer*innen involviert sind.

Falls sich das vorhergesagte Problem nicht bestätigen lässt, wird es als *falscher Alarm* (auch *False Positive* bzw. Fehler des Typs 1) bezeichnet. Werden Probleme, die bei Nutzer*innen auftreten, von Expert*innen nicht gefunden, wird von *verpassten Usability-Problemen* (auch *False Negative* bzw. Fehler des Typs 2) gesprochen. So stellt sich die Frage, ob ein Problem wirklich ein Problem ist, wenn Nutzer*innen diesem nie begegnen. Sauro (2016) umschreibt dies folgendermassen: «Inspection methods in particular have often been criticized for identifying problems that are potentially false positives. After all, when there's no actual evidence that a user encounters a problem—only an expert's opinion of a potential problem—is it really a problem?»

Aus wirtschaftlicher Sicht können beide Fehlertypen problematisch sein. Falsche Alarme können dazu führen, dass ein Unternehmen Zeit und Ressourcen in die Behebung von Problemen investiert, die nicht existieren. Im schlimmsten Fall werden durch diese Anpassungen neue Probleme geschaffen. Auf der anderen Seite können verpasste Probleme beispielsweise zu Schwierigkeiten im späteren Betrieb des Systems führen und einen Mehraufwand durch eine erhöhte Anzahl Supportanfragen verursachen. Abbildung 7 zeigt die Einordung der Begriffe Hit, falscher Alarm und verpasste Probleme in einer Übersicht.

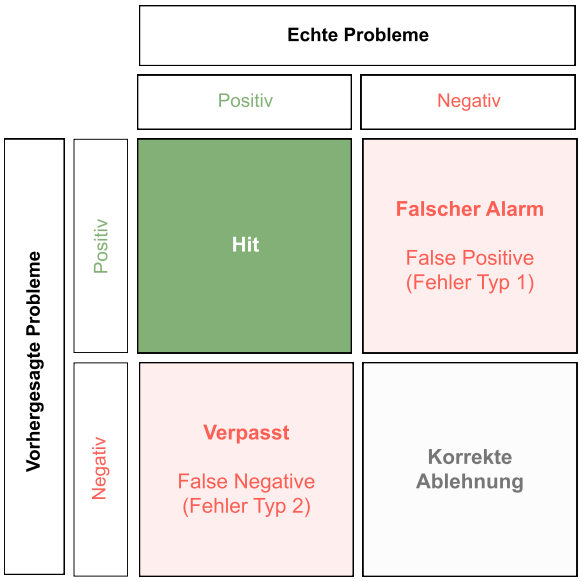


Abbildung 7. Visualisierung der Begriffe Hit, falscher Alarm und verpasstes Usability-Problem nach Sauro (2016). Eigene Darstellung.

2.3.3 Schweregrad und Relevanz

Usability-Evaluationen können eine Vielzahl von Erkenntnisse und Hinweise generieren. Aufgrund von begrenzten Ressourcen und zeitlichen Einschränkungen können Unternehmen nicht alle Resultate bearbeiten und dafür Lösungen implementieren. Durch die Festlegung, wie schwerwiegend das Problem ist, wird eine Priorisierung möglich. Dies wird als *Schweregrad* (engl. *severity rating*) bezeichnet.

Nach Koutsabasis et al. (2007) bezieht sich die *Relevanz* (engl. *relevance* oder *realness*) eines Usability-Problems darauf, ob es sich um ein Problem handelt, das Nutzer*innen tatsächlich beeinträchtigt, sowie auf das Ausmass, in welchem das Problem die Nutzer*innen in der Effektivität, der Effizienz und der Zufriedenstellung bei der Aufgabenbearbeitung beeinflusst.

Es existieren verschiedene Vorlagen für die Kategorisierung des Schweregrads, wie die von Nielsen (1994b, zitiert nach Moser, 2012), welche in Tabelle 3 ersichtlich ist.

Tabelle 3

Schweregrad-Einstufung nach Nielsen (1994b, zitiert nach Moser, 2012).

Stufe	Beschrieb
0	kein Problem erkennbar
1	reine Kosmetik
2	kleines Usability-Problem
3	grosses Usability-Problem
4	fatales Usability-Problem

Eine solche Kategorisierung ist jedoch insofern problematisch, da sie viel Interpretationsspielraum zulässt. Zum Beispiel ist nicht klar, was *gross* oder *klein* genau bedeutet. Die Kategorisierung unterliegt somit einer hohen Subjektivität. Travis und Hodgson (2019) liefern ein Vorgehen, nach welchem der Schweregrad eines Problems eruiert werden kann. Anhand des Einflusses, der Anzahl betroffener Nutzer*innen und des dauerhaften Bestehens des Problems wird eine Beurteilung vorgenommen, wie Abbildung 8 zeigt.

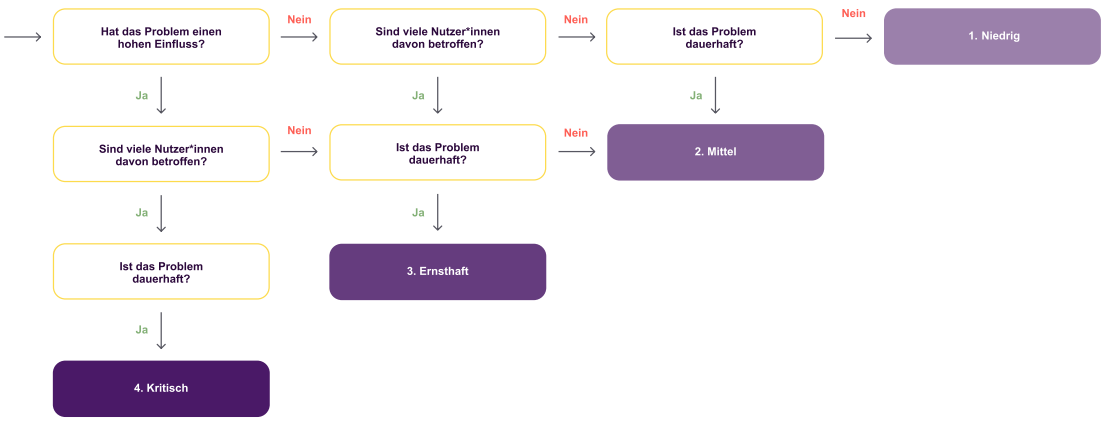


Abbildung 8. Entscheidungsbaum zur Klassifizierung des Schweregrads eines Usability-Problems nach Travis und Hodgson (2019). Eigene Darstellung und Übersetzung.

Mit diesem Ansatz lässt sich der der Interpretationsspielraum zwar etwas reduzieren, jedoch nicht ganz ausschliessen.

2.4 Magic Number 5

Zur Festlegung, wie viele Personen für einen Usability-Test rekrutiert werden sollen, wird in der Praxis oft die Regel *Magic Number 5* angewendet die auf Nielsen und Landauer (1993) zurückgeht. Diese besagt, dass in einem formativen Usability-Test bereits mit fünf Personen 80 % der Probleme einer Benutzerschnittstelle gefunden werden können. Diese Regel ist aus ökonomischer Sicht attraktiv, da mit wenigen Personen eine Mehrheit der Probleme identifiziert werden kann, ohne dass hohe Kosten und grosser Zeitaufwand entstehen. Diese Regel steht in einem gewissen Konflikt zur psychologischen Perspektive, wonach andere Ansätze zur Bestimmung einer Stichprobengrösse gewählt werden, wie statistische Genauigkeit oder A-priori-Teststärkeanalysen (Lakens, 2022).

Die Herleitung der mathematischen Formel, die der Magic Number 5 zugrunde liegt, erfolgt unter der Annahme einer bestimmten Wahrscheinlichkeit, mit der ein Usability-Problem durch eine einzelne Person entdeckt wird. Dieser Wert wird als *Discovery Likelihood* bzw. p bezeichnet[1] und steht für die Wahrscheinlichkeit, dass das Ereignis in *n* Versuchen mindestens einmal auftritt.

$$P(x \geq 1) = 1 - (1 - p)^n$$

Formel 1. Formel zur Berechnung der Wahrscheinlichkeit, ein Usability-Problem zu finden (Nielsen & Landauer, 1993, zitiert nach Sauro & Lewis, 2016).

Zum besseren Verständnis dient ein Rechenbeispiel: Man gehe davon aus, dass ein nicht funktionierender Link auf einer Webseite eine von zehn Personen verwirrt (p = 0.1 bzw. 10 %). Wenn nun ein Usability-Test mit insgesamt fünf Personen ($n = 5$) durchgeführt wird, dann errechnet sich aus der Formel eine Wahrscheinlichkeit von ca. 41 %, dass der tote Link in diesem Usability-Test auch gefunden wird.

[1] Es ist anzumerken, dass dies nicht mit dem p-Wert von Signifikanztests zu verwechseln ist. Zur besseren Unterscheidung wird p für die Discovery Likelihood in dieser Arbeit nicht kursiv geschrieben.

Welcher Wert für p in einem Usability-Test angenommen werden soll, unterscheidet sich je nach Quelle. Jedoch liegt der Durchschnitt bei 30 % (Turner, Nielsen & Lewis, 2002, zitiert nach Albert & Tullis, 2022). Die ursprünglichen Autoren der Formel, Nielsen und Landauer, gehen von 31 % aus (1993, zitiert nach Albert & Tullis, 2022). In diesen Quellen wird jedoch nicht nach dem Schweregrad differenziert. Dieser Umstand ist zugleich eine Kritik von Lewis (1994, zitiert nach Barnum, 2003) an der Magic Number 5. Wenn Probleme eine tiefere Wahrscheinlichkeit haben, entdeckt zu werden, sind mehr als fünf Personen notwendig, um 80 % der Probleme aufzudecken. Dies ist z. B. der Fall, wenn interaktive Systeme bereits eine hohe Usability aufweisen. Weiter führt Lewis aus, dass die Wahrscheinlichkeit der Entdeckung eines Problems auch durch die Auswahl, die Konstruktion und die Anzahl der Aufgaben in einem Usability-Test beeinflusst wird. So werden Probleme eher entdeckt, wenn es mehrere Testaufgaben gibt.

Zusätzlich sei angemerkt, dass offensichtlichere Probleme nicht unbedingt auch schwerwiegendere sein müssen (Sauro, 2014). Beispielsweise mag ein Link in einem Onlineshop nicht funktionieren, hindert Personen aber nicht daran, weiter Bestellungen abzusetzen. Umgekehrt wäre es vorstellbar, dass aufgrund eines seltenen Problems (z. B. dass zwei verschiedene Personen den exakt gleichen Benutzernamen verwenden) die Bestellungen oder Zahlungen falsch zugeordnet werden.

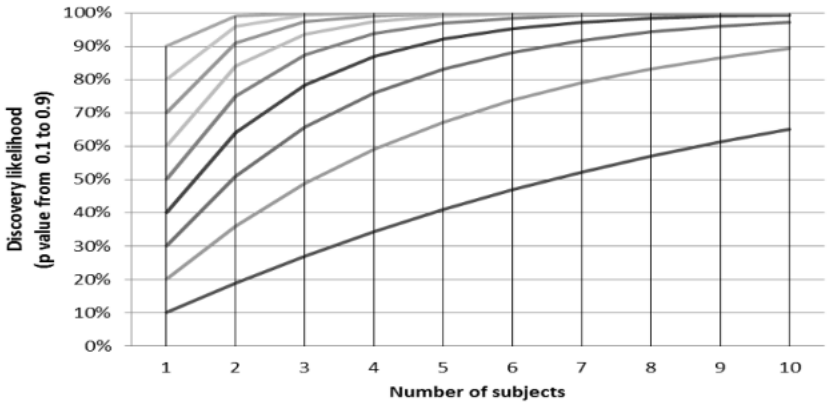
Um nun eine Stichprobengrösse berechnen zu können, kann die Formel nach *n* aufgelöst werden (Sauro & Lewis, 2016):

$$n = \frac{\ln(1 - P(x \geq 1))}{\ln(1 - p)}$$

Formel 2. Formel zur Berechnung der Stichprobengrösse nach Nielsen und Landauer (1993, zitiert nach Sauro & Lewis, 2016).

Hierzu wird wiederum p benötigt (die Annahme, wie viele Personen dem Problem begegnen) sowie die Festlegung der Wahrscheinlichkeit $P(x \geq 1)$, also wie viel Sicherheit beim Resultat vorhanden sein sollte, dies im gegebenen Test beobachten zu können. Wie die Discovery Likelihood

und die Anzahl Testpersonen zusammenhängen, wird in Abbildung 9 von Borsci et al. (2013) veranschaulicht.



Number of Subjects										
p value	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0.10	10%	19%	27%	34%	41%	47%	52%	57%	61%	65%
0.20	20%	36%	49%	59%	67%	74%	79%	**83%**	**87%**	**89%**
0.30	30%	51%	66%	76%	**83%**	**88%**	**92%**	**94%**	**96%**	**97%**
0.40	40%	64%	78%	**87%**	**92%**	**95%**	**97%**	**98%**	**99%**	**99%**
0.50	50%	75%	**88%**	**94%**	**97%**	**98%**	**99%**	**99.9%**	**99.9%**	**99.9%**
0.60	60%	**84%**	**94%**	**97%**	**99%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**
0.70	70%	**91%**	**97%**	**99%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**
0.80	**80%**	**96%**	**99%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**
0.90	**90%**	**99%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**

Abbildung 9. Die Wahrscheinlichkeit, ein Usability-Problem zu identifizieren, in Abhängigkeit von der Anzahl Testpersonen und dem p-Wert; Borsci et al. (2013) mit eigenen Ergänzungen. Copyright 2013 Brunel University.

Es wird nun verständlich, woher die bekannte Regel stammt, dass mit fünf Personen 80 % der Probleme gefunden werden kann. Unter der Annahme, dass ein Problem eine von drei Personen betrifft (bzw. p = 0.30), besteht mit fünf Personen eine Wahrscheinlichkeit von 83 %, das Problem in einem Test anzutreffen. Wenn aber das Bedürfnis nach mehr Sicherheit im Resultat (höhere Wahrscheinlichkeit, das Problem zu finden) oder seltenere Probleme gefunden werden sollen, sind mehr als fünf Personen notwendig. Daher sollte immer je nach Situation und Kontext entschieden werden, wie viele Testpersonen notwendig sind, um ein aussagekräftiges Ergebnis zu erzielen.

2.5 Wirtschaftlichkeit im Kontext von Usability und UX

Der Einsatz von Usability-Evaluationsmethoden steht in Konkurrenz mit anderen Aktivitäten in einem Softwareentwicklungsprozess. So werden Usability-Aktivitäten oft als Erstes vernachlässigt, wenn Zeit und Ressourcen knapp werden, z. B. kurz vor einer Markteinführung (Rajanen & Iivari, 2007). Es existieren insgesamt nur wenige Publikationen zum Thema der Kosten-Nutzen-Analyse im Kontext von Usability (Rajanen, 2011, 2020). Auch Mutschler und Reichert (2004) weisen darauf hin, dass bekannte Vorgehensmodelle, wie der *Usability Engineering Lifecycle*, keine Zusammenhänge zur gesamtwirtschaftlichen Perspektive aufzeigen.

Nachfolgend wird die Wirtschaftlichkeitsanalyse (Kosten-Nutzen-Analyse) zunächst erläutert und anschliessend in den Kontext von Usability bzw. UX gesetzt.

2.5.1 Wirtschaftlichkeitsanalyse

Mutschler und Reichert (2004) beschreiben die Analyse der Wirtschaftlichkeit anhand der drei Säulen Kosten-, Nutzen- und Risikoanalyse. Bei der Kostenanalyse werden die tatsächlichen Projektkosten sowie anfallende Folgekosten berücksichtigt. Hierzu gehören die Bereiche Hardware, Software, Personal, Infrastruktur sowie externe Dienstleistungen. Dies wird auch als Kostenstruktur bezeichnet.

Im Rahmen der Nutzenanalyse werden monetäre (z. B. Umsatz) sowie nichtmonetäre Nutzen (z. B. Kundenzufriedenheit) unterschieden. Identisch zur Kostenstruktur wird hier von einer Nutzenstruktur gesprochen.

Bei der dritten Säule, Risikoanalyse, ist das Ziel, vorhersehbare Probleme zu erkennen und deren Eintrittswahrscheinlichkeit abzuschätzen. Die daraus entstehenden Massnahmen können dann besser kalkuliert und geplant werden.

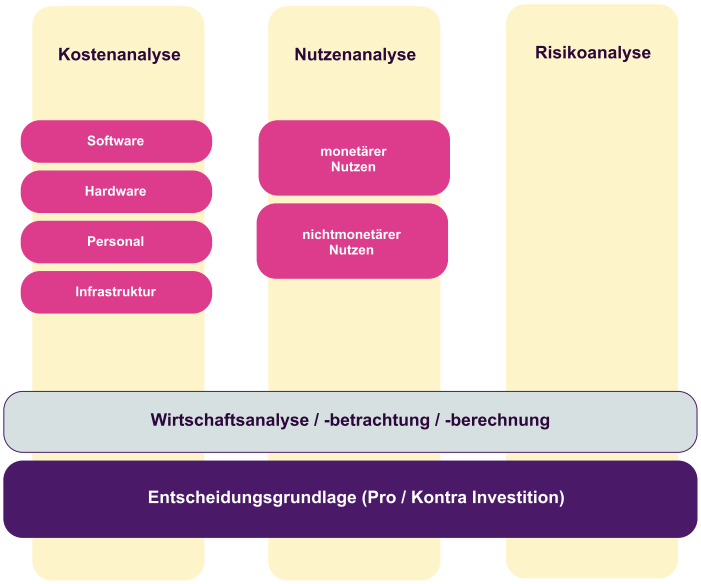


Abbildung 10. Kosten-Nutzen-Analyse mit den drei Säulen Kosten-, Nutzen- und Risikoanalyse nach Mutschler und Reichert (2004). Eigene Darstellung.

In Übereinstimmung mit diesem Modell definieren Burill und Ellsworth (1980, zitiert nach Rajanen, 2020) drei Vorgehensschritte, wonach als Erstes die zu erwartenden Kosten und Nutzen eines Vorhabens (wie Projekt, Produkt etc.) identifiziert werden, danach die Analyse des Verhältnisses zwischen diesen Kosten und Nutzen stattfindet und als Drittes eine Entscheidung über die Investition getroffen wird.

2.5.2 Kontextualisierung

In einer Übersichtsarbeit von Rajanen und Iivari (2007) werden die Kosten und die Nutzen von Usability-Aktivitäten dargestellt. Dabei unterscheiden sie in die Entwicklungsphase sowie in den eigentlichen Betrieb. Auf der Kostenseite stehen einmalige und wiederkehrende Kosten sowie anfallende Kosten für Überarbeitung des Systems basierend auf den gewonnenen Erkenntnissen. Nach Rajanen (2020) sind messbare Kosten, die (1) direkt entstehenden Aufwände (z. B. Mitarbeiterkosten, Projektausgaben, Agenturleistungen etc.), (2) einmalige Käufe (z. B. Equipment, Material etc.), (3) einmalige Entwicklungskosten und (4) kontinuierliche Kosten (z. B. Training des Personals etc.). Auf

der Nutzenseite stehen (1) die verbesserte Effizienz (z. B. bessere Nutzung der Ressourcen), (2) die verbesserte Effektivität (z. B. erhöhte Produktivität) sowie (3) die indirekten Vorteile (z. B. Nutzung der Erkenntnisse aus der Usability-Evaluation für Prozessoptimierungen).

Mutschler und Reichert (2004) verweisen darauf, dass die Informationen, die für eine Kosten-Nutzen-Analyse verwendet werden, oftmals aus dem Controlling eines Unternehmens stammen und einen Stand in der Vergangenheit zeigen. Somit ist kritisch in Frage zu stellen, ob diese Informationen als Entscheidungsgrundlage für zukünftige Investitionen sinnvoll sind. Mutschler und Reichert (2004) sagen dazu jedoch, dass es schwer nachvollziehbar sei, «[…] den Nutzen einer Investition ausschliesslich zum Zeitpunkt der Investition selbst zu betrachten, denn schliesslich ist es die Aussicht auf spätere, vorteilhafte Änderungen vor allem in der Kostenstruktur eines Unternehmens, die eine entscheidende Motivation ist.» (S.8). Weitaus schwieriger ist die Messung der abstrakten Kosten, z. B. reduzierte Produktivität beim Personal, und solcher Kosten, die aufgrund der Bindung der Ressourcen für die Implementierung oder durch mangelnden Wissenstransfer aufgrund Personalfluktuation entstehen. Dies sind erhebliche indirekte Kosten, die sich nur sehr schwer abschätzen und beziffern lassen.

Managemententscheidungen werden meist basierend auf Kennzahlen (sogenannten *Key-Performance-Indicators*, KPI) wie Umsatz oder Gewinn getroffen, um schlussendlich den Nutzen einer Investition (*Return on Investment*, ROI) beurteilen zu können. Es gilt daher, UX-Metriken, wie Bearbeitungszeit, Erfolgsquoten etc., in managementgerechte KPI zu übersetzen (Schrepp, Hinderks & Thomaschewski, 2017).

2.6 Actual Effectiveness und Actual Efficiency

Zum Vergleich der UEM in Bezug auf ihre Effektivität sowie Effizienz dienen die Kennzahlen *Actual Effectiveness* (dt. tatsächliche Effektivität) und *Actual Efficiency* (dt. tatsächliche Effizienz) nach Law und Hvannberg (2004). Diese sagen aus, wie gut die Methode dafür geeignet ist, Usability-Probleme zu finden (Effektivität) bzw. wie ressourcenschonend eine Methode ist (Effizienz).

Um die Actual Effectiveness einer Methode beurteilen zu können, muss zunächst die *Thoroughness* (dt. Gründlichkeit) eruiert werden. Die Thoroughness einer UEM nimmt zu, je mehr der echten Probleme identifiziert werden. Da nicht alle Probleme gleich bedeutsam sind, dient der Schweregrad als Gewichtung.

$$Thoroughness = \frac{Summe\ Schweregrad\ der\ Hits}{Summe\ Schweregrad\ aller\ echten\ Probleme}$$

Formel 3. Thoroughness (dt. Gründlichkeit) einer UEM nach Law und Hvannberg (2004).

Die echten Probleme stellen Probleme dar, die tatsächlich bei Nutzer*innen auftreten. Diese werden beispielsweise aus Helpdesk-Datenbanken oder nutzerbasierten Methoden (Interviews, Tagebuchstudien oder Usability-Tests) erhoben.

Die *Validity* (dt. Validität) gibt das Verhältnis der echten Probleme (Hits) zu den angenommenen Problemen an, die durch die UEM vorhergesagt wurden. Da Usability-Tests mit Nutzer*innen durchgeführt werden, sind alle durch diese Methode identifizierten Probleme definitionsgemäss als Hits zu betrachten. Dies bedeutet somit, dass es für diese Methode keine falschen Alarme gibt und daher ihre Validity immer perfekt ist (= 1).

$$Validity = \frac{Anzahl\ Hits}{Anzahl\ der\ vorhergesagten\ Probleme}$$

Formel 4. Validity (dt. Validität) einer UEM nach Law und Hvannberg (2004).

Aus dem Produkt der beiden Kennzahlen Thoroughness und Validity wird nun die Actual Effectiveness berechnet. Damit wird ausgedrückt, wie wirkungsvoll eine Methode darin war, Probleme in einem interaktiven System aufzudecken.

$$Actual\ Effectiveness = Thoroughness \times Validity$$

Formel 5. Actual Effectiveness nach Law und Hvannberg (2004)

Um die Actual Efficency zu berechnen, also herauszufinden, wie ressourcenschonend eine UEM war, werden die Anzahl Hits ins Verhältnis zum investierten Arbeitsaufwand gesetzt.

$$Actual\ Efficiency = \frac{Anzahl\ Hits}{Anzahl\ Arbeitsstunden}$$

Formel 6. Actual Efficiency nach Law und Hvannberg (2004).

3 METHODISCHES VORGEHEN

Im folgenden Kapitel wird das methodische Vorgehen beschrieben. Um eine klarere Vorstellung zu erhalten, werden zunächst die Herleitung der Hypothesen und das Forschungsdesign erklärt, dann werden das Testobjekt und anschliessend die Operationalisierungen vorgestellt. Danach folgen die Planung, die Auswahl und der Beschrieb der Stichprobe. Abschliessend werden die Datenerhebung und der Auswertungsprozess im Detail erläutert.

3.1 Herleitung der Hypothesen

Basierend auf den theoretischen Grundlagen erfolgt die Herleitung der Hypothesen in Bezug auf die beiden Fragestellungen. Die erste Fragestellung widmet sich dem Unterschied der gefundenen Usability-Probleme je nach UEM. An der Methode HE wird kritisiert, dass diese zu viele falsche Alarme generiere und nicht wirkungsvoll sei, um tieferliegende Probleme im Arbeitsprozess und -kontext zu finden (Hartson & Pyla, 2012; Law & Hvannberg, 2004). Zu viele falsche Alarme bergen das Risiko, dass falsche Entscheidungen in Bezug auf die Problembehebung getroffen werden oder dadurch sogar neue Probleme kreiert werden (Sauro, 2016). In einer Studie von Thyvalikakath, Monaco, Thambuganipalle und Schleyer (2009) wurden die Resultate von HE und Usability-Tests anhand von vier interaktiven Systemen miteinander verglichen. So konnten durchschnittlich 50 % der Usability-Probleme mit der HE identifiziert werden. Auch Moser (2012) weist darauf hin, dass nutzerbasierte Methoden aufgrund ihres partizipativen Charakters zuverlässigere und glaubwürdigere Resultate liefern.

In aller Regel sind bei einem Usability-Tests mehr Personen involviert als bei einer HE. Gestützt auf die Discovery Likelihood ist zu vermuten, dass dadurch auch mehr Usability-Probleme entdeckt werden. Weiter gibt es bei einem Usability-Test per Definition keine falschen Alarme. Aus diesen geschilderten theoretischen Überlegungen wird nun angenommen, dass, wenn Evaluationen mittels Usability-Tests durchgeführt werden, relevantere Ergebnisse gefunden werden. Die empirischen Hypothesen lauten:

H1a: Wenn eine Usability-Evaluation mittels eines Usability-Tests durchgeführt wird, werden mehr Usability-Probleme gefunden als bei einer heuristischen Evaluation.

H1b: Wenn eine Usability-Evaluation mittels eines Usability-Tests durchgeführt wird, dann werden Usability-Probleme mit einem höheren Schweregrad gefunden als bei einer heuristischen Evaluation.

Die zweite Fragestellung bezieht sich auf die Kostenanalyse. Aus ökonomischen und praktischen Gründen werden heute Usability-Tests vermehrt in Form eines sogenannten Remote-Testing (z. B. via Videokonferenztools) durchgeführt (Chynał & Sobecki, 2015). Daraus ergeben sich verschiedene Vorteile, wie das Wegfallen von Anfahrtswegen und der aufwändigen Infrastruktur eines Usability-Labors sowie eine flexiblere Einbettung in den Arbeitsalltag der Testpersonen.

Erkenntnisse von Hertzum, Molich und Jacobsen (2014) legen nahe, dass nichtmoderierte Usability-Tests kosteneffektiver sind und die Anzahl identifizierter Usability-Probleme ähnlich ist wie bei der moderierten Variante.

Aus diesen Überlegungen wird folgende empirische Hypothese abgeleitet:

H2: Wenn eine Usability-Evaluation mittels Usability-Tests durchgeführt wird, sind die Kosten pro gefundenem Usability-Problem tiefer als bei einer heuristischen Evaluation.

3.2 Forschungsdesign

Zur Untersuchung der beschriebenen Forschungsfragen und Hypothesen wurde ein sequenzielles Mixed-Methods-Forschungsdesign anhand eines einfaktoriellen Between-Subject-Feldexperiments mit drei Abstufungen gewählt. Die Abstufungen stellten die gewählten UEM dar mit den Ausprägungen HE, Usability-Test automoderiert (UTA) und Usability-Test moderiert (UTM). Zur Beantwortung der ersten Fragestellung wurde das Testobjekt mittels den drei UEM untersucht. Diese Ergebnisse wurden anschliessend konsolidiert nach Hartson und Pyla (2012), damit ermittelt werden konnte, wie viele Probleme durch welche Methoden gefunden wurden sowie um Aussagen in Bezug

auf deren Schweregrad machen zu können. Um die zweite Fragestellung zu beantworten, wurden die Arbeitsaufwände der Schritte Vorbereitung, Rekrutierung, Durchführung sowie Konsolidierung protokolliert. Die Drittkosten für die Rekrutierung der Testpersonen sowie Lizenzkosten für Solid wurden ebenfalls festgehalten. Dies erlaubte die Berechnung der Aufwände pro gefundenem Usability-Problem pro UEM.

Somit lässt sich das vorliegende Design als «*QUAL → quant*» beschreiben (Creswell & Plano Clark, 2018). Das Vorgehen wird dadurch begründet, dass die qualitativen Ergebnisse der drei UEM nicht ausreichen, um eine Aussage im Hinblick auf die gestellten Fragestellungen treffen zu können. Die nachgelagerten quantitativen Methoden ermöglichten eine nähere Untersuchung der Unterschiede und die Bewertung der formulierten Hypothesen.

Die Yousty AG hatte ein starkes Interesse daran, die konkreten Usability-Probleme möglichst rasch zu erfahren, um den NAP zeitnah weiterzuentwickeln. Um diesem Anspruch gerecht zu werden, führte der Autor eine *Quick and Dirty*-Analyse (QD) im Anschluss an die Erhebungsphase des UTM durch. Dabei wurden basierend auf den Erinnerungen sowie den eigenen Notizen die subjektiv zentralsten Usability-Probleme zusammengefasst. Dies widerspiegelt zugleich auch ein typisches Vorgehen bei der Praxispartnerin sinnhaft GmbH, wonach moderierte Usability-Tests und deren Auswertung aus Gründen der Ressourcen oftmals nur durch eine Person durchgeführt werden können. Aus diesem Grund wurde QD als relevant erachtet und mitaufgenommen. Die QD stellt somit keine eigene Methode im Sinne des Experiments dar, sondern eine Variante des UTM, bei der die Auswertung gekürzt ist.

Abbildung 11 zeigt zusammenfassend das beschriebene Forschungsdesign. Nachfolgend wird das Testobjekt, die Stichprobenplanung- und -auswahl, Vorbereitung, Pretests sowie Datenbereinigung, -aufbereitung und -auswertung näher beschrieben.

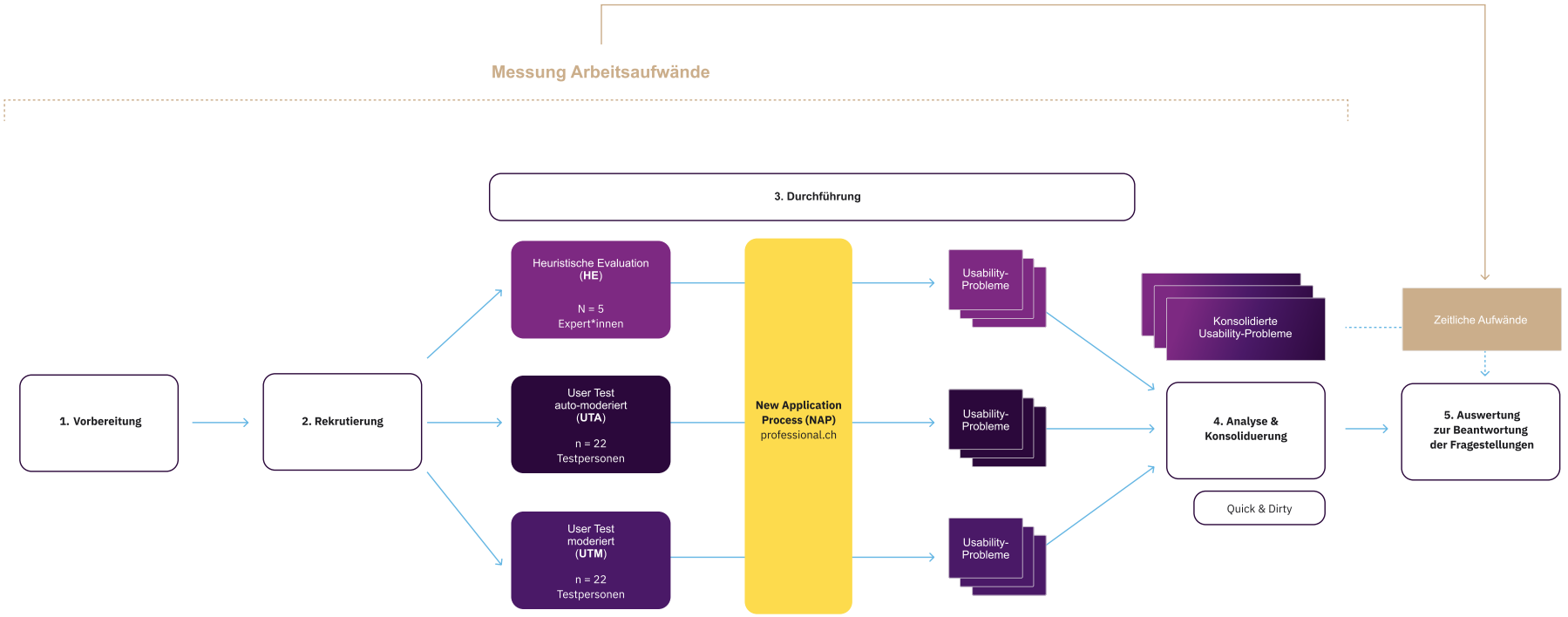


Abbildung 11. Mixed-Methods-Forschungsdesign für die vorliegende Arbeit.

3.3 Testobjekt New Application Process (NAP)

Wie in der Einleitung sowie im Forschungsdesign dieser Arbeit erläutert wurde, diente der New Application Process (NAP) von professional.ch als Testobjekt für die vorliegende Studie. Dabei handelt es sich um einen neuen Online-Bewerbungsprozess von professional.ch. Der NAP war zum Zeitpunkt der Erhebung als klickbarer HTML-Prototyp in deutscher Sprache verfügbar. So konnten die Testpersonen mit dem Prototyp interagieren und tatsächliche Eingaben machen (z. B. Suche nach Berufsbezeichnungen bei der Eingabe des Berufswegs). Aus technischen Gründen waren einige Stellen im Prozess mit erfundenen Daten (engl. *mocked data*) vorabgefüllt.

Der typische Ablauf des NAP war wie folgt: Nutzer*innen starteten den Prozess bei einem fiktiven Stellenangebot und kamen für die Bewerbung in eine Chat-artige Begrüssung mit ersten Informationen zum Ablauf. Anschliessend wurden persönliche Daten abgefragt, wie Geschlecht, Name oder Adresse. In einem nächsten Schritt wurden Angaben zur Ausbildung und zum Berufsweg gemacht. Es folgten zusätzliche Informationsabfragen zur Person, wie spezielle Fähigkeiten, Sprachen, absolvierte Kurse usw., um das Profil abzurunden. Abschliessend konnten die Eingaben in einer Übersicht in Form eines automatisch erstellten Lebenslaufs (CV) überprüft und bestätigt werden.

Zusätzlich gab es die Variante *Quick-Bewerbung*. Dieser Ablauf ist identisch wie oben beschrieben, jedoch ohne die zusätzlichen Informationen zur Person (siehe Tabelle 4, Schritt 7 «Das macht mich einzigartig»). Dies soll einen vereinfachten Bewerbungsprozess erlauben, insbesondere für Arbeitsstellen, bei denen zusätzliche Informationen zur Person in einem ersten Selektionsprozess nicht relevant sind. Für die vorliegenden Erhebung wurden beide Varianten (regulär sowie Quick-Bewerbung) als Testaufgaben (Szenarien) verwendet. Tabelle 4 zeigt die wesentlichen Schritte des NAP und sind exemplarisch in Abbildung 12 visualisiert.

Tabelle 4

Schritte des New Application Process.

Schritt		Beschrieb
1	Inserat	Stellenbeschrieb und Informationen zum Arbeitgeber
2	Willkommen	Begrüssung in einer Chat-artigen Darstellung
3	Meine Daten	Angaben zu Geschlecht, Name, Adresse usw.
4	Ausbildung	Angaben zu höchster Ausbildung und dazu, wann diese abgeschlossen wurde
5	Berufsweg 1	Berufserfahrung im relevanten Job
6	Berufsweg 2	Angaben zum letzten Arbeitgeber sowie Tätigkeitsbeschrieb
7	Einzigartig	Zusätzliche Informationen zur Person («Das macht mich einzigartig»)
7a	Skills	Spezielle Fähigkeiten, z. B. Führung eines Teams oder Ähnliches
7b	Sprachen	Muttersprache und Fremdsprachen
7c	Kurse und Zertifikate	Dokumente von Weiterbildungen, Zusatzkursen oder Ähnlichem
7d	Interessen und Hobbys	Tätigkeiten und Abrundung des Profils
7e	Projekte	Zusätzliche Aktivitäten, wie ein eigener Blogs etc.
8	CV-Übersicht	Zusammenfassung der Eingaben in der Darstellung als Lebenslauf
8a	Abschluss	Weiterführende Infos und Angaben zu den nächsten Schritten
8b	Bestätigung	Bestätigung in Form einer Animation (Konfetti)
9	Profil und Login	Profilerstellung und Login

Da vor allem die Anwendung des NAP auf mobilen Geräten für die Yousty AG von Interesse war, wurden sämtliche Datenerhebungen auf Apple iPhones durchgeführt. Damit konnte der Einfluss des Gerätetyps reduziert werden. Weiter setzt Apple keine Limite bei der Dauer der Bildschirmaufnahmen wie es bei anderen Anbietern, z. B. Samsung (vgl. «Galaxy phone stops recording videos after 10 minutes», Samsung, 2022), der Fall ist.

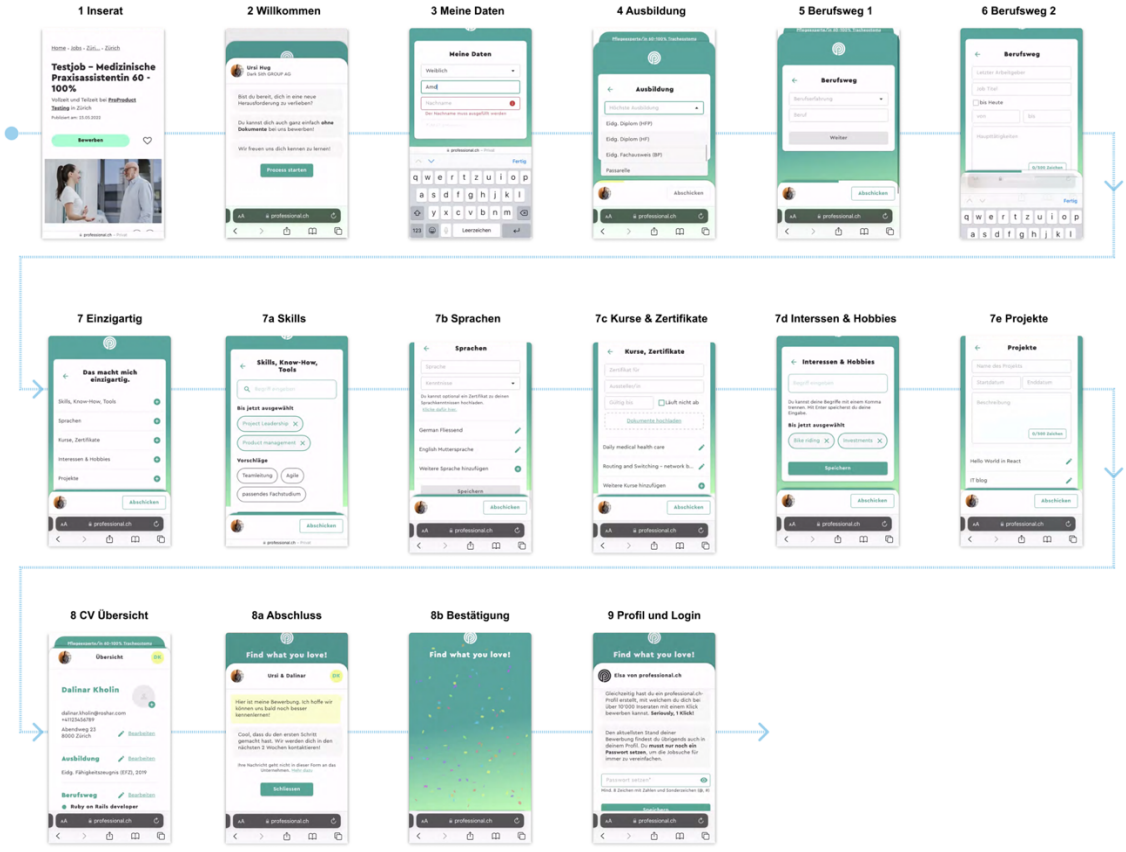


Abbildung 12. Exemplarischer Ablauf des New Application Process.

3.4 Operationalisierung

Die abhängigen Variablen waren die Anzahl gefundener Usability-Probleme (AV1.1), der Schweregrad der gefundenen Probleme (AV1.2), ihr Auftrittsort im Prototyp (AV1.3) sowie der Arbeitsaufwand pro Methode (AV2). Die Kosten pro Usability-Problem wurden dann aus dieser Variablen mit den Rekrutierungskosten sowie mit den Lizenzkosten verrechnet. Sie stellen somit keine eigene unabhängige Variable dar.

Zur Kontrolle von möglichen Einflussfaktoren (siehe Abschnitt 3.4.8) und aus dem beschriebenen Interesse der Praxispartnerin an der Variante UTA wurden ergänzend bei den Methoden UTA und UTM der emotionale Zustand (AV3.1), die subjektive Zufriedenheit (AV3.2) und die mentale Belastung (AV3.3) der Testpersonen bei der Aufgabenbearbeitung gemessen. Da diese Werte bei einer HE typischerweise nicht untersucht werden, wurden diese Variablen (AV3.1–3.4) bei dieser Methode nicht in die Erfassung miteinbezogen.

Tabelle 5

Abhängige Variablen der vorliegenden Studie in der Übersicht.

	Variable
AV1.1	Anzahl gefundener Usability-Probleme
AV1.2	Schweregrad
AV1.3	Auftrittsort im Prototyp
AV2	Arbeitsaufwand
AV3.1	emotionaler Zustand
AV3.2	subjektive Zufriedenheit
AV3.3	mentale Belastung

Nachfolgend wird auf die Operationalisierung der abhängigen Variablen näher eingegangen.

3.4.1 Anzahl gefundene Probleme

Hier wurde jeweils die konsolidierten Usability-Probleme gezählt, die durch die Methoden UTA, UTM und HE identifiziert wurden. Eine detaillierte Beschreibung des Konsolidierungsprozesses folgt in Abschnitt 3.11. Um eine möglichst objektive Erhebung sicherzustellen, kamen zwei unabhängige Evalutor*innen zum Einsatz.

3.4.2 Schweregrad

Als Basis für die Operationalisierung des Schweregrads eines Usability-Problems wurde das Schema nach Travis und Hodgson (2019) verwendet. Dieses wurde zusätzlich im Sinne von Nielsen (1994b) mit einer Stufe für positive Hinweise ergänzt.

Tabelle 6

Klassifikationsschema für den Schweregrad von Usability-Problemen nach Travis und Hodgson (2019) und Nielsen (1994b).

Stufe	Beschrieb
4	*Sehr hoch/Kritisch* Katastrophe bei der Benutzerfreundlichkeit: Dies muss unbedingt behoben werden, bevor das Produkt freigegeben werden kann.
3	*Hoch/Ernsthaft* Schwerwiegendes Problem der Benutzerfreundlichkeit: Wichtig zu beheben, daher sollte dem Problem hohe Priorität eingeräumt werden.
2	*Mittel/Geringfügig* Geringfügiges Problem der Benutzerfreundlichkeit: Die Behebung dieses Problems sollte mit niedriger Priorität behandelt werden.
1	*Tief/Kosmetisch* Nur kosmetisches Problem: Muss nicht behoben werden, es sei denn, es steht zusätzliche Zeit für das Projekt zur Verfügung.
0	*Kein Problem/Positiver Aspekt* Ich bin nicht der Meinung, dass dies ein Problem der Benutzerfreundlichkeit ist.

3.4.3 Auftrittsort

Auf Basis der konsolidierten Liste und des Video-Materials wurden die Probleme jeweils einem der Schritte des NAP-Prozesses (siehe Abschnitt 3.3) zugewiesen.

3.4.4 Arbeitsaufwand

Alle Arbeitsaufwände aus den drei Methoden wurden dokumentiert. Neben der aufgewendeten Zeit wurde festgehalten, ob es sich hierbei um die Phase Vorbereitung, Durchführung, Konsolidierung oder Analyse handelte. So konnte jeder Zeiteintrag den Methoden zugeordnet und der jeweilige Gesamtaufwand pro Methode berechnet werden.

3.4.5 Emotionaler Zustand

Zur Messung des emotionalen Zustands wurde der Fragebogen PANAS verwendet. Dieser umfasst zwanzig Items, die mit einer fünfstufigen Likert-Skala beantwortet werden. Die Reliabilität des PANAS ist hoch (Cronbachs Alpha, PA $\alpha = 0.89$ und NA $\alpha = 0.85$) (Crawford & Henry, 2004). In der vorliegenden Arbeit kam die deutsche Version zur Anwendung (Krohne, Egloff, Kohlmann & Tausch, 1996). Die Messung erfolgte im Anschluss an die Testsitzung bei der Nachbefragung. Der verwendete PANAS befindet sich in Anhang F.

3.4.6 Subjektive Zufriedenheit

Zur Messung der subjektiven Zufriedenheit wurde der Post-Study System Usability Questionnaire (PSSUQ) verwendet, da dieser Fragebogen die höchste Reliabilität (Cronbach Alpha = 0.94) aufweist sowie kostenlos verfügbar ist (Sauro & Lewis, 2016).

Zudem erschien die Anzahl der Items sowie deren inhaltliche Formulierungen für die vorliegende Erhebung und das Testobjekt am zutreffendsten.

Der PSSUQ umfasst insgesamt 16 Items, die mit einer siebenstufigen Likert-Skala beantwortet werden. Die Fragen decken verschiedene Aspekte ab und sind in drei Subskalen unterteilt:

1. Systemqualität (engl. *System Quality*)
2. Informationsqualität (engl. *Information Quality*)
3. Qualität der Benutzerschnittstelle (engl. *Interface Quality*)

Der PANAS wurde von Watson, Clark und Tellegen (1988) entwickelt. In der vorliegenden Untersuchung kam die deutsche Version nach Kaminski (2018) zur Anwendung, wobei die Fragen auf die Verwendung des NAP geringfügig angepasst wurden (z. B. wurde das Wort «System» mit «professional.ch» ersetzt).

Die Messung erfolgte im Anschluss an die Testsitzung bei der Nachbefragung. Der verwendete PSSUQ befindet sich in Anhang F.

3.4.7 Mentale Belastung

Der Fragebogen *NASA Task Load Index* (NASA-TLX) erfasst die kognitive Belastung mittels sechs Items für die Bereiche geistige Anforderung, körperliche Anforderung, Zeitdruck, Leistung, Anstrengung und Frustration. Dabei wird jede Frage angelehnt an eine Likert-Skala in einem Bereich von 0 bis 100 Punkten mit Fünferschritten bewertet (Hart, 2006). Der NASA-TLX zeigt eine hohe Reliabilität (Cronbachs Alpha = 0.84). Es kam die deutsche Version (Flägel, Galler, Steinhäuser & Götz, 2019) zum Einsatz. Aufgrund der technischen Limitierung des verwendeten Fragebogentools wurden die Fragen mit einer elfstufigen Likert-Skala abgebildet (Werte 0 bis 10).

Die Messung erfolgte im Anschluss an die Testsitzung bei der Nachbefragung. Der verwendete NASA-TLX befindet sich in Anhang F.

3.4.8 Kontextuelle Einflussfaktoren

Zur Kontrolle von möglichen Einflussfaktoren dient als Grundlage das *Four-Factor Framework of Contextual Fidelity* (4FFCF, adaptierte Version, Sauer et al., 2019), wonach die Eigenschaften der Testpersonen, die Aufgabenstellung, das Testobjekt (Prototyp) sowie die Testumgebung einen Einfluss auf die Resultate eines Usability-Tests haben. Demnach wird z. B. die Leistung, die wahrgenommene Benutzerfreundlichkeit oder die emotionale Reaktion einer Testperson durch diese Faktoren beeinflusst.

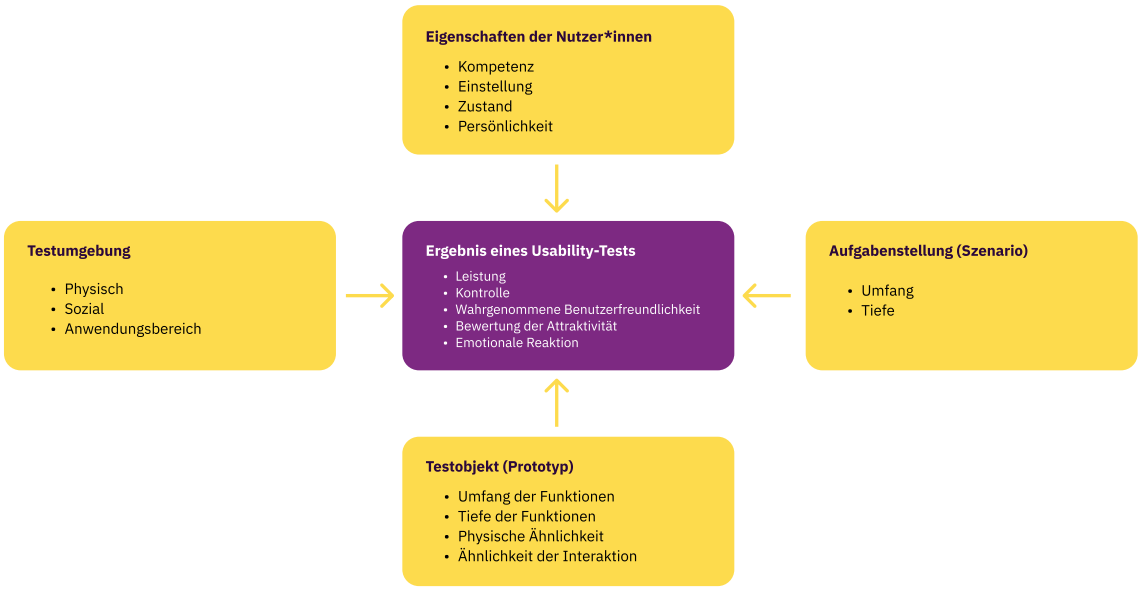


Abbildung 13. Four-Factor Framework of Contextual Fidelity nach Sauer et al. (2019). Eigene Darstellung und Übersetzung.

Eigenschaften der Testperson. Mittels der Vor- bzw. Nachbefragung wurden das Alter, Geschlecht, die persönliche Relevanz des Themas Bewerbung, die Erfahrung mit professional.ch sowie die Selbsteinschätzung der digitalen Affinität erfasst. Da die Tageszeit einen Einfluss auf den emotionalen Zustand einer Person haben kann (Egloff, Tausch, Kohlmann & Krohne, 1995), wurde dies ebenfalls erfasst.

Obwohl Ergebnisse von Kortum und Oswald (2018) darauf hinweisen, dass die Persönlichkeitsstruktur, insbesondere die Offenheit für Erfahrungen und die Verträglichkeit, einen Einfluss auf die subjektive Bewertung der Usability hat, konnte im Rahmen dieser Arbeit keine umfassende Erhebung der Persönlichkeitsstruktur vorgenommen werden, da dies den Rahmen überstiegen hätte.

Aufgabenstellungen und Testobjekt (Prototyp). Da die Yousty AG vorgegebene Anwendungsfälle mit dem NAP-Prototyp testen wollte, waren der Prototyp sowie die Szenarien für diese Studie vorgegeben und konnten nicht kontrolliert werden.

Testumgebung. Je nach verwendeter UEM unterscheidet sich der soziale oder physische Kontext. So kann die Präsenz einer beobachtenden Person die Emotionen und Leistung einer Testperson beeinflussen (Sonderegger & Sauer, 2009). Vollständigkeitshalber sei hier darauf hingewiesen, dass die UEM die unabhängige Variable ist, die im Rahmen dieser Erhebung manipuliert wird.

Erkenntnisse von Sauer et al. (2019) sowie Tullis, Fleischman, McNulty, Cianchette und Bergel (2002) weisen darauf hin, dass die Umgebung (Feld- oder Labortests) keinen wesentlichen Einfluss hat, solange die sonstigen Bedingungen günstig sind, wie dass keine mehrdeutigen Aufgabenstellungen vorhanden sind.

3.5 Heuristiken und Testaufgaben (Szenarien)

In den nachfolgenden Abschnitten werden die gewählten Heuristiken für die Heuristische Evaluation erläutert und begründet.

3.5.1 Heuristiken

Für die vorliegende Studie wurden die Heuristiken nach Nielsen (2020) ausgewählt. Diese Heuristiken schienen passend auf das Testobjekt und es konnte aufgrund ihrer Bekanntheit angenommen werden, dass sie den meisten UX-Professionals vertraut sind.

1. Sichtbarkeit des Systemstatus
2. Übereinstimmung zwischen dem System und der Wirklichkeit
3. Kontrolle und Freiheit
4. Konsistenz und Standards
5. Fehlervermeidung
6. Wiedererkennen statt Erinnern
7. Flexibilität und Effizienz
8. Ästhetisches und minimalistisches Design
9. Hilfe beim Erkennen und Beheben von Fehlern

10. Hilfe und Dokumentation

Um sicherzustellen, dass die Heuristiken allen Expert*innen gleichermassen bekannt und präsent waren, wurde zusätzlich eine Beschreibung inklusive Erklärung für deren Anwendung mitgeliefert. Diese ist Anhang J zu entnehmen.

3.5.2 Testaufgaben und Szenarien

Um eine gewisse Aufgabenbreite sicherzustellen, kamen vier verschiedene Testaufgaben im Usability-Testing zum Einsatz. Diese dienten ebenfalls den Expert*innen als Szenarien bei der HE.

Die vier Aufgaben (Szenarien) waren:

1. Bewerbung auf ein vorgegebenes Job-Inserat
2. Bewerbung auf ein vorgegebenes Job-Inserat (Variante Quick-Bewerbung)
3. Login auf der Webseite professional.ch
4. Bewerbung auf ein Inserat mittels hinterlegter Profildaten

Die detaillierte Ausformulierung der Aufgabenstellung findet sich in Anhang I.

Es fand keine Randomisierung der Aufgaben statt, da die gewählte Reihenfolge der Customer-Journey von professional.ch entspricht.

3.6 Stichprobenplanung

Im Folgenden wird zunächst die Stichprobenplanung der Testpersonen bei den Methoden UTA und UTM beschrieben und anschliessend die der Expert*innen der HE. Dabei werden die Ansätze erläutert und das Vorgehen begründet.

3.6.1 Usability-Testing (UTA, UTM)

Nach Lakens (2022) gibt es verschiedene Ansätze, um eine Stichprobengrösse zu bestimmen. Neben statistischen Argumenten sind auch andere Begründungen zulässig, beispielsweise ökonomische oder die Anwendung von allgemein akzeptierten Heuristiken bzw. Normen. Weiter argumentiert Lakens, dass in der Forschung besonders die ökonomischen Faktoren zu wenig Beachtung finden, obwohl alle Forschenden mit der Thematik von begrenzten Ressourcen konfrontiert sind. Die Begrenzung der Ressourcen wird durch die Abwägung zwischen den Kosten der Datensammlung und dem Nutzen der gewonnenen Informationen gerechtfertigt.

Im Rahmen der Auseinandersetzung mit der Thematik der Wirtschaftlichkeit und dem Anspruch auf Praxisnähe dieser Arbeit wurde die im theoretischen Teil beschriebene Regel Magic Number 5 als Grundlage für die Planung der Stichprobengrösse verwendet, da diese die Argumente der begrenzten Ressourcen sowie die Anwendung von allgemein akzeptierten Normen miteinschliesst. Da jedoch für die vorliegende Erhebung eine höhere Discovery Likelihood als 80% und tiefere p-Werte angestrebt wurden, musste die Stichprobengrösse entsprechend berechnet werden.

Zunächst musste der Wert für p bestimmt werden. Sauro und Lewis (2016) liefern Angaben, wie häufig Usability-Probleme in bestimmten Anwendungsbereichen vorkommen (siehe Tabelle 7).

Tabelle 7

p-Werte für verschiedene Anwendungsbereiche nach Sauro und Lews (2016).

		95 % KI	
Anwendungsbereich	**p-Wert**	**unten**	**oben**
Business-Applications	0.37	0.25	0.50
Consumer-Software	0.23	0.13	0.33
Webseiten	0.04	0.03	0.06

Sauro und Lewis (2016) weisen jedoch darauf hin, dass diese Daten aus eigenen Beobachtungen stammen und nicht repräsentativ sind. Aufgrund des Fehlens von weiteren Quellen zu p-Werten der verschiedenen Anwendungsbereiche stützen sich die nachfolgenden Berechnungen auf diese Daten.

Weiter gibt es ebenfalls keine klaren Kriterien, wann etwas als Business-Application, Consumer-Software oder Webseite zu betrachten ist. Der NAP wurde im Rahmen dieser Stichprobenplanung nach eigenem Ermessen als *Consumer-Software* kategorisiert, mit der Begründung, dass Privatpersonen eine Bewerbung für eine Arbeitsstelle verschicken und ihr Bewerbungsprofil verwalten. Für eine konservative Festlegung von p dient der untere Bereich des Konfidenzintervalls (95 % KI [0.13 - 0.33]). Als Wahrscheinlichkeit $P(x \geq 1)$ wird 95 % angenommen. Daraus resultiert folgende Berechnung:

$$n = \frac{\ln(1 - P(x \geq 1))}{\ln(1 - p))}$$

$$21.511 = \frac{\ln(1 - .95))}{\ln(1 - .13))}$$

Formel 7. Berechnung der Stichprobengrösse für die beiden Usability-Tests (UTA, UTM).

Die resultierende Stichprobe wird aufgerundet (Diamond 1981, zitiert nach Sauro & Lewis, 2016) und ergibt somit $n = 22$ für jeden der beiden Usability-Tests. Dies liegt weit über der Anzahl Testpersonen, die häufig in der Praxis bei Usability-Tests zum Einsatz kommen (Barnum, 2003). Jedoch erlaubt diese Stichprobengösse eine bessere Aussagekraft, und mittels Bootstrapping-Verfahren können Vergleiche mit kleineren Stichproben gezogen werden.

3.6.2 Heuristische Evaluation (HE)

Die Autoren der Methode, Nielsen und Landauer, haben sich intensiv mit der Frage nach der optimalen Anzahl Expert*innen auseinandergesetzt (Nielsen, 1992, 1994a; Nielsen & Landauer, 1993) und kommen zum Schluss, dass drei bis fünf Personen ausreichend sind. Mit einer höheren Anzahl Personen ist nach Nielsen und Landauer kein relevanter Wissenszuwachs zu generieren. Die Anzahl von fünf Expert*innen hat sich als Industriestandard etabliert (Tan, Liu & Bishu, 2009).

Weiter hat gemäss Nielsen und Mack (1994, zitiert nach Tan et al., 2009) das Domänenwissen Einfluss auf die benötigte Anzahl Testpersonen. Somit genügen bereits zwei bis drei Personen als sogenannte *Double Experts*, um 60 % der Probleme zu finden. Dies sind Personen, die zum einen über fundiertes Wissen im Bereich UX bzw. Usability verfügen und zum anderen spezifische Kenntnisse im Anwendungsbereich haben.

Aufgrund dieser Überlegungen wird der Standard übernommen und die Stichprobengrösse der Expert*innen auf fünf gesetzt ($N = 5$). Somit war es nicht zwingend notwendig, dass die rekrutierten Personen Double Experts sind.

3.7 Stichprobenauswahl

Für die Rekrutierung der Testpersonen in den beiden Usability-Tests wurde auf den Drittanbieter TestingTime zurückgegriffen. Dies ermöglichte ein effizientes und praxisnahes Vorgehen. Für die HE wurden Expert*innen aus dem beruflichen Umfeld des Autors rekrutiert.

Zunächst wird die Stichprobe der Testpersonen der Methoden UTA und UTM dargestellt, danach folgen die Expert*innen der HE. Dabei werden die relevanten Kriterien wie Alter, Geschlecht, Erfahrung usw. beschrieben.

3.7.1 Testpersonen Usability-Testing (UTA/UTM)

Um die Zielgruppe des NAP optimal abbilden zu können, wurden in Absprache mit der Yousty AG folgende Kriterien zur Teilnahme an der Studie festgelegt:

- Alter zwischen 18 und 30 Jahren
- ausgeglichene Geschlechterverteilung
- persönliche Relevanz (innerhalb der letzten zwölf Monate z. B. online neue Stelle gesucht, Bewerbung geschrieben, an Bewerbungsgesprächen teilgenommen etc.)
- berufliche Tätigkeit in den Branchen Pflege, Handwerk, kaufmännische Berufe sowie Gastronomie

Beim Alter ist anzumerken, dass auf Wunsch der Yousty AG eine möglichst gleichmässige Verteilung zwischen den Altersgruppen 18–20 Jahre, 21–25 Jahre sowie 26–30 Jahre angestrebt wurde. Zusätzlich wurden aus den erwähnten technischen Gründen nur Personen mit einem iPhone rekrutiert. Damit diese Kriterien sichergestellt werden konnten und sich die Erhebung nicht über einen zu langen Zeitraum erstreckte, wurde die Rekrutierung und Vergütung der Testpersonen mit dem externen Dienstleister TestingTime abgewickelt. Dies spiegelt auch ein praxisnahes Vorgehen wider, wonach aus Kostengründen oft mit Drittanbietern für den Rekrutierungsprozess gearbeitet wird. Die Yousty AG stellte die notwendigen finanziellen Mittel zur Verfügung. Eine eigenhändige Rekrutierung im Umfeld des Autors wurde in Betracht gezogen, jedoch hätte dies einen grösseren

Zeitaufwand erfordert und die Rekrutierungskriterien hätten nicht im gleichen Masse kontrolliert werden können. Gemäss Stichprobenplanung wurden pro Usability-Test-Methode 22 Personen rekrutiert, was somit total 44 Personen ($N = 44$) waren.

Alter. Das Durchschnittsalter der Testpersonen über beide Gruppen betrug 24.2 Jahre ($SD = 3.28$). Die jüngste Person war 18 und die älteste 29 Jahre alt. Zwischen den Gruppen UTA und UTM bestand kein Altersunterschied ($t(42) = -0.87$, $p = 0.39$). Die von der Yousty AG gewünschten Altersgruppen (18–20 Jahre, 21–25 Jahre, 26–30 Jahre) konnten gewährleistet werden. 14 Personen waren 18–20 Jahre, 14 Personen 21–25 Jahre und 16 Personen 26–29 Jahre alt.

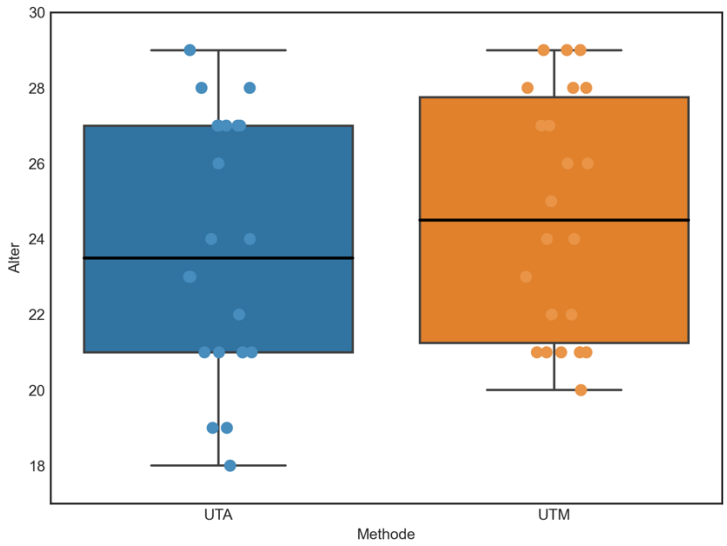


Abbildung 14. Altersverteilung der Testpersonen.

Geschlecht. Bei der Rekrutierung wurde ein ausgeglichenes Geschlechterverhältnis angestrebt. Insgesamt identifizierten sich 27 Personen als Frauen, 16 als Männer und keine als divers. In Bezug auf die Verteilung zwischen den beiden Gruppen UTA und UTM zeigt ein Chi-Quadrat-Anpassungstest, dass die beobachteten Häufigkeiten nicht signifikant von den erwarteten Häufigkeiten abwichen ($\chi 2$ (1, $n = 43$) = 0.26, $p = 0.61$).

Tabelle 8

Kontingenztabelle Geschlecht der Testpersonen.

Geschlecht	**Methode**		**Total**
	UTA	UTM	
Männlich	7	9	17
Weiblich	14	13	27
Divers	0	0	0
Total	21	22	43

Persönliche Relevanz. Eine weitere Voraussetzung zur Teilnahme war die persönliche Relevanz, konkret die Erfahrung mit dem Schreiben von Bewerbungen. Dieses Kriterium wurde durch TestingTime sichergestellt, jedoch im Rahmen der Vorbefragung der Studie nochmals detailliert abgefragt. Personen der Gruppe UTA haben in den letzten zwölf Monaten durchschnittlich einmal pro Monat eine Bewerbung verschickt (*Md* = 4, *SD* = 1.27), Personen der Gruppe UTM mehrmals pro Monat (*Md* = 5, *SD* = 0.96). Diese Unterschiede zwischen den Gruppen waren nicht signifikant (exakter Mann-Whitney-U-Test, *U* = 211.00, *p* = 0.62). Die Häufigkeiten sind in Tabelle 9 ersichtlich.

Tabelle 9

Kontingenztabelle Persönliche Relevanz (Bewerbungen) der Testpersonen.

	«Wie häufig hast du dich in den letzten 12 Monaten mit dem Thema Bewerbung auseinandergesetzt?»							
	Keine Antwort /Weiss nicht	*Seltener oder nie*	*Mehrmals pro Jahr*	*Einmal pro Monat*	*Mehrmals pro Monat*	*Mehrmals pro Woche*	Fehlend	Total
	1	*2*	*3*	*4*	*5*	*6*		
UTA	0	1	6	5	4	5	1	22
UTM	0	1	3	4	13	1	0	22
Total	0	2	9	11	17	6	1	44

Bekanntheit professional.ch. Die Bekanntheit und die Vorerfahrung mit professional.ch wurden ebenfalls abgefragt. Insgesamt traf dies nur auf eine Person (*n* = 1) zu.

Ausschlusskriterien. Als Ausschlusskriterien galten eine Beschäftigung bei einer Job-Plattform sowie ein Wohnsitz ausserhalb der Schweiz. Diese Voraussetzungen wurden durch TestingTime sichergestellt.

3.7.2 Expert*innen der HE

Die fünf rekrutierten Expert*innen (*N* = 5) waren im Durchschnitt 33 Jahre alt (*SD* = 3.85) und wiesen im Bereich Usability bzw. UX eine durchschnittliche Berufserfahrung von 7,9 Jahren (*SD* = 4.64) auf. Drei Personen identifizierten sich als weiblich (*n* = 3) und zwei als männlich (*n* = 2). Die selbsteingeschätzte Expertise (*Md* = 5, *SD* = 0.45) sowie die digitale Affinität (*Md* = 6, *SD* = 0.84) waren insgesamt hoch (siebenstufige Likert-Skala, 1 = *sehr niedrig*, 7 = *sehr hoch*). Keinem der Expert*innen war professional.ch vorher bekannt. Die Expert*innen sind in den Branchen Banking, IT, E-Commerce, Softwareentwicklung sowie UX-Design tätig.

Tabelle 10

*Expert*innen der heuristischen Evaluation.*

	Alter	**UX-Erfahrung in Jahre**	**Selbsteinschätzung UX-Expertise** 1 = *sehr niedrig*, 7 = *sehr hoch*	**Selbsteinschätzung digitale Affinität** 1 = *sehr niedrig*, 7 = *sehr hoch*	**Geschlecht**
Expert*in 1	38	15	5	6	w
Expert*in 2	37	10	5	7	m
Expert*in 3	30	3.5	5	6	w
Expert*in 4	32	5	5	6	m
Expert*in 5	30	6	6	7	w
M	33.4	7.9	5.2	6.2	–
Md	32	6	5	6	–
SD	3.85	4.64	0.45	0.84	–

Die Rekrutierung fand im beruflichen Umfeld des Autors sowie durch Direktanschrift via LinkedIn statt. Die fünf Expert*innen erhielten eine pauschale Entschädigung von CHF 200 mit Ausnahme der beiden Expert*innen, die Mitarbeitende der Yousty AG bzw. sinnhaft GmbH sind, da diese die Analyse im Rahmen ihrer Arbeitszeit erledigen konnten. Diese Entschädigung wurde später in der Kostenberechnung nicht berücksichtigt, da sie ein symbolischer Betrag war. Bei der Kostenanalyse wurden die geleisteten Stunden mit einem Stundensatz berechnet.

3.8 Vorbereitung und Pretests

Im Rahmen der Vorbereitung wurden die notwendigen Vorkehrungen für die Durchführung der UEM getroffen. Dazu gehörten beispielsweise die Auswahl von passenden Heuristiken, die Rekrutierung der Testpersonen und Expert*innen, die Koordination mit professional.ch, die Konfiguration der Software Solid User Tests usw.

Um ein möglichst standardisiertes Vorgehen zu gewährleisten und die Qualität sicherzustellen (Koutsabasis et al., 2007), wurden für die drei Methoden – HE, UTA und UTM – schriftliche Unterlagen erstellt, die den Ablauf beschrieben und Instruktionen enthielten, wie Leitfäden, Heuristiken oder Aufgabenbeschrieb. Zur Vergleichbarkeit wurden in allen Methoden die gleichen Szenarien bzw. Testaufgaben verwendet. Diese Dokumente nach Lavery, Cockton & Atkinson (1997), Nielsen (1994b), Travis (2009) sowie Travis und Hodgson (2019) sind in Anhang J, I sowie L zu finden.

Anschliessend an die Vorbereitung wurden für alle drei Methoden Pretests durchgeführt, um einen reibungslosen Ablauf sowie inhaltliche Klarheit sicherzustellen. Abgesehen von kleineren sprachlichen Präzisierungen bei den Aufgabestellungen mussten keine Veränderungen vorgenommen werden. Es zeigte sich jedoch, dass aufgrund der Begrüssung, der Einführung und der Verabschiedung in einer UTM-Sitzung mehr Zeit benötigt wurde als bei einer UTA-Sitzung. So wurden die Dauern für eine UTA-Sitzung auf 30 Minuten und für eine UTM-Sitzung auf 45 Minuten festgelegt. Dies war insofern relevant, da die Sitzungsdauer ein Kostenfaktor bei der Bestellung bei TestingTime ist.

3.9 Durchführung

Der Erhebungszeitraum der beiden Usability-Tests lag zwischen dem 7. Juni und dem 29. Juni 2022 und erstreckte sich über insgesamt 23 Tage. Die HE fanden zwischen dem 14. Juni und dem 25. Juni 2022 statt und dauerten 11 Tage.

Die Einverständniserklärung wurde direkt über TestingTime abgewickelt. Die Testpersonen wurden in der Testsitzung explizit nochmals um Erlaubnis bezüglich der Videoaufnahme gefragt (Bild und Ton) und das Einverständnis wurde auf Video festgehalten. Auf die Aufnahme des Kamerabilds wurde verzichtet, da es für die Beantwortung der Fragestellung als nicht notwendig erachtet wurde.

Alle Daten wurden in anonymer Form festgehalten. Damit bei der Auswertung die Aufnahmen und Antworten aus den Fragebögen zugeordnet werden konnten, wurden die Testpersonen gebeten, einen eindeutigen Code anzugeben. Dieser setzte sich aus den ersten zwei Buchstaben des Vornamens, den ersten zwei Buchstaben des Nachnamens sowie den letzten zwei Ziffern des Jahrgangs zusammen. Aufgrund der technischen Limitierung des Fragebogen-Tools, konnten keine komplett zufälligen Codes generiert werden, um die Daten der Vor- und Nachbefragung miteinander zu verknüpfen.

Bei den UTM übernahm der Autor der vorliegenden Thesis die Moderation. Daneben waren keine Drittpersonen anwesend. Die Sitzungen wurden mittels des Videokonferenz-Tools *Zoom* durchgeführt und aufgezeichnet. Für eine standardisierte Durchführung sorgte ein Leitfaden, der die wesentlichsten Schritte vor den Tests, während der Tests und nach den Tests definierte. Der Leitfaden für die Moderation findet sich in Anhang L.

Bei den UTA mussten die durch die Testpersonen eingereichten Aufnahmen jeweils auf die Qualität überprüft werden. Im Falle von unzureichender Qualität (z. B. fehlende Video-Datei) wurde bei TestingTime eine Ersatzperson angefordert. Um eine Beeinflussung der UTA-Sitzungen zu vermeiden, hat der Moderator die UTA-Videos im Voraus inhaltlich nicht angeschaut, damit diese die Moderation bzw. die UTM-Resultate nicht beeinflussten. Abbildung 15 visualisiert den Ablauf des Usability-Tests in den Varianten UTA und UTM.

Abbildung 15. Ablauf des Usability-Tests (UTA, UTM).

Nach Abschluss der Erhebung wurden alle Videodateien (UTM, UTA) zwei Evaluator*innen für die Auswertung zugestellt. Diese analysierten die Videos und dokumentierten die beobachteten Probleme in einer Excel-Liste. Dabei wurden eine kurze Beschreibung des Problems, eine Einschätzung des Schweregrads sowie bei Bedarf ein ergänzender Kommentar festgehalten. Ebenfalls wurde dokumentiert, bei welcher Testperson das Problem aufgetreten ist. Um Reihenfolgeneffekte bzw. Positionseffekte (Wirtz, 2021) zu vermeiden, wurde für beide Evaluator*innen eine zufällige Reihenfolge der Videos für die Analyse aus beiden Methoden UTA und UTM bestimmt.

Die fünf Expert*innen der HE führten die Evaluationen basierend auf den definierten Heuristiken individuell durch. Dazu dienten die erwähnten schriftlichen Unterlagen (siehe Anhang J) als Grundlage. Die gefundenen Probleme wurden in Form von fünf einzelnen Excel-Listen an den Autor zur Konsolidierung übergeben. Aufgrund der begrenzten zeitlichen Verfügbarkeit der Expert*innen konnte keine gemeinsame Konsolidierung vorgenommen werden.

3.10 Datenbereinigung

Die Datenbereinigung umfasste die Reduktion der Daten durch Entfernen von positiven Kommentaren sowie Prüfung auf Durchkreuzer und Datenqualität. Im Folgenden werden diese Aspekte kurz erläutert.

3.10.1 Usability-Probleme

In den Usability-Tests und bei der HE wurden basierend von den Testpersonen bzw. Expert*innen dokumentiert auf dem vorgegebenen Schema für den Schweregrad auch positive Kommentare, z. B. «schönes Design», «sehr angenehm zum Ausfüllen» oder «cool, dass man sich mobil bewerben kann». Für Unternehmen können solche Erkenntnisse wertvoll sein und ausschlaggebende Hinweise für die Weiterentwicklung des Produkts bzw. der Dienstleistung liefern. Um jedoch eine Aussage im Sinne der in dieser Arbeit verwendeten Definition von Usability-Problemen machen zu können und eine zu bewältigbare Datenmenge zu haben, wurden alle Kommentare, die positive Einträge enthielten, von der Analyse ausgeschlossen. Weiter wurden Einträge, die nicht eindeutig zuzuordnen waren oder zu viel Spielraum für Interpretationen liessen, ebenfalls ausgeschlossen (z. B. «war sehr schnell, aber nicht sicher ob seriös ausgefüllt»).

3.10.2 Durchkreuzer

Zur Kontrolle, ob Testpersonen Antworten in einem der Fragebögen durchgekreuzt haben (sogenanntes *straight-lining*), kam die Methode *Intra-Individual Response Variability* (IRV) nach Dunn, Heggestad, Shanock und Theilgard (2018) zur Anwendung. Es konnten keine Durchkreuzer festgestellt werden und es wurden daher keine Antworten ausgeschlossen.

3.10.3 Videoqualität

Eine Videoaufnahme ($n = 1$) musste aufgrund der Qualität (fehlende Audioaufnahme und fehlende Teile) von der Analyse ausgeschlossen werden. Dies fiel erst in der Analysephase auf. Daher wurde keine Ersatzperson bei TestingTime bestellt.

3.11 Datenaufbereitung und -auswertung

Der Auswertungsprozess orientierte sich an Hartson und Pyla (2012). In einem ersten Schritt wurde das Videomaterial wie oben beschrieben begutachtet und die Beobachtungen wurden festgehalten (*evaluator comments*). Da Ergebnisse einer Evaluation stark von Evaluator*innen abhängen können, wie die *Comparative Usability Evaluations* (CUE) Studien von Molich et al. (2010) zeigen, wurden zwei unabhängige Personen für die Auswertung eingesetzt. Analog dazu gaben in der HE die fünf Expert*innen ihre Einschätzungen ab. Die Beobachtungen bzw. Einschätzungen wurden dann in einem zweiten Schritt vom Autor zu Usability-Problemen (*UX problem instances*) konsolidiert. In einem nächsten Schritt wurde daraus eine *Master-Liste* erstellt (*UX problem records*). In einem letzten Schritt werden gewöhnlich die Usability-Probleme gruppiert (*UX problem groups*). In der vorliegenden Studie wurde auf diesen Schritt verzichtet, da die konkrete Anzahl der Usability-Probleme von Interesse war und nicht die Problemgruppen.

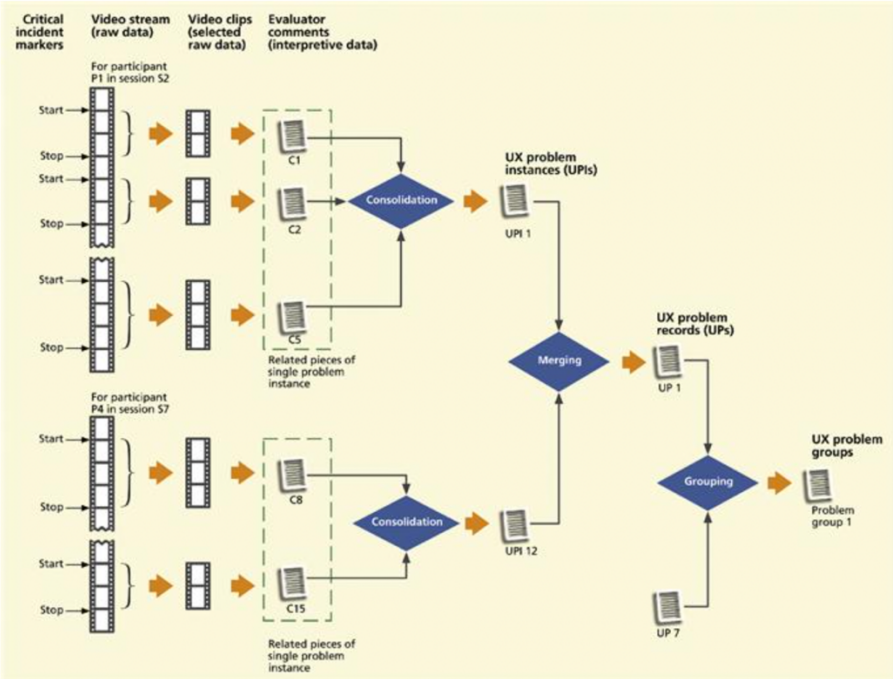


Abbildung 16. Konsolidierung, Zusammenführung und Gruppierung von UX-Problemen aus Hartson und Pyla (2012, S. 562). Copyright 2012 bei Elsevier.

Diese Master-Liste (siehe Anhang A) enthielt pro Problem einen kurzen Beschrieb, bei welchem Schritt im NAP das Problem vorkam, was der Schweregrad war, bei welcher Testperson das Problem auftrat, von welchen Expert*innen das Problem vorhergesagt wurde und von welchen Evaluator*innen es beobachtet wurde. Der Schweregrad wurde unter Berücksichtigung der Einschätzungen der beiden Evaluator*innen und der fünf Expert*innen vom Autor festgelegt. Weiter wurde pro Problem festgehalten, ob es sich um einen Hit, einen falschen Alarm oder ein verpasstes Problem handelte. Der geschilderte Auswertungsprozess ist in Abbildung 17 dargestellt.

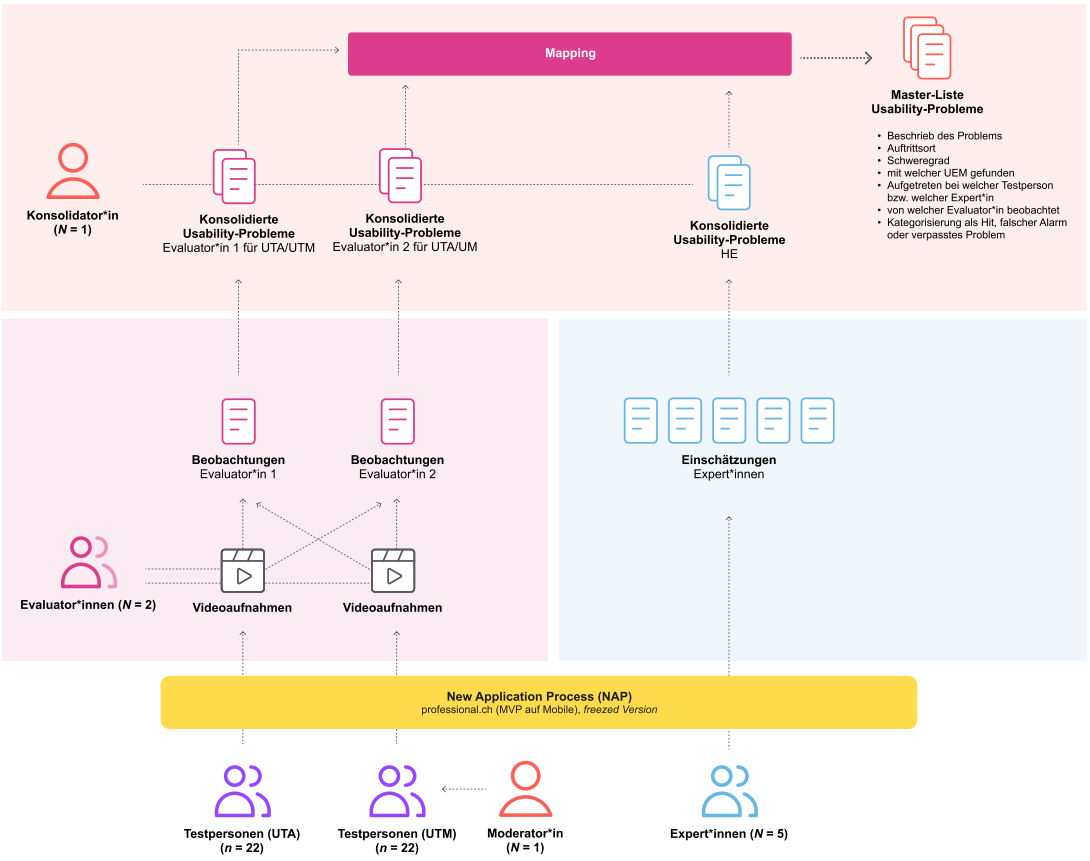


Abbildung 17. Visualisierung des Auswertungsprozesses der vorliegenden Studie.

Die Konsolidierung der Usability-Probleme erfolgte nahe am Datenmaterial der Expert*innen und Evaluator*innen. Dies bedeutet, dass Kommentare möglichst als separate Probleme betrachtet und nicht zu übergeordneten Problemgruppen zusammengefasst wurden (z. B. Kommentare zur Berufserfahrung wie «Berufserfahrung unklar, da nicht gearbeitet», «[…] unklar, da verschiedene

Erfahrungen» und «[…] unklar, ob Lehre dazugehört» wurden als separate Probleme betrachtet und nicht als ein übergeordnetes Problem gruppiert, z. B. «Berufserfahrung unklar»).

Die anschliessenden quantitativen Auswertungen der Master-Liste erfolgten in JASP. Für spezifische Analysen, z. B. das Bootstrapping der Anzahl der Usability-Probleme, kam Python zum Einsatz. Die verwendeten statistischen Methoden sind im nachfolgenden Abschnitt 3.12 näher beschrieben.

Zur Berechnung der Gesamtkosten pro Methode im Sinne der zweiten Fragestellung, wurde anhand von Erfahrungswerten der Praxispartnerin ein Kostenband für den Stundensatz mit einer unteren und oberen Grenze von CHF 165 bzw. CHF 210 definiert. Zudem wurden auch die Kosten für die Rekrutierung der Testpersonen bei TestingTime sowie Lizenzkosten für Solid mitberücksichtigt. Somit konnte pro UEM für alle zeitlichen und externen Aufwände die Kosten pro Usability-Problem berechnet werden.

Um die Ergebnisse aus den beiden Fragestellungen zu erweitern und weiterführende Erkenntnisse für Handlungsempfehlungen zu gewinnen, wurden verschiedene zusätzliche Analysen durchgeführt. So wurden die Kennzahlen Actual Effectiveness und Actual Efficiency berechnet. Für jede UEM wurde analysiert, welche Probleme ausschliesslich durch eine Methode aufgedeckt wurden (*einzigartige Probleme*) und welche durch die anderen Methoden unentdeckt blieben. Ausserdem wurde untersucht, ob die UEM Usability-Probleme in unterschiedlichen Schritten des NAP identifizierten.

3.12 Statistische Methoden

In folgendem Abschnitt werden die statistischen Methoden erläutert, die zur Verifikation der Hypothesen sowie für die weiterführende Analyse angewendet wurden.

3.12.1 Anzahl Usability-Probleme

Um das Konfidenzintervall für die Anzahl gefundener Usability-Probleme pro zu berechnen, wurde das *angepasste Konfidenzintervall nach Wald* (engl. *adjusted-Wald binomial confidence interval*) verwendet, da es sich besonders gut für kleinere Stichproben ($n < 100$) eignet (Sauro & Lewis, 2005, 2016). Das angepasste Wald-Konfidenzintervall berücksichtigt die Tatsache, dass die Schätzwerte (*Maximum Likelihood Estimator*, MLE) nicht immer normalverteilt sind und korrigiert die Standardfehler dahingehend.

Wie im theoretischen Teil erläutert wurde, werden in der Praxis Usability-Tests oftmals mit nur wenigen Personen durchgeführt (Magic Number 5, vgl. Barnum, 2003). In der vorliegenden Arbeit wurde eine grössere Stichprobe gezogen. Um eine Übertragbarkeit der Ergebnisse sicherzustellen, wird *Bootstrapping* angewendet, um die Frage zu beantworten, was passiert wäre, wenn aus der vorliegenden Gesamtstichprobe im Usability-Test zufällig beispielsweise nur fünf Personen befragt worden wären.

Bootstrapping ist eine statistische Methode, bei der wiederholte Stichproben aus einer Datenmenge gezogen werden (sogenanntes *Resampling*) (Efron, 1979, 2000). Basierend darauf kann eine zuverlässige Schätzung von Kennzahlen, wie der Mittelwert oder die Standardabweichung, berechnet werden. Die Methode eignet sich dann, wenn die Stichprobe sehr klein oder die Verteilung der Datenmenge unbekannt ist.

3.12.2 Schweregrad der Usability-Probleme

Zur Analyse der Unterschiede der drei UEM in Bezug auf den Schweregrad wurden der Kruskal-Wallis-Test sowie anschliessend der Dunn-Bonferroni-Test angewendet, da die abhängige Variable ordinalskaliert war.

3.12.3 Kosten pro Usability-Problem

Für die Untersuchung der Unterschiede in Bezug auf die Kosten pro Usability-Problem wurde ebenfalls ein Kruskal-Wallis-Test mit anschliessendem Dunn-Bonferroni-Test durchgeführt. Da die Voraussetzung der Normalverteilung nicht gegeben war, konnte keine ANOVA durchgeführt werden. Zur Prüfung der Normalverteilung kam ein Shapiro-Wilk-Test (Shapiro & Wilk, 1965) zur Anwendung.

3.12.4 Reliabilitätsanalyse

Zur Beurteilung, wie einig die Expert*innen bzw. Evaluator*innen sich in ihren Bewertungen waren, wurde Krippendorffs Alpha (Hayes & Krippendorff, 2007) angewendet. Diese Methode ermöglicht die Berücksichtigung von mehreren Bewerter*innen (engl. *rater*) sowie nominalskalierten Daten.

3.12.5 Weiterführende Analysen

Zur Prüfung von Zusammenhängen der Einflussfaktoren und der Anzahl Usability-Probleme pro Person wurde eine Regressionsanalyse oder bei fehlenden Voraussetzungen (z. B. Skalenniveau) eine Rangkorrelation nach Spearman bzw. Korrelation nach Bravais-Pearson berechnet.

Zur Prüfung auf Unterschiede bei Häufigkeiten, wurde ein Pearson Chi-Quadrat-Anpassungstest verwendet. Dieser kam z. B. bei der Anzahl einzigartiger Probleme sowie dem Auftrittsort von Problemen je nach UEM zum Einsatz.

Bei Gruppenvergleichen mit intervallskalierten Variablen (z. B. Alter) wurde ein Student t-Test angewendet, da die Stichprobenverteilung bei $n > 30$ aufgrund des zentralen Grenzwertsatzes annähernd normalverteilt ist und daher auf eine Überprüfung der Normalverteilung verzichtet werden kann (Bortz & Schuster, 2010). Falls die Varianzhomogenität verletzt war, wurde stattdessen ein Welch-Test durchgeführt. Im Falle von ordinalskalierten Variablen wurde der Mann-Whitney-U-Test angewendet.

4 ERGEBNISSE

Im Nachfolgenden werden die Ergebnisse nach Fragestellung gegliedert dargestellt und mit weiterführenden Analysen abgerundet. Die Diskussion der Ergebnisse erfolgt im Kapitel 5.

4.1 Fragestellung 1: Inwiefern unterscheiden sich die gefundenen Usability-Probleme, die durch die verschiedenen Usability Methoden erhoben wurden?

Zunächst wird die Anzahl der Hits je nach UEM und dann deren Schweregrad näher analysiert. Anschliessend erfolgt jeweils die statistische Hypothesenprüfung.

4.1.1 Anzahl der Probleme

Insgesamt wurden mit allen drei UEM zusammen 150 Usability-Probleme aufgedeckt, wovon 115 in den Varianten des Usability-Tests mit Testpersonen zum Vorschein kamen. Durch die Methode UTM wurden mit 89 die meisten Usability-Problemen (95.0 % KI [79.1, 96.9]) identifiziert, gefolgt vom UTA mit 59 Problemen (95.0 % KI [48.6, 69.4]). Am wenigsten echte Probleme wurden mit 19 durch die HE gefunden (95.0 % KI [12.4, 28.1]).

Zur Prüfung der Hypothese H1a, dass mittels eines Usability-Tests mehr echte Probleme gefunden werden als mit einer HE, wurden die Häufigkeiten und deren Konfidenzintervalle verglichen, wie Abbildung 18 zeigt. Demnach werden mit Usability-Tests signifikant mehr Probleme identifiziert als mit einer HE.

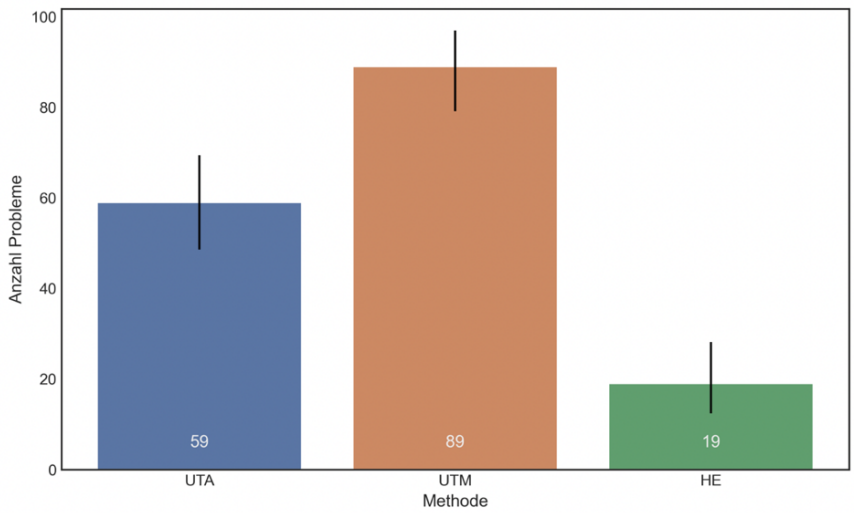


Abbildung 18. Anzahl gefundener Usability-Probleme (Hits) nach Methode mit 95% KI (angepasstes Konfidenzintervall nach Wald).

Neben den 19 Hits (12.6 %) der HE waren 35 Probleme falsche Alarme (23.2 %). Insgesamt gab es 96 Probleme (64.0 %), durch die beiden Usability-Tests aufgedeckt wurden, mit der HE jedoch verpasst wurden.

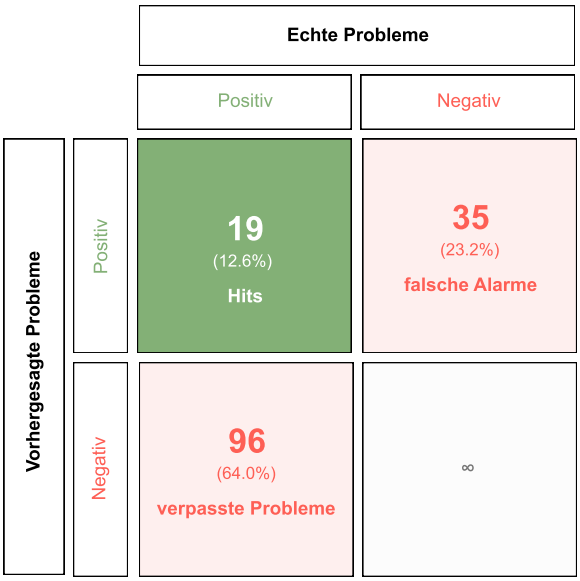


Abbildung 19. Anzahl Hits, falscher Alarme und verpasster Probleme der HE.

Die Bootstrapping-Resultate zeigten zwei Aspekte. Erstens, dass bei wenigen Personen aufgrund der überlappenden Konfidenzintervalle noch keine klaren Unterschiede erkennbar waren. Eine Ausnahme ergab sich unter Berücksichtigung der Hits. Hier war ersichtlich, dass über die HE bereits bei fünf Personen signifikant weniger Probleme auffindbar waren als über den UTM. Zweitens zeigte sich, dass zwischen dem UTA und dem UTM ab circa 10 Personen Unterschiede auftraten. Abbildung 20 veranschaulicht die Bootstrapping-Resultate mit den Konfidenzintervallen.

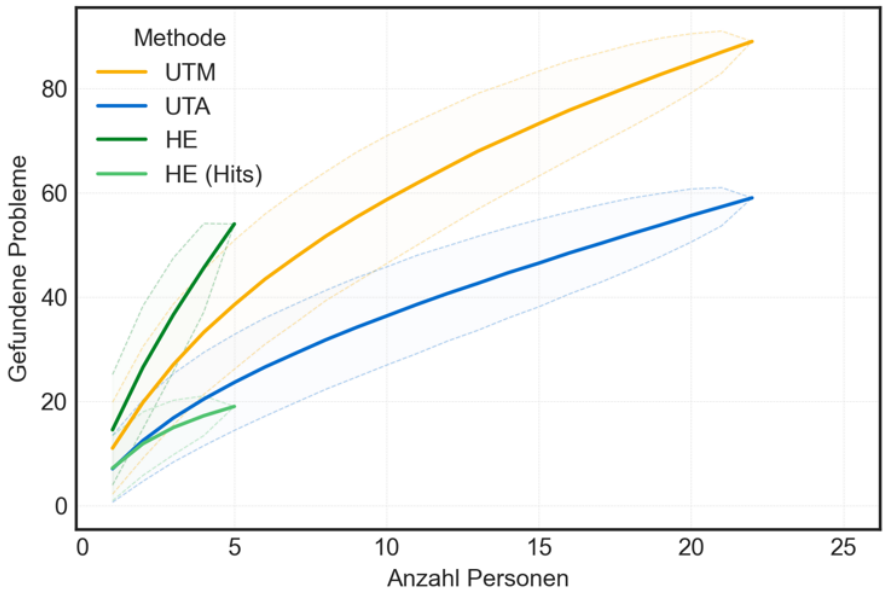


Abbildung 20. Vergleich der Anzahl Usability-Probleme der Methoden UTM, UTA und HE mit dem 95% Konfidenzintervall basierend auf 10'000 Bootstrap-Replikationen (ohne Zurücklegen).

Eine einzelne Expert*in identifizierte im Durchschnitt 7 Probleme, eine Testperson beim UTM 11 Probleme und beim UTA 7 Probleme. Die Tabelle mit den detaillierten Bootstrap-Resultaten ist in Anhang B zu finden.

4.1.2 Schweregrad

Der Median des Schweregrads der Probleme, die mittels der Methode HE gefunden wurden, liegt bei 4 (*SD* = 1.3), beim UTA liegt er bei 3 (*SD* = 1.0) und beim UTM bei 2 (*SD* = 1.04).

Zur Prüfung der Hypothese H1b, ob sich der Schweregrad der Usability-Probleme je nach Methode unterscheidet, wurde aufgrund der Ordinalskala des Schweregrads ein Kruskal-Wallis-Test durchgeführt. Dieser zeigte, dass der Schweregrad durch die verwendete UEM beeinflusst wurde ($\chi^2(2) = 6.74$, $p = .034$). Die im Anschluss durchgeführten Post-hoc-Tests (Dunn-Bonferroni) zeigten, dass sich nur die HE und der UTM ($z = 2.49$, $p = .019$) signifikant unterschieden. Es handelt sich dabei nach Cohen (1992) um einen mittleren Effekt ($r = 0.24$). Zwischen HE und UTA ($p = 0.13$) sowie UTA und UTM ($p = 0.13$) war kein Unterschied feststellbar.

Somit wurden mit der HE eher schwerwiegendere Probleme gefunden als mit den beiden Usability-Test-Methoden. Dies steht im Widerspruch zur Hypothese H1b und die Nullhypothese muss beibehalten werden.

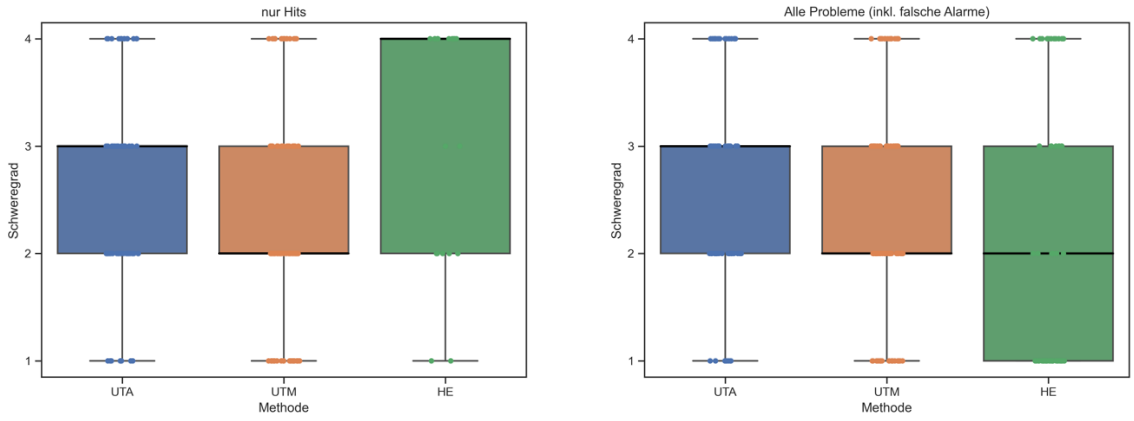


Abbildung 21. Schweregrad der Usability-Probleme nach Methode.

Ergänzend wurde analysiert, ob Unterschiede bestehen, wenn bei der HE auch die falschen Alarme einbezogen werden. Der Median für die HE lag nun bei 2 ($SD = 1.25$). Ein Kruskal-Wallis-Test bestätigte, dass Unterschiede bestanden ($\chi^2(2) = 6.196$, $p = .045$). Die Post-hoc Tests (Dunn-Bonferroni) zeigten Unterschiede zwischen der HE und dem UTA ($z = -2.49$, $p = .019$). Somit wurden durch den UTA eher schwerwiegendere Probleme gefunden als durch die HE. Zwischen HE sowie UTM ($p = 0.23$) und UTA sowie UTM ($p = 0.28$) konnten keine signifikanten Unterschiede festgestellt werden.

Die Tabelle mit den vollständigen Resultaten ist in Anhang C zu finden.

4.2 Fragestellung 2: Was kostet das Auffinden eines einzelnen Usability Problems bzw. wieviel Aufwand verursacht eine Usability-Evaluationsmethode?

Insgesamt war die HE die sparsamste Methode mit knapp 27 Stunden und der UTM mit gut 79 Stunden die aufwändigste. Der Aufwand für die Vorbereitung war bei allen drei Methoden mit ca. 12–13 Stunden ähnlich. Unterschiede zeigten sich in der Phase Durchführung sowie Analyse und Konsolidierung. So war die Methode UTA mit den automatisierten Testsitzungen in der Durchführung mit etwas mehr als zwei Stunden von den drei UEM am effizientesten. Auch die Analyse und die Konsolidierung der gefundenen Usability-Probleme dauerte weniger lang als bei der moderierten Variante. Wurde der Aufwand mit der Anzahl der gefundenen Usability-Probleme ins Verhältnis gesetzt, lag er pro gefundenem Problem für die HE bei 1.41 Stunden und war somit höher als bei den anderen Methoden. Die beiden Usability-Test-Varianten UTA und UTM waren mit 0.81 bzw. 0.89 Stunden pro Problem ähnlich aufwändig.

Tabelle 11

Aufwand der einzelnen UEM und pro Usability-Problem (Hits).

	UTA	**UTM**	**HE**
gefundene Probleme (Hits)	59	89	19
Vorbereitung [h]	13.18	12.73	12.08
Durchführung [h]	2.25	17.75	11.79
Analyse und Konsolidierung [h]	32.08	48.64	3.00
Total [h]	47.52	79.12	26.88
Aufwand pro Usability-Problem [h]	**0.81**	**0.89**	**1.41**

Unter Annahme des Kostenbands war die Methode HE absolut gesehen die günstigste Methode mit circa CHF 4’440–5’650. Der UTA war etwas mehr als doppelt so teuer (ca. CHF 9’700–11’800) und der UTM kostete fast knapp das Vierfache einer HE (CHF 17’100–20’660). Jedoch waren die Kosten für das Auffinden eines einzelnen Usability-Problems mit der Methode UTA mit ca. CHF 165–200 am günstigsten, gefolgt vom UTM mit ca. CHF 190–235. Am teuersten war die HE mit ca. CHF 235–300.

Tabelle 12

Gesamtkosten sowie Kosten pro gefundenem Usability-Problem (Hits), aufgeschlüsselt nach den Methoden HE, UTA und UTM.

	UTA	**UTM**	**HE**
gefundene Probleme (Hits)	59	89	19
Arbeitsaufwand [h]	47.5	79.1	26.9
Stundensatz [CHF]			
CHF 165	7’838	13’052	4’439
CHF 210	9’975	16’611	5’649
Rekrutierungskosten [CHF]	1'816	4'039	0
Lizenzkosten [CHF]	1’080	0	0
Total [CHF]	10’734 – 12’871	17'090 – 20’650	4'439 – 5’649
Kosten pro gefundenem Problem (Hit) [CHF]	**182 – 218**	**192 – 232**	**234 – 297**

Da die Annahme der Normalverteilung verletzt war (Shapiro-Wilks-Test; $p < .001$), wurde zur Überprüfung der Hypothese 2, ein Kruskal-Wallis-Test durchgeführt. ($\chi^2(2) = 38.12$, $p < .001$). Die daraufhin durchgeführten Post-hoc-Tests (Dunn-Bonferroni-Tests) zeigten, dass die HE gegenüber dem UTA ($z = 6.11$, $p < .001$, $r = 0.57$) sowie dem UTM ($z = 4.9$, $p < .001$, $r = 0.47$) höhere Kosten aufwies. Hierbei handelt es sich nach Cohen (1992) um starke Effekte. Zwischen UTA und UTM

konnte kein Unterschied festgestellt werden (z = -1.61, p =0.21). In Übereinstimmung mit der Hypothese 2 waren somit die Kosten pro Problem bei Usability-Tests tiefer als bei einer HE.

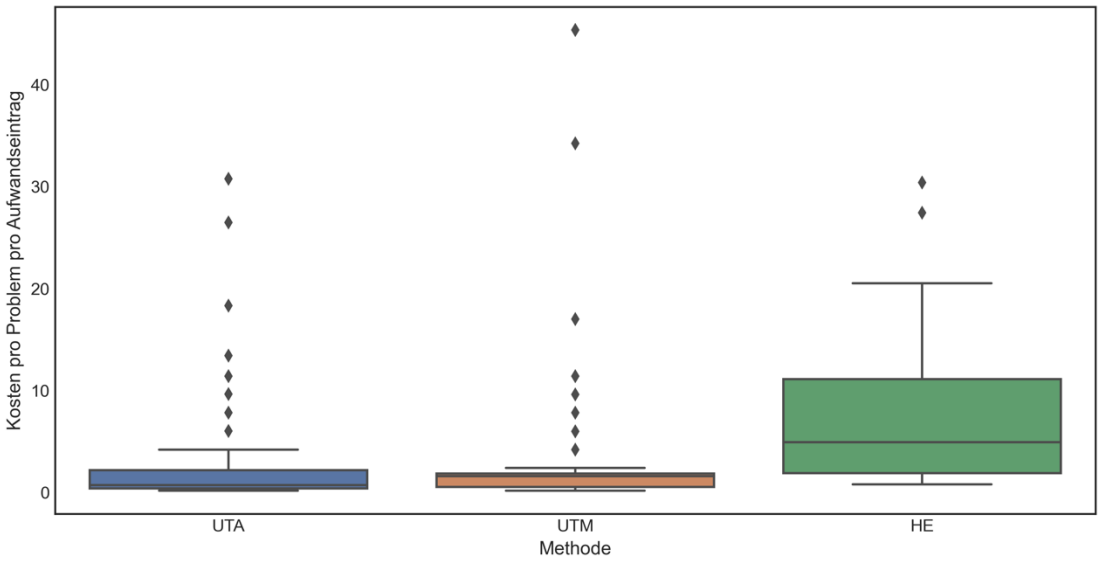


Abbildung 22. Kosten pro Problem pro Aufwandseintrag gruppiert nach den Methoden HE, UTA und UTM.

4.3 Weiterführende Analysen

Nachfolgend werden die Auswertungen der Einflussfaktoren, der Reliabilität und der Kennzahlen Actual Effectiveness sowie Actual Efficiency vorgestellt. Danach werden die einzigartigen Probleme, der Auftrittsort der Probleme sowie abschliessend die QD ausgeführt.

4.3.1 Einflussfaktoren

Zur Prüfung auf mögliche Zusammenhänge mit der Anzahl gefundener Usability-Problemen in Bezug auf die persönliche Relevanz, die digitale Affinität, die Tageszeit sowie die Antworten aus den Instrumenten PANAS, PSSUQ und NASA-TLX statistische Analysen durchgeführt.

Persönliche Relevanz. Testpersonen bei der Methode UTA wiesen eine leicht geringere persönliche Relevanz auf (*Md* =4, *SD* = 1.27) gegenüber Personen der Methode UTM (*Md* =5, *SD* = 0.96). Ein exakter Mann-Whitney-U-Test zeigte jedoch keine signifikanten Unterschiede

(U = 211.00, p = 0.62). Es war keine signifikante Korrelation zwischen der Anzahl gefundener Usability-Probleme pro Person und Anzahl Bewerbungen feststellbar (r_s = -0.15, p = 0.33, n = 44).

Digitale Affinität. Testpersonen in der Gruppe UTA (*Md* =6.0, *SD* = 0.63) und UTM (*Md* =5.5, *SD* = 0.80) wiesen eine ähnliche digitale Affinität auf. Hier konnten ebenfalls keine Unterschiede festgestellt werden (Exakter Mann-Whitney-U-Test, U = 263.00, p = 0.40). Es war keine signifikante Korrelation zwischen der Anzahl gefundener Usability-Probleme und der digitalen Affinität feststellbar (r_s = -0.016, p = 0.92, n = 44).

Tageszeit. Personen der Gruppe UTA haben den Test eher gegen Abend (M = 16.48, SD = 5.1) durchgeführt, Personen der Gruppe UTM am Mittag (M = 11.84, SD = 3.2). Da die Varianzhomogenitätsannahme verletzt war (Levene-Test; p = .042), wurde ein Welch-Test durchgeführt (t(31.51) = 3.46, p = .002). Demnach haben Testpersonen beim UTA die Sitzungen signifikant zu einer anderen Tageszeit durchgeführt als beim UTM. Dieser Effekt ist nach Cohen (1992) stark (r = 0.53). Es konnte keinen Einfluss der Tageszeit auf die Usability-Probleme pro Person angenommen werden (F(1, 40) = 3.08, p = 0.09).

Affekt, subjektive Zufriedenheit und mentale Belastung. Bei den Instrumenten zur subjektiven Zufriedenheit (PSSUQ) und mentale Belastung (NASA-TLX) gab es keine signifikanten Unterschiede zwischen den Gruppen. Beim emotionalen Zustand (PANAS) zeigten sich welche. So fühlten sich Testpersonen in der Gruppe UTM nach dem Usability-Test signifikant aktiver, interessierter, freudig erregter, angeregter, stolzer, begeisterter, wacher, nervöser, entschlossener und aufmerksamer als die Personen aus der Gruppe UTA (siehe Tabelle 13). Die Effektstärke nach Cohen (1992) entsprachen mittleren bis starken Effekten. Die vollständige Tabelle findet sich in Anhang G.

Tabelle 13

Gruppenunterschiede der emotionalen Zustände im UTA und UTM.

	U	*p*	*r*	95% KI Unten	Oben
aktiv	122.5	0.010*	-0.44	-0.68	-0.12
interessiert	138	0.032*	-0.37	-0.63	-0.04
freudig erregt	143.5	0.048*	-0.35	-0.62	-0.01
angeregt	108	0.004**	-0.51	-0.72	-0.21
stolz	128.5	0.016*	-0.42	-0.66	-0.09
begeistert	132.5	0.023*	-0.40	-0.65	-0.07
wach	121.5	0.010*	-0.45	-0.68	-0.13
nervös	155	0.035*	-0.30	-0.58	0.05
entschlossen	114	0.006**	-0.48	-0.71	-0.17
aufmerksam	105.5	0.003**	-0.52	-0.73	-0.22

** Signifikant α =0.05, ** signifikant α =0.01*
r = .30 entspricht einem mittleren Effekt, r = .50 entspricht einem starken Effekt (Cohen, 1992)

Die Anzahl Usability-Probleme pro Person korrelierten signifikant mit den Items des PANAS: verärgert (r_s = 0.35. p = 0.02), erschrocken (r_s = 0.33, p =0.03), gereizt (r_s = 0.38, p = .012) und ängstlich (r_s = 0.41, p = .007). Beim PSSUQ korrelierten die Items «Die Informationen halfen mir erfolgreich dabei, die Aufgaben zu lösen» (r_s = 0.35, p = 0.02) sowie «Die Bedienung von professional.ch ist angenehm» (r_s = 0.34, p = 0.03).

4.3.2 Reliabilität

Expert*innen. Krippendorffs Alpha lag bei -0.116, 95 % KI [-0.19, -0.06] und zeigte somit eine unzureichende Übereinstimmung (Hayes & Krippendorff, 2007). Werden nur die 19 Hits betrachtet, zeigte sich eine leichte Verbesserung, jedoch war diese Übereinstimmung immer noch unzureichend (Krippendorffs Alpha = -0.05, 95 % KI [-0.135, 0.024]).

Keines der durch die Expert*innen vorhergesagten Probleme wurde von allen fünf Expert*innen genannt. Nur zwölf Probleme wurden von zwei oder mehr Personen vorhergesagt. Am häufigsten

wurden niederschwellige Probleme von jeweils nur einer Person genannt. Abbildung 23 zeigt zusammenfassend eine *Heatmap* mit der Anzahl Probleme (inkl. falscher Alarme), die von mehreren Expert*innen gefunden wurden, aufgeschlüsselt nach dem Schweregrad.

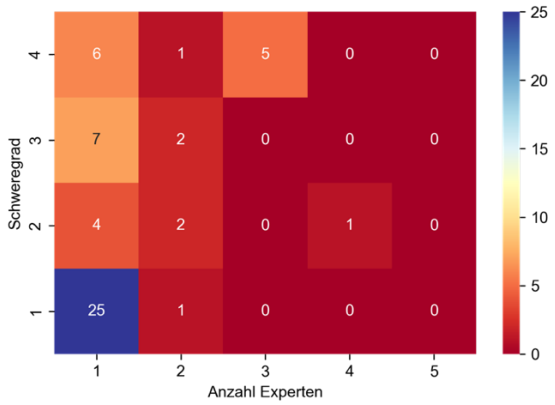


*Abbildung 23. Heatmap Anzahl Usability-Probleme (inkl. falsche Alarme) die von mehreren Expert*innen gefunden wurden, aufgeschlüsselt nach Schweregrad.*

Testpersonen. Die Testpersonen beim UTA (Krippendorffs Alpha = 0.19, 95 % KI [0.095, 0.28]) sowie beim UTM (Krippendorffs Alpha = 0.118, 95 % KI [0.06, 0.18]) zeigten jeweils eine leichte Übereinstimmung.

Evaluator*innen. Auch in diesem Fall wurde Krippendorffs Alpha berechnet. Die Evaluator*innen zeigten eine schlechte Übereinstimmung (α = -0.14, 95 % KI [-0.55, -0.34]).

4.3.3 Actual Effectiveness und Actual Efficiency

Der UTM war über alle Schweregrade am wirkungsvollsten, um Probleme zu identifizieren. Die HE war zum Auffinden von schwerwiegenden Problemen (Stufe 4) am effizientesten. Mittlere Probleme (Schweregrad 2 und 3) wurden mit dem UTA am effizientesten identifiziert.

Tabelle 14

Actual Effectiveness und Actual Effiency der Methoden UTA, UTM und HE aufgeschlüsselt nach Schweregrad.

Kennzahl	**Schweregrad**			
	1	2	3	4
Actual Effectiveness				
UTA	0.29	0.5	0.62	0.74
UTM	**0.75**	**0.71**	**0.85**	**0.84**
HE	0.0	0.09	0.02	0.44
Actual Efficiency				
UTA	0.17	**0.44**	**0.34**	0.29
UTM	**0.27**	0.38	0.28	0.20
HE	0.07	0.19	0.07	**0.37**

fett *= höchster Wert*

4.3.4 Einzigartige Probleme

Es zeigte sich, dass durch den UTM mit 38 am meisten Probleme gefunden wurden, die durch die beiden anderen Methoden unentdeckt blieben. Der UTA deckte 14 einzigartige Probleme auf und die HE nur eines. Diese Unterschiede sind jedoch nicht signifikant, wie ein Chi-Quadrat-Anpassungstest zeigt (χ^2 (6, n = 53) = 11.825, p = 0.07).

Tabelle 15

Einzigartige Usability-Probleme, die jeweils nur durch eine Methode gefunden wurden.

	Schweregrad				
Methode	1	2	3	4	Total
UTA	4	6	2	2	14
UTM	15	12	9	2	38
HE	0	0	0	1	1

4.3.5 Auftrittsort

Am häufigsten traten Probleme beim Schritt 6 «Berufsweg 1» mit 22 auf, gefolgt von Schritt 3 «Meine Daten» mit 17 Problemen und Schritt 9 «Profil und Login» mit 16 Problemen. Eine ausführliche Übersicht ist in Abbildung 24 dargestellt. Ein Chi-Quadrat-Anpassungstest zeigte, dass die beobachteten Häufigkeiten nicht signifikant von den erwarteten Häufigkeiten abwichen (χ^2(28, n = 202) = 43.1, p = 0.07). Damit kann angenommen werden, dass der Auftrittsort der gefundenen Probleme nicht von der jeweiligen UEM abhängig ist.

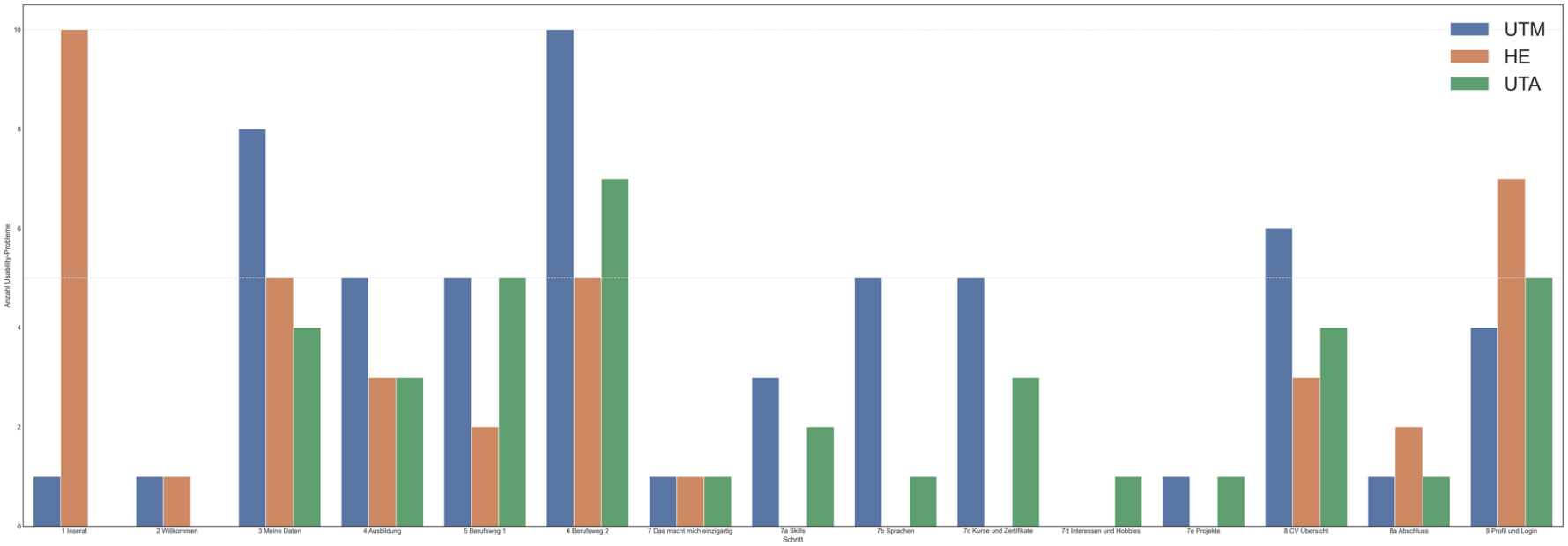


Abbildung 24. Auftrittsort der Usability-Probleme im Prototyp, gruppiert nach den UEM.

4.3.6 Quick and Dirty-Auswertung

Bei der QD wurden 36 Usability-Probleme identifiziert, was circa 24 % aller Usability-Probleme bzw. 31 % der Hits entspricht. Tabelle 16 zeigt die Anzahl Probleme aufgeschlüsselt nach Schweregrad.

Tabelle 16

Identifizierte Usability-Probleme in der Variante Quick and Dirty aufgeschlüsselt nach Schweregrad.

	Schweregrad				
	1	2	3	4	*Total*
Usability-Probleme	3	10	12	11	36

Der Gesamtaufwand für QD betrug 35.4 Stunden, was knapp einer Stunde pro Usability-Problem entspricht. Die kalkulierten Gesamtkosten für diese Variante lagen bei CHF 9'880–11'470. Dies entspricht ca. CHF 275–320 pro gefundenem Usability-Problem. Die Methode QD war demnach die teuerste Methode zum Auffinden eines einzelnen Usability-Problems.

5 DISKUSSION

In diesem vorletzten Kapitel werden zunächst die Ergebnisse entlang den Fragestellungen und Hypothesen interpretiert und diskutiert. Anschliessend folgen die kritische Reflexion und Limitierung. Abschliessend werden die Handlungsempfehlungen aufgestellt.

5.1 Zusammenfassung und Interpretation

Nachfolgend werden die gewonnenen Erkenntnisse erläutert und in Bezug auf ihre Anwendbarkeit in der Praxis interpretiert.

Im Sinne von **Fragestellung 1**, inwiefern sich die gefundenen Usability-Probleme je nach UEM unterscheiden, zeigen die Ergebnisse, dass im UTM mit 89 signifikant am meisten relevante Usability-Probleme identifiziert wurden, gefolgt vom UTA mit 59. Am wenigsten Probleme wurden in der HE mit 19 gefunden. Dies steht im Einklang mit der Hypothese H1a, dass mittels Usability-Tests mehr relevante Probleme gefunden werden als mit einer HE. Da aufgrund des Charakters der jeweiligen UEM eine unterschiedliche Anzahl Personen involviert war und es somit naheliegend ist, dass die Anzahl Probleme sich unterscheidet, wurde mittels Bootstrapping-Verfahren die durchschnittliche Anzahl Probleme pro Person berechnet. Dabei zeigte sich, dass es keine signifikanten Unterschiede bei einer geringeren Anzahl Personen gibt, ausser es werden nur die Hits betrachtet. In diesem Fall ist bereits bei fünf Personen ein UTM effektiver als eine HE. Zwischen dem UTA und dem UTM sind ab circa zehn Personen Unterschiede feststellbar.

Anschliessend wurde untersucht, ob sich der Schweregrad der gefundenen Usability-Probleme je nach Methode unterscheidet. Die HE findet signifikant schwerwiegendere Probleme, wenn nur Hits angeschaut werden. Die Resultate stehen somit im Widerspruch zur Hypothese H1b, wonach Usability-Tests mehr schwerwiegende Probleme identifizieren. Es wurde ein Effekt identifiziert, jedoch nicht in die angenommene Richtung. Dieser Umstand bedarf einer weiterführenden Analyse (siehe Kapitel 6 Fazit und Ausblick). Hingegen sind die Probleme insgesamt signifikant weniger

schwerwiegend, wenn alle Probleme (inkl. der falschen Alarme) berücksichtigt werden. In der Praxis fehlt jedoch oft das Wissen, ob es sich um Hits handelt, da es meist an Vergleichsdaten mangelt.

Zwischen dem UTA und dem UTM waren keine Unterschiede in Bezug auf den Schweregrad feststellbar. Dies steht im Widerspruch zu den Ergebnissen von Hertzum et al. (2015), wonach Personen in nichtmoderierten Tests einen höheren Prozentsatz an hochrelevanten Äusserungen machen. Dies wäre demnach ein wertvolles Feld für weitere Forschung, siehe Kapitel 6.

Wird der Fokus auf die schwerwiegenden Probleme gelegt, wurden mit der HE 10 der insgesamt 19 Hits mit dem Schweregrad 4 identifiziert. Es ist aber zu betonen, dass die HE mit fünf Expert*innen durchgeführt wurde. Es ist anzuzweifeln, ob in der Praxis jeweils so viele Personen zum Einsatz kommen. Viel wahrscheinlicher ist es zu vermuten, dass solche Evaluationen jeweils nur von einer Person durchgeführt werden. Wie die Bootstrapping-Resultate gezeigt haben, erzielten einzelne Expert*innen im Schnitt nur sieben Hits der insgesamt 115 echten Probleme. Somit zeigt sich, dass Usability-Tests mit einer kleinen Stichprobe von fünf Personen mehr Hits aufdecken als einzelne Expert*innen. Werden die Anzahl der Hits, der falschen Alarme sowie der verpassten Probleme der HE verglichen, sind die Ergebnisse in Bezug auf die Hits zwar konservativer, aber decken sich sonst mit Daten aus einer Analyse von (Sauro, 2012), wie Tabelle 17 zeigt.

Tabelle 17

Hits, falsche Alarme und verpasste Usability-Probleme im Vergleich von Sauro (2016) und eigenen Ergänzungen.

	Hits	**Falscher Alarm**	**Verpasst**	**Anzahl Expert*innen**	**Anzahl Testpersonen**
vorliegende Studie	*13%*	*23%*	*64%*	*5*	*44*
Sauro, 2016	32%	9%	68%	4	50
Doubleday et al., 1997	36%	40%	39%	5	20
Law & Hvannberg 2002	30%	38%	32%	2	10
Law & Hvannberg 2004	43%	46%	48%	18	19
Hvannverg et al. 2006	40%	37%	60%	10	10
M	*32%*	*32%*	*52%*	7.3	*25.5*

Sauro (2012) weist darauf hin, dass Probleme, die mit der HE, aber nicht mittels Usability-Tests gefunden wurden, nicht zwingend falsche Alarme sein müssen. Es könnte auch sein, dass diese Probleme schlicht bei viel weniger Nutzenden auftreten (tiefer p-Wert). Sauro (2012) plädiert daher dafür, diese Probleme eher als Probleme mit einer sehr tiefen Auftrittswahrscheinlichkeit (sogenannte *Long-Tail-Probleme*) zu klassifizieren. Jedoch bleibt das Risiko, dass dadurch womöglich falsch priorisiert wird und Ressourcen nicht effektiv investiert werden. Im schlimmsten Fall resultieren aus diesen Anpassungen neue Probleme.

Weiter unterstützen die vorliegenden Erkenntnisse die Kritik an der Magic Number 5 (vgl. Barnum, 2003). Wären bei den Usability-Tests jeweils fünf Personen eingesetzt worden, wird anhand der Bootstrapping-Resultate ersichtlich, dass mittels UTA circa 21 % bzw. beim UTM ca. 34 % der insgesamt 115 echten Probleme hätten identifiziert werden können. Auch bei der HE haben die fünf

Expert*innen zusammen mit 19 Hits nur ca. 16 % identifiziert. Tabelle 18 zeigt die Resultate für jeweils fünf Personen.

Tabelle 18

Anzahl gefundener Usability-Probleme. Bootstrap-Resultate basierend auf 10'000 Replikationen.

Anzahl Personen	**Methode**		
	UTA	UTM	HE
5	24 (20.9%)	39 (33.2)%	19 (16.5)%
7	29 (25.2%)	48 (41.0%)	-
12	41 (35.6%)	65 (56.5%)	-
15	47 (40.9%)	73 (63.5%)	-

Für die festgestellten emotionalen Unterschiede in den Gruppen UTA und UTM kann es mehrere Erklärungen geben. Zum einen könnte die Interaktion mit dem Moderator im Sinne eines *Interviewer-Effekts* (VandenBos, 2016) einen Einfluss gehabt haben, wonach die Präsenz und das Auftreten des Moderators Einfluss auf das Verhalten und auf die Antworten der Testpersonen hat. Andererseits könnte der Avatar des automoderierten Usability-Tests im Sinne des *Uncanny-Valley-Effekts* (Skjuve, Haugstveit, Følstad & Brandtzaeg, 2019) einen negativen Einfluss auf die Testpersonen gehabt haben. Dieser Effekt beschreibt, dass menschenähnliche Computeranimationen, die einen zu hohen Grad an Ähnlichkeit mit dem menschlichen Aussehen oder Verhalten aufweisen (z. B. ein Avatar), unangenehme Gefühle auslösen können. Es wäre auch eine Kombination der beiden Gründe denkbar.

Zwar haben Testpersonen der beiden Gruppen UTA und UTM die Testsitzungen an unterschiedlichen Tageszeiten durchgeführt, jedoch gab es keinen Hinweis auf einen Zusammenhang zur Anzahl Usability-Probleme pro Testperson. Eine Erklärung für die unterschiedlichen Tageszeiten

ist, dass moderierte Sitzungen eher während Arbeitszeiten stattfinden, hingegen automoderierte Sitzungen zu einer von der Testperson beliebig gewählten Tageszeit.

Wie mit der Reliabilitätsanalyse aufgezeigt werden konnte, weisen die Expert*innen eine ungenügende Übereinstimmung auf. Zusammenfassend lässt sich sagen, dass die Resultate der HE eine tiefere Validität als die vom UTA bzw. vom UTM haben und eine tiefe Reliabilität (Wirtz, 2021) aufweisen. So kann zwar eine höhere Variabilität in einem formativen Usability-Evaluationsprozess von Vorteil sein, da dadurch viele verschiedene Probleme zum Vorschein kommen, jedoch hat sich gezeigt, dass im vorliegenden Fall diese Ergebnisse tendenziell auch verzerrt sind, da nur 19 der 54 (ca. 35 %) durch die HE identifizierten Probleme auch Hits waren. Auch bei den Evaluator*innen zeigt sich eine unzureichende Übereinstimmung. Dies verdeutlicht die Notwendigkeit, jeweils mehrere Expert*innen bzw. Evaluator*innen für eine Beurteilung hinzuzuziehen.

Werden alle Probleme (inkl. falscher Alarme) betrachtet, war neben der tiefen Reliabilität und den falschen Alarmen auffallend, dass die von den Expert*innen genannten Probleme inhaltlich teilweise sehr detailliert waren. So wurde z. B. bemängelt, dass im Job-Inserat die Rolle der Ansprechperson nicht ersichtlich gewesen sei oder dass der Cookie-Banner farblich zu wenig als solcher erkennbar gewesen sei. Eine mögliche Erklärung ist, dass Expert*innen bestrebt sind, konsistent zu erscheinen. Dies erhöht den Druck, Ergebnisse zu liefern (*kognitiven Dissonanz*, Wirtz, 2021). Weiter besteht das Risiko, dass Expert*innen davon ausgehen, dass die eigenen Meinungen und Überzeugungen auch von anderen Personen geteilt werden, im Sinne des False-Consensus-Effekts (Ross et al., 1977). Im Kontext einer HE bedeutet dies, dass Expert*innen bei der Bewertung eines interaktiven Systems ihre eigenen Ansichten übermässig auf die potenziellen Nutzer*innen projizieren und z. B. eine Funktion fälschlicherweise als intuitiv bzw. problematisch erachten.

Im Rahmen der weiterführenden Analysen wurde überprüft, welche Probleme nur durch eine einzige Methode aufgefunden wurden und durch die beiden anderen unentdeckt blieben. Hier ist der UTM am effektivsten, jedoch sind die Unterschiede statistisch nicht signifikant. Ebenfalls wurde analysiert, ob die UEM bei verschiedenen Schritten im Prototyp Probleme identifizieren. Hier sind keine signifikanten Unterschiede feststellbar.

Gestützt auf die Erkenntnisse von Hertzum et al. (2015) kann eine Erklärung, weshalb mittels des UTM signifikant mehr Usability-Probleme aufgedeckt werden konnten, die Möglichkeit zum Nachfragen sein (engl. *Probing*, Birns, Joffre, Leclerc & Paulsen, 2002). So können verbale Äusserungen von Testpersonen aufgegriffen und vertieft werden. Dadurch können neue Perspektiven entwickelt werden. Beispielsweise erzählte eine Testperson, dass es bei der hohen Anzahl an Bewerbungen schwer war, einen Überblick zu behalten. Auf Nachfrage des Moderators, wie die Testperson damit umgegangen sei, erwähnte die Testperson eine selbsterstellte Excel-Datei, in der die Bewerbungen verwaltet wurden. Solche Materialien können wertvolle Hinweise für das Verständnis des Nutzungskontexts liefern und für die Weiterentwicklung von professional.ch von Nutzen sein. Es ist anzunehmen, dass eine solche Entdeckung in einem nichtmoderierten Test oder einer HE nicht zum Vorschein gekommen wäre.

Fragestellung 2 zu den Kosten des Auffindens eines einzelnen Usability-Problems bzw. dazu, welche Kosten eine UEM verursacht, konnte mittels einer Kostengegenüberstellung in Abschnitt 4.2 beantwortet werden. Dabei wurden die direkt messbaren, einmaligen Kosten berücksichtigt, die für die Durchführung einer UEM notwendig sind, wie der Arbeitsaufwand, die Rekrutierungskosten sowie die Lizenzkosten.

Der Vorbereitungsaufwand war bei allen Methoden ähnlich, aber es zeigten sich Unterschiede in der Durchführung, der Analyse und der Konsolidierung. Die direkten Kosten wurden auf Basis von Erfahrungswerten der Praxispartnerin berechnet. Ebenfalls wurden die Rekrutierungskosten des Drittanbieters TestingTime berücksichtigt. Die deskriptiven Ergebnisse zeigten, dass in Bezug auf die Gesamtkosten die HE die günstigste Methode ist, gefolgt vom UTA und vom UTM. Jedoch sind die Kosten pro gefundenem Usability-Problem beim UTA am tiefsten, gefolgt vom UTM. Am teuersten ist die HE. Es konnte die Hypothese bestätigt werden, dass die beiden Usability-Test-Varianten pro gefundenem Usability-Problem signifikant günstiger sind als die HE. Zwischen den beiden Usability-Test-Methoden konnten hingegen keine Unterschiede festgestellt werden.

Tabelle 19

Gesamtvergleich der Usability-Evaluationsmethoden UTA, UTM, HE und QD.

	UTA	**UTM**	**HE**	**QD**
Identifizierte Probleme	59	89	54	36
Hits	59	89	19	36
Falsche Alarme	0	0	35	0
Verpasste Hits	56	26	96	79
Verpasst gesamt	92	62	96	114
Gesamtkosten [CHF]	9'654 –11’791	17'090 – 20’650	4'439 – 5’649	9'880 – 11’470
Kosten pro Hit [CHF]	164 – 200	192 –232	234 – 297	274 – 318

Anhand der Kennzahlen Actual Effectiveness und Actual Efficiency wurde deutlich, dass von den drei UEM der UTM Usability-Probleme am effektivsten identifiziert. Die HE ist insbesondere dann ressourcenschonend, wenn es sich um schwerwiegende Probleme handelt. Zusammenfassend zeigt Tabelle 19 den Gesamtvergleich aller verwendeten UEM sowie der Quick & Dirty-Auswertung.

5.2 Kritische Reflexion und Limitierung

Nachfolgend wird auf verschiedene Aspekte dieser Masterarbeit kritisch eingegangen. Diese werden reflektiert und es werden Limitierungen aufgezeigt.

Anhand des Four-Factor Framework of Contextual Fidelity (Sauer et al., 2019) wurde aufgezeigt, dass es eine Reihe von Einflussfaktoren gibt, welche die Resultate beeinflussen können. So stellt sich z. B. die Frage, ob mit einem weniger ausgereiften Prototypen vergleichbare Resultate aufgetreten wären. Es ist offen, ob die vorliegenden Ergebnisse mit Produkten repliziert werden können, die noch weit am Anfang einer Entwicklung stehen und wenig fortgeschritten sind. Der vorliegende Prototyp war bereits ausgereift und stand als klickbare HTML-Version zur Verfügung.

Weiter steht in Bezug auf die Testumgebung offen, ob die Resultate z. B. mit einer anderen Moderation ähnlich ausgefallen wären oder inwiefern die Aufgabenstellungen einen Einfluss hatten. Bei der Datenerhebung ist kritisch anzumerken, dass der beschriebene Konsolidierungsprozess sowie die definitive Vergabe des Schweregrads nur durch den Autor durchgeführt wurde. Die Zuordnung von einzelnen Beobachtungen zu einem Usability-Problem bzw. die Konsolidierung aus mehreren Usability-Methoden war dabei nicht immer trennscharf und unterlag einer gewissen Subjektivität.

In dieser Studie wurde die Konsolidierung der Probleme nahe an den Daten gemacht, was tendenziell mehr Usability-Probleme generiert. Zum Beispiel wurden Kommentare wie «Farben im Chat sind verwirrend» und «Chat ist einfach aufgepoppt» als separate Usability-Probleme definiert. Eine etwas grosszügigere Kategorienbildung (z. B. dass alle Kommentare zur Berufserfahrung einem Problem zugeordnet werden), hätte dazu geführt, dass die Gesamtzahl der identifizierten Probleme gesunken wäre und sich die Trefferquote der HE erhöht hätte.

Aufgrund des Charakters der HE und der kleinen Stichprobengrösse wurden die Instrumente PANAS, PSSUQ und NASA-TLX bei den Expert*innen nicht erhoben. Somit können zu möglichen Einflüssen in Bezug auf den Affekt, die Zufriedenheit und die mentale Belastung zwischen den Methoden HE und UTA bzw. UTM keine Aussagen gemacht werden.

In Bezug auf die Stichproben wurden zwar klare Selektionskriterien definiert, jedoch stellt sich die Frage, inwieweit die Erkenntnisse generalisiert werden können, wenn Testpersonen von einem professionellen Anbieter wie TestingTime rekrutiert werden können. Denkbar ist auch, dass solche Personen in Bezug auf Usability-Testing affiner sind und somit tendenziell mehr Kommentare sowie Hinweise liefern. Dies kann im vorliegenden Rahmen nicht bestätigt oder ausgeschlossen werden.

Bei der Kostenanalyse wurden die direkt messbaren Kosten, wie einmalige Arbeitsaufwände, Lizenzgebühren und Rekrutierungskosten berücksichtigt. Jedoch konnten die einmaligen Kosten, wie Equipment oder kontinuierliche Kosten, wie Serverkosten nicht berücksichtigt werden. Ebenfalls fehlten die Entwicklungskosten, die aus der Behebung der Usability-Probleme resultieren. Diese Daten standen seitens der Yousty AG nicht zur Verfügung.

Was bei der Kostenkalkulation ebenfalls nicht berücksichtigt wurde, ist der Schweregrad der Usability-Probleme. Ein Problem mit einem höheren Schweregrad hat für ein Unternehmen eine grössere Relevanz (Koutsabasis et al., 2007). So konnten zwar die Kosten zum Auffinden eines einzelnen Problems beziffert werden, jedoch sind die Kosten und Nutzen der Problembehebung, nicht bei jedem Problem gleich.

Es ist weiter unklar wie die Anzahl von zwei Evaluator*innen die Kosten beim UTA und beim UTM beeinflusste. Es ist weiter offen, wie Kosten ausgesehen hätten, wenn nur mit fünf Testpersonen getestet worden wäre (Magic Number 5) bzw. ob diese einfach linear sind. Es ist denkbar, dass die Vorbereitungskosten ähnlich sind und nicht stark von der Anzahl Testpersonen abhängen. Dies ist jedoch eine Vermutung, die es zu prüfen gilt.

Für die HE wurden die Heuristiken nach Nielsen (2020) verwendet. Hier stellt sich die Frage, ob andere Resultate entstanden wären, wenn andere Heuristiken, z. B. nach Shneiderman (Bader et al., 2017), zum Einsatz gekommen wären.

Ebenfalls ist hier die bereits erwähnte Kritik von Lewis (1994, zitiert nach Barnum, 2003) zu erwähnen, dass durch die Konstruktion und die Anzahl der Aufgaben die Wahrscheinlichkeit beeinflusst wird, dass Usability-Probleme gefunden werden. Eine andere Formulierung und eine

abweichende Anzahl von Aufgaben hätten möglicherweise einen Einfluss auf die Anzahl gefundener Probleme gehabt.

Obwohl ein digitales Tool zur Erfassung der Zeit zum Einsatz kam, steht offen, wie genau diese Messungen sind. Trotz Sorgfalt ist es nicht auszuschliessen, dass diese nicht immer exakt waren oder dass Einträge vergessen wurden. Weiter rapportierten die Expert*innen ihre Aufwände selbstständig und dies konnte nicht kontrolliert werden. Hingegen kann angemerkt werden, dass sich die Zeitaufwände für den UTA, den UTM und die HE mit den Erfahrungswerten der Praxispartnerin sowie mit Daten aus der CUE-4 Studie von Molich und Dumas (2008) decken.

Bei sechs Testpersonen aus der Gruppe UTA ($n = 6$) fiel auf, dass die Antworten im PSSUQ sehr negativ ausfielen und nicht mit dem beobachtbaren Verhalten im Video übereinstimmten. Der offensichtliche Grund dafür war, dass der PPSUQ eine ungewohnte Reihenfolge der Likert-Skala verwendet (1 = *trifft völlig zu* bzw. 7 = *trifft gar nicht zu*). Diese Datensätze wurden dann so ausgewertet, als wäre die Skala in die andere Richtung verlaufen. Für zuenkünftige Erhebung wäre es sinnvoll, die Skalen bereits im Voraus umzudrehen, was nach Albert und Tullis (2022) sowie Sauro (2019) ein zulässiges Vorgehen wäre. Dieses Phänomen konnte auch in der Gruppe UTM beobachtet werden, führte dort aber nicht zu Problemen, da der Moderator Hinweise geben konnte, falls Missverständnisse auftraten.

Die Reihenfolge der Messinstrumente PANAS (Affekt), PSSUQ (subjektive Benutzerfreundlichkeit) und NASA-TLX (mentale Belastung) wurden nicht randomisiert und es ist daher nicht auszuschliessen, ob z. B. das Ausfüllen des PANAS und des PSSUQ nacheinander einen Einfluss auf die Antworten im NASA-TLX gehabt haben.

Weiter kann nicht ausgeschlossen werden, dass die leichten Anpassungen am PSSUQ und am NASA-TLX einen Einfluss hatten. Insbesondere sei hier auf die technischen Limitierungen des Fragebogen-Tools hingewiesen, wonach die Fragen des NASA-TLX mit einer elfstufigen Likert-Skala abgebildet werden mussten.

Der Einfluss der Persönlichkeit auf die subjektive Bewertung der Usability (Kortum & Oswald, 2018) konnte im Rahmen dieser Arbeit nicht kontrolliert werden. Es ist somit nicht auszuschliessen, dass die Persönlichkeitsmerkmale *Offenheit für Erfahrungen* und *Verträglichkeit* eine Auswirkung auf die Resultate im PSSUQ hatten.

Die Problematik der tiefen Reliabilität der Expert*innen hätte minimiert werden können, wenn gemäss dem typischen Ablauf einer HE die Expert*innen zusammengekommen wären, um einen Konsens zu bilden und die gefundenen Probleme zu konsolidieren. Dies wurde in der vorliegenden Studie aufgrund der zeitlichen Verfügbarkeit der Expert*innen nicht gemacht. Es ist aber zu vermuten, dass dieser Schritt zwar die Anzahl der gefundenen Probleme durch die HE reduziert und somit die Validity bzw. die Actual Effectiveness verbessert hätte, jedoch dadurch nicht mehr Hits aufgetreten wären. Lindgaard (2006) kritisiert die Messung der Actual Effectiveness und Actual Efficiency, da es nicht möglich sei, wirklich alle Probleme in einem interaktiven System zu finden. Lindgaard begründet dies wie folgt: «In the absence of a complete usability problem set, the notions of thoroughness and efficiency are meaningless and also impossible to calculate» (S.1073). Demnach kann zwar die Validität einer Methode beurteilt werden, jedoch nicht ihre Gründlichkeit. Die Bekanntheit aller Probleme würde bedingen, dass eine Evaluation so lange wiederholt wird, bis keine neuen Probleme mehr gefunden werden. Ein solches Vorgehen ist in der Praxis aufgrund der begrenzten Ressourcen jedoch nicht anwendbar. Weiter ist anzumerken, dass bei der Kennzahl Actual Efficiency zusätzliche Kosten, z. B. für die Rekrutierung, nicht miteinberechnet wurden. Hier wäre eine Erweiterung des Konstrukts sinnvoll, bei der nicht nur die aufgewendeten Stunden mitberücksichtigt werden.

5.3 Handlungsempfehlungen

Auf Basis der Erkenntnisse dieser Arbeit werden nachfolgend vier praktische Handlungsempfehlungen abgeleitet, die UX-Professionals im Evaluationsprozess unterstützen sollen.

5.3.1 Empfehlung 1: Usability-Tests bevorzugen

Moderierte Usability-Tests sind wirkungsvoll, um Probleme in interaktiven Systemen aufzudecken. Als kostengünstigere Alternative können automoderierte Usability-Tests in Betracht gezogen werden. Tests mit wenigen Testpersonen können bereits mehr echte Probleme aufdecken als einzelne Expert*innen. Insgesamt weisen Usability-Tests ein besseres Kosten-Nutzen-Verhältnis auf als heuristische Evaluationen. Bei begrenzten finanziellen Ressourcen kann der Einsatz einer heuristischen Evaluation in Erwägung gezogen werden, jedoch sollte dabei auch das Risiko von falschen Alarmen berücksichtigt werden.

5.3.2 Empfehlung 2: Vorsicht bei der Anwendung der Magic Number 5

Zur Bestimmung der Anzahl Testpersonen für ein Usability-Testing ist es ratsam, nicht einfach die Regel Magic Number 5 anzuwenden. Stattdessen sollte gefragt werden, ob eher die offensichtlicheren oder die versteckteren Probleme gefunden werden sollen und wie hoch das Mass an Sicherheit sein soll, das bei den Ergebnissen erwartet wird. Basierend auf diesen Informationen sollte dann eine geeignete Stichprobe berechnet werden.

5.3.3 Empfehlung 3: Mehraugenprinzip anwenden

Um sicherzustellen, dass Usability-Tests aussagekräftig sind und eine optimale Kosten-Nutzen-Bilanz bieten, ist es empfehlenswert, Auswertungen jeweils von mehreren Personen durchführen zu lassen. Eine schnelle und oberflächliche Auswertung kann das Risiko erhöhen, wichtige Fehler zu übersehen und somit die Aussagekraft der Tests zu beeinträchtigen. Auch bei einer heuristischen Evaluation ist es angezeigt, die Durchführung durch mehrere Personen machen zu lassen.

5.3.4 Empfehlung 4: Fokus auf schwerwiegende Probleme bei der heuristischen Evaluation

Wenn eine heuristische Evaluation durchgeführt wird, sollten bei der anschliessenden Bearbeitung nur die schwerwiegendsten Probleme im Fokus stehen. Dadurch erhöht sich die Wahrscheinlichkeit, dass die identifizierten Probleme auch tatsächlich Schwierigkeiten darstellen, mit denen Nutzer*innen auch konfrontiert sein werden.

6 FAZIT UND AUSBLICK

Die vorliegende Studie untersuchte die inhaltlichen und finanziellen Unterschiede der drei Usability-Evaluationsmethoden UTM, UTA und HE. Dank der Ergebnisse konnten die beiden Fragestellungen beantwortet werden. Mithilfe weiterführender Analysen konnten ergänzende Erkenntnisse herausgearbeitet werden. Die Ergebnisse weisen darauf hin, dass die Methode UTM am meisten Probleme aufdeckt, während die HE Probleme höherem Schweregrads aufzeigt. Die Kosten pro identifiziertem Problem sind beim UTA am geringsten.

Mit Blick auf die eingangs beschriebenen Zielsetzungen dieser Arbeit, konnten für die sinnhaft GmbH in Bezug auf die Kosten und Nutzen von UEM wichtige Erkenntnisse gewonnen werden, die z. B. im Angebotsprozess genutzt werden können. Dank der umfangreichen Ergebnisse aus den drei durchgeführten UEM wurde der Yousty AG eine Entscheidungsgrundlage für die Weiterentwicklung von professional.ch geliefert, welche Teile und Funktionen überarbeitet sowie weiterentwickelt werden sollen.

Im Folgenden werden Vorschläge für weitere Forschungsvorhaben präsentiert, die auf dieser Studie aufbauen können.

Die Hypothese H1b, wonach mittels Usability-Tests schwerwiegendere Usability-Probleme gefunden werden, konnte nicht verifiziert werden. Es wurde das Gegenteil beobachtet, nämlich dass durch die HE schwerwiegendere Probleme gefunden wurden. Es wäre wertvoll, weitere Forschungen durchzuführen, um herauszufinden, durch welche Evaluationsmethoden unter welchen Bedingungen welche Schweregrade gefunden werden. Wie in der Kritik erwähnt wurde, steht es offen, inwiefern sich die Resultate generalisieren lassen. So wäre eine Replikation der Studie interessant, um ein besseres Verständnis dafür zu erhalten, welche UEM und Einflussfaktoren wie auf die Resultate wirken (z. B. die Anzahl der Usability-Probleme oder der Arbeitsaufwand). An dieser Stelle wäre es interessant, eine eingehendere Auseinandersetzung darüber zu führen, wie sich neuartige Ansätze wie das automoderierte Usability-Testing oder andere zukünftige Anwendungen mit künstlicher Intelligenz auf Testpersonen bzw. die Ergebnisse auswirken.

Eine weitere interessante Forschungsfrage wäre, wie insbesondere die Verzerrung der HE-Ergebnisse minimiert werden könnte. Auch die CUE-8 Studie von Molich et al. (2010) zeigen, dass es eine grosse Variabilität bei der Auswertung von Usability-Tests gibt. So werden unterschiedliche Evaluator*innen auch unterschiedliche Probleme identifizieren, auch wenn diese das gleiche Ausgangsmaterial sehen. Es wäre nützlich, weiter zu untersuchen, wie diese Unterschiede verringert werden können.

Die Entscheidung, wie gross eine Stichprobe ist, wird immer getroffen. Es ist nur die Frage, wer dies mit welchem Wissen tut (vgl. Lakens, 2022; Mutschler & Reichert, 2004). Oftmals bestimmen hier wirtschaftliche und zeitliche Aspekte. So führt eine zu kleine Stichprobe zu wenig akkuraten Ergebnissen und erhöht das Risiko, falsche Entscheidungen zu treffen. Eine zu grosse Stichprobe erhöht die Kosten ohne zusätzliche Aussagekraft. Aus wirtschaftlicher Sicht wäre eine weitere Forschungsfrage, ob und an welchem Punkt das Hinzunehmen von weiteren Testpersonen einen Mehrwert bietet. Hinweise dazu liefert das neue und wenig erforschte Modell *Cube Root Problem Discovery Model* (Sauro & Lewis, 2022). Im Modell wird postuliert, dass mittels der Ergebnisse der ersten Nutzer*innen berechnet werden kann, wie viele Usability-Probleme mit einer grösseren Stichprobe auftreten sollten. So wäre es denkbar, dass Usability-Tests mit kleineren Stichproben durchgeführt werden und dann jeweils kalkuliert wird, ob die Hinzunahme von weiteren Personen lohnend ist. Dahingehend wertvoll wäre ein verlässliches Instrument zum Abschätzen der p-Werte der Usability-Probleme. Dieses würde die Berechnung einer jeweils passenden Stichprobengrösse ermöglichen. Sauro und Lewis (2016) führen zwar solche Werte auf, weisen aber explizit darauf hin, dass es sich hierbei nur um Erfahrungswerte handelt.

Unternehmen müssen Entscheidungen trotz Unsicherheit treffen (vgl. Wilkinson & Klaes, 2022), und verpasste Probleme sowie falsche Alarme können dazu führen, dass Ressourcen unangemessen eingesetzt werden und daraus kein Nutzen entsteht. Es muss auch in Frage gestellt werden, ob die Kosten der einzige Einflussfaktor sind sowie ob und wie UEM in einem Entwicklungsprozess eingesetzt werden. Dieser traditionelle Ansatz, bei dem die Kosten begründet und die Gewinne hervorgehoben werden, steht im Gegensatz zu einem durch Aly und Sturm (2019)

neuartigen vorgeschlagenen Ansatz eines *Fear-Setting*, bei dem Verluste und Kosten durch eine Untätigkeit unterstrichen werden. Es wird angenommen, dass dieser Ansatz tendenziell ältere und erfahrenere Entscheidungsträger*innen eher dazu bewegt, in die Usability-Aktivitäten zu investieren und dadurch mögliche Verluste sowie verpasste Gelegenheiten zu vermeiden. Dies deckt sich mit Bias und Mayhew (2005), die betonen, dass die Glaubwürdigkeit von Usability-Aktivitäten vor allem von einem *wahrgenommenen* Return on Investment abhängt. Dieses Gebiet wäre somit ein wertvolles Feld für weitere Forschungen.

Wenn die Kosten infrage gestellt werden, kann dies auch ein Ausdruck des Misstrauens gegenüber dem Mehrwert solcher Methoden sein. Es ist somit eine Aufgabe von Usability-Professionals, Unternehmen dahingehend zu beraten und zu begleiten.

Insgesamt ist jede Bemühung zur Verbesserung des Nutzungserlebnisses wertvoll. Es ist besser, irgendetwas zu tun, als gar nichts zu unternehmen. Um mit den Worten von Nielsen (1997) einen passenden Abschluss zu finden: «Just do it. The true choice is not between discount and deluxe usability engineering. If that were the choice, I would agree that the deluxe approach would bring better results. The true choice, however, is between doing something and doing nothing. Perfection is not an option. My choice is to do something!».

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

FORMELVERZEICHNIS

LITERATURVERZEICHNIS

Albert, B. & Tullis, T. (2022). *Measuring the user experience* (3rd ed.). Cambridge, MA: Morgan Kaufmann. https://doi.org/10.1016/C2018-0-00693-3

Aly, M. & Sturm, C. (2019). Hacks for cost-justifying usability: „Fear-setting“ vs. „goal-setting“. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (article no. 77, pp. 1–10). New York, NY: ACM. https://doi.org/10.1145/3338286.3347544

Bader, F., Schön, E. M. & Thomaschewski, J. (2017). Heuristics considering UX and quality criteria for heuristics. *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*, 48. https://doi.org/10.9781/ijimai.2017.05.001

Barnum, C. (2003). The ‘magic number 5’: Is it enough for web-testing? *Information Design Journal*, *11*, 160–170. https://doi.org/10.1075/idj.11.2.08bar

Barnum, C. M. (2020). *Usability testing essentials* (2nd ed.). Cambridge, MA: Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-816942-1.00001-0

Baumgartner, J., Sonderegger, A. & Sauer, J. (2017). *Stop reading, start looking: A pictorial workload scale for the evaluation of interactive products*. Poster presented at the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference, Rome. https://doi.org/10.13140/RG.2.2.13273.57440

Bias, R. G. & Mayhew, D. J. (Eds.). (2005). *Cost-justifying usability: An update for an internet age* (2nd ed.). San Francisco, CA: Morgan Kaufmann.

Birns, J., Joffre, K., Leclerc, J. & Paulsen, C. (2002). Getting the whole picture: Collecting usability data using two methods: Concurrent think aloud and retrospective probing. Retrieved from https://www.researchgate.net/publication/228865154_Getting_the_Whole_Picture_Collecting_Usability_Data_Using_Two_Methods--Concurrent_Think_Aloud_and_Retrospective_Probing

Borsci, S., Macredie, R. D., Barnett, J., Martin, J., Kuljis, J. & Young, T. (2013). Reviewing and extending the five-user assumption: A grounded procedure for interaction evaluation. *ACM Transactions on Computer-Human Interaction*, *20*, article no. 29, 1–23. https://doi.org/10.1145/2506210

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-12770-0

Breyer, B. & Bluemke, M. (2016). *Deutsche Version der Positive and Negative Affect Schedule PANAS (GESIS Panel)*. Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS) der GESIS Leibniz Institute for the Social Sciences. https://doi.org/10.6102/ZIS242

Budiu, R. (2017, October 22). You are not the user: The false-consensus effect. Nielsen Norman Group. Verfügbar unter: https://www.nngroup.com/articles/false-consensus/

Chynał, P. & Sobecki, J. (2015). Statistical verification of remote usability testing method. In *Proceedings of the Mulitimedia, Interaction, Design and Innnovation on ZZZ - MIDI ’15* (article no. 12, pp. 1–7). New York, NY: ACM. https://doi.org/10.1145/2814464.2814476

Cohen, J. (1992). Statistical Power Analysis. Current Directions in Psychological Science, 1(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Crawford, J. R. & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, *43*, 245–265. https://doi.org/10.1348/0144665031752934

Creswell, J. W. & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Los Angeles: SAGE.

Deutsches Institut für Normung (DIN). (2020a). *DIN EN ISO 9241-210:2020-03, Ergonomie der Mensch-System-Interaktion_- Teil_210: Menschzentrierte Gestaltung interaktiver Systeme (ISO_9241-210:2019).* Berlin: Beuth. https://doi.org/10.31030/3104744

Deutsches Institut für Normung (DIN). (2020b). *DIN EN ISO 9241-110:2020-10, Ergonomie der Mensch-System-Interaktion_- Teil_110: Interaktionsprinzipien (ISO_9241-110:2020)*. Berlin: Beuth. https://doi.org/10.31030/3147467

Dunn, A. M., Heggestad, E. D., Shanock, L. R. & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology*, *33*, 105–121. https://doi.org/10.1007/s10869-016-9479-0

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*, 1– 26. https://doi.org/10.1214/aos/1176344552

Efron, B. (2000). The bootstrap and modern statistics. *Journal of the American Statistical Association*, *95*, 1293–1296. https://doi.org/10.1080/01621459.2000.10474333

Egloff, B., Tausch, A., Kohlmann, C.-W. & Krohne, H. W. (1995). Relationships between time of day, day of the week, and positive mood: Exploring the role of the mood measure. *Motivation and Emotion*, *19*, 99–110. https://doi.org/10.1007/BF02250565

Fernandez, A., Insfran, E. & Abrahão, S. (2011). Usability evaluation methods for the web: A systematic mapping study. *Information and Software Technology*, *53*, 789–817. https://doi.org/10.1016/j.infsof.2011.02.007

Flägel, K., Galler, B., Steinhäuser, J. & Götz, K. (2019). Der „National Aeronautics and Space Administration-Task Load Index“ (NASA-TLX): Ein Instrument zur Erfassung der Arbeitsbelastung in der hausärztlichen Sprechstunde: Bestimmung der psychometrischen Eigenschaften. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, *147–148*, 90–96. https://doi.org/10.1016/j.zefq.2019.10.003

Gebauer, H., Krempl, R. & Fleisch, E. (2008). Erfolgsfaktoren der Dienstleistungsentwicklung in technologieorientierten Unternehmen. In C. Marxt & F. Hacklin (Hrsg.), *Business Excellence in technologieorientierten Unternehmen* (S. 119–129). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-73881-7_10

Georgsson, M. (2019). NASA RTLX as a novel assessment for determining cognitive load and user acceptance of expert and user-based evaluation methods exemplified through a mHealth diabetes self-management application evaluation. *Studies in Health Technology and Informatics*, *261*, 185–190. Retrieved from https://ebooks.iospress.nl/publication/51533

Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: John Wiley & Sons.

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*, 904–908. https://doi.org/10.1177/154193120605000909

Hartson, H. R. & Pyla, P. S. (2012). *The UX Book: Process and guidelines for ensuring a quality user experience*. Waltham, MA: Morgan Kaufmann.

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*, 77–89. https://doi.org/10.1080/19312450709336664

Hertzum, M., Borlund, P. & Kristoffersen, K. B. (2015). What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions. *International Journal of Human-Computer Interaction*, *31*, 557–570. https://doi.org/10.1080/10447318.2015.1065691

Hertzum, M., Molich, R. & Jacobsen, N. E. (2014). What you get is what you see: Revisiting the evaluator effect in usability tests. *Behaviour & Information Technology*, *33*, 144–162. https://doi.org/10.1080/0144929X.2013.783114

Kaminski, C. P. (2018). *Gebrauchstauglichkeitsanalyse zur Qualitätssicherung im medizinischen Kontext* (Unveröffentlichte Dissertation). Eberhard Karls Universität Tübingen.

Kortum, P. & Oswald, F. L. (2018). The impact of personality on the subjective assessment of usability. *International Journal of Human-Computer Interaction*, *34*, 177–186. https://doi.org/10.1080/10447318.2017.1336317

Koutsabasis, P., Spyrou, T. & Darzentas, J. (2007). Evaluating usability evaluation methods: Criteria, method and a case study. In Julie A. Jacko (Ed.), *Human-computer interaction: Interaction design and usability* (pp. 569–578). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-73105-4_63

Krohne, H. W., Egloff, B., Kohlmann, C.-W. & Tausch, A. (1996). Untersuchungen mit einer deutschen Version der „Positive and Negative Affect Schedule“ (PANAS). *Diagnostica*, *42*, 139–156. https://doi.org/10.1037/t49650-000

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*, article 33267. https://doi.org/10.1525/collabra.33267

Lavery, D., Cockton, G. & Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology*, *16*, 246–266. https://doi.org/10.1080/014492997119824

Law, E. L.-C. & Hvannberg, E. T. (2004). Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. In *Proceedings of the third Nordic conference on Human-computer interaction – NordiCHI ’04* (pp. 241–250). New York, NY: ACM. https://doi.org/10.1145/1028014.1028051

Lindgaard, G. (2006). Notions of thoroughness, efficiency, and validity: Are they valid in HCI practice? *International Journal of Industrial Ergonomics*, *36*, 1069–1074. https://doi.org/10.1016/j.ergon.2006.09.007

Madan, A. & Kumar, S. (2012). Usability evaluation methods: A literature review. *International Journal of Engineering Science and Technology*, *4*, 590–599. Retrieved from https://www.researchgate.net/publication/266874640_Usability_evaluation_methods_a_literature_review

Mator, J. D., Lehman, W. E., McManus, W., Powers, S. A., Tiller, L. N., Unverricht, J. et al. (2021). Usability: Adoption, measurement, value. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *63*, 956–973. https://doi.org/10.1177/0018720819895098

Molich, R., Chattratichart, J., Hinkle, V., Jensen, J. J., Kirakowski, J., Sauro, J. et al. (2010). Rent a car in just 0, 60, 240 or 1,217 seconds? Comparative usability measurement, CUE-8. *Journal of Usability Studies*, *6*, 8–24. Retrieved from https://uxpajournal.org/wp-content/uploads/sites/7/pdf/JUS_Molich_November_2010.pdf

Molich, R. & Dumas, J. S. (2008). Comparative usability evaluation (CUE-4). *Behaviour & Information Technology*, *27*(, 263–281. https://doi.org/10.1080/01449290600959062

Moser, C. (2012). *User Experience Design: Mit erlebniszentrierter Softwareentwicklung zu Produkten, die begeistern*. Berlin, Heidelberg: Springer.

Mutschler, B. & Reichert, M. (2004). Usability-Metriken als Nachweis der Wirtschaftlichkeit von Verbesserungen der Mensch-Maschine-Schnittstelle. In *Proceedings of the IWSM / MetriKon Workshop on Software Metrics (IWSM / MetriKon'04)* (S. 407–418). Königs Wusterhausen: Shaker-Verlag. Verfügbar unter http://dbis.eprints.uni-ulm.de/160/

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems – CHI '92* (pp. 373–380). New York, NY: ACM. https://doi.org/10.1145/142750.142834

Nielsen, J. (1994a, November 1). How to conduct a heuristic evaluation. Nielsen Norman Group. Retrieved from https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/

Nielsen, J. (1994b, November 1). Severity ratings for usability problems. Nielsen Norman Group. Retrieved from https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/

Nielsen, J. (1997, January 1). Discount usability for the web. Nielsen Norman Group. Retrieved form https://www.nngroup.com/articles/web-discount-usability/

Nielsen, J. (2020, November 15). 10 usability heuristics for user interface design. Nielsen Norman Group. Retrieved from https://www.nngroup.com/articles/ten-usability-heuristics/

Nielsen, J. & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the SIGCHI conference on Human factors in computing systems – CHI '93* (pp. 206–213). New York, NY: ACM. https://doi.org/10.1145/169059.169166

Nielsen, J. & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems Empowering people – CHI '90* (pp. 249–256). New York, NY: ACM. https://doi.org/10.1145/97243.97281

professional.ch. (2023). Unsere Mission. Verfügbar unter https://www.professional.ch/unsere-mission

Rajanen, M. (2007). Usability cost-benefit models: Different approaches to usability cost analysis. In *Proceedings of the 9th International Conference on Enterprise Information Systems (ICEIS 2007)* (pp. 332–336). New York, NY: ACM. https://doi.org/10.13140/2.1.1718.6086

Rajanen, M. (2011). *Applying usability cost-benefit analysis: Explorations in commercial and open source software development contexts* (Dissertation). University of Oulu.

Rajanen, M. (2020). Usability cost-benefit analysis for information technology applications and decision making: In E. C. Idemudia (Ed.), *Advances in Business Strategy and Competitive Advantage* (pp. 136–152). Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-7998-3351-2.ch008

Rajanen, M. & Iivari, N. (2007). Usability cost-benefit analysis: How usability became a curse word? (Lecture notes in computer science). In C. Baranauskas, P. Palanque, J. Abascal & S. D. J. Barbosa (Eds.), *Human-Computer Interaction – INTERACT 2007* (pp. 511–524). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-74800-7_47

Raskin, J. (2000). *The humane interface: New directions for designing interactive systems*. New York, NY: ACM Press/Addison-Wesley Publishing Co.

Richter, M. & Flückiger, M. (2016). *Usability und UX kompakt: Produkte für Menschen* (4. Aufl.). Berlin, Heidelberg: Springer Vieweg. https://doi.org/10.1007/978-3-662-49828-6

Ross, L., Greene, D. & House, P. (1977). The “false consensus effect”: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*, 279–301. https://doi.org/10.1016/0022-1031(77)90049-X

Samsung. (2022, November 13). Galaxy phone stops recording videos after 10 minutes. Samsung Support. Retrieved from https://www.samsung.com/us/support/troubleshooting/TSG01001538/

Sarodnick, F. & Brau, H. (2011). *Methoden der Usability Evaluation: Wissenschaftliche Grundlagen und praktische Anwendung* (2. Aufl.). Bern: Verlag Hans Huber.

Sauer, J., Sonderegger, A., Heyden, K., Biller, J., Klotz, J. & Uebelbacher, A. (2019). Extra-laboratorial usability tests: An empirical comparison of remote and classical field testing with lab testing. *Applied Ergonomics*, *74*, 85–96. https://doi.org/10.1016/j.apergo.2018.08.011

Sauer, J., Sonderegger, A. & Schmutz, S. (2020). Usability, user experience and accessibility: towards an integrative model. *Ergonomics*, *63*, 1207–1220. https://doi.org/10.1080/00140139.2020.1774080

Sauro, J. (2012, September 6). How effective are heuristic evaluations? MeasuringU. Retrieved from https://measuringu.com/effective-he/

Sauro, J. (2013, July 30). Rating the severity of usability problems. MeasuringU. Retrieved from https://measuringu.com/rating-severity/

Sauro, J. (2014). The relationship between problem frequency and problem severity in usability evaluations. *Journal of Usability Studies*, *10*, 17–25. Retrieved from http://uxpajournal.org/the-relationship-between-problem-frequency-and-problem-severity-in-usability-evaluations

Sauro, J. (2016, February 2). Managing false positives in UX research. MeasuringU. Retrieved from https://measuringu.com/false-positives/

Sauro, J. (2019, December 18). 10 things to know about the post study system usability questionnaire. MeasuringU. Retrieved from https://measuringu.com/pssuq/

Sauro, J. & Lewis, J. (2022, April 5). A new statistical approach for predicting usability problems. MeasuringU. Retrieved from https://measuringu.com/cube-root-problem-discovery-model/

Sauro, J. & Lewis, J. R. (2005). Estimating completion rates from small samples using binomial confidence intervals: Comparisons and recommendations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*, 2100–2103. https://doi.org/10.1177/154193120504902407

Sauro, J. & Lewis, J. R. (2016). *Quantifying the user experience: practical statistics for user research* (2nd ed.). Cambridge, MA: Morgan Kaufmann.

Schrepp, M., Hinderks, A. & Thomaschewski, J. (2017). Die UX KPI-Wunsch und Wirklichkeit: Kann man User Experience in einer einzigen Kennzahl zusammenfassen? In S. Hess & H. Fischer (Hrsg.), *Mensch und Computer 2017: Usability Professionals* (S. 117–125). Regensburg: Gesellschaft für Informatik e.V. https://doi.org/10.18420/MUC2017-UP-0100

Skjuve, M., Haugstveit, I. M., Følstad, A. & Brandtzaeg, P. B. (2019). Help! Is my chatbot falling into the uncanny valley? An empirical study of user experience in human-chatbot interaction. *Human Technology*, *15*, 30–54. https://doi.org/10.17011/ht/urn.201902201607

Tan, W., Liu, D. & Bishu, R. (2009). Web evaluation: Heuristic evaluation vs. user testing. *International Journal of Industrial Ergonomics*, *39*, 621–627. https://doi.org/10.1016/j.ergon.2008.02.012

Thyvalikakath, T. P., Monaco, V., Thambuganipalle, H. & Schleyer, T. (2009). Comparative study of heuristic evaluation and usability testing methods. *Studies in health technology and informatics*, *143*, 322–327. http://dx.doi.org/10.3233/978-1-58603-979-0-322

Tomczak, T., Walter, B. & Henkel, S. (2011). Strategisches Employer Branding. *GfM-Forschungsreihe*, *6*, 1–12. Verfügbar unter https://www.steauf.de/wp-content/uploads/2015/05/Embloyee_Branding_06_2011-1.pdf

Travis, D. (2009, October 5). How to prioritise usability problems. User Focus. Retrieved from https://www.userfocus.co.uk/articles/prioritise.html

Travis, D. & Hodgson, P. (2019). *Think like a UX researcher: How to observe users, influence design, and shape business strategy*. Boca Raton, FL: CRC Press.

Tullis, T., Fleischman, S., McNulty, M., Cianchette, C. & Bergel, M. (2002). An empirical comparison of lab and remote usability testing of web sites. In *Usability Professionals' Association 2002 Conference Proceedings* (CD-ROM). Bloomingdale, IL: Usability Professionals' Association.

UXQB e.V. (Hrsg.). (2020, 1. November). CPUX-UT Curriculum: Certified Professional for Usability and User Experience – Usability Testing and Evaluation (Version 1.18 DE). Verfügbar unter: https://uxqb.org/public/documents/CPUX-UT_DE_Curriculum.pdf

VandenBos, G. R. (Ed.). (2016). *APA College Dictionary of Psychology* (2nd ed.). Washington, DC: American Psychological Association.

Vermeeren, A. P. O. S., Law, E. L.-C., Roto, V., Obrist, M., Hoonhout, J. & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction Extending Boundaries - NordiCHI '10* (pp. 521–530). New York, NY: ACM. https://doi.org/10.1145/1868914.1868973

Watson, D., Clark, L. A. & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Wilkinson, N. & Klaes, M. (2022). *An introduction to behavioral economics* (3rd ed.). London, New York: Bloomsbury Academic.

Wirtz, M. A. (Hrsg.). (2021). *Dorsch: Lexikon der Psychologie* (20. Aufl.). Bern: Hogrefe.