



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Universität Hamburg

Fachbereich Informatik

Masterarbeit

Usability-Testmethoden für mobile Anwendungen im Vergleich

vorgelegt von

Peter Schmidt

geb. am 27. Dezember 1984 in Lauterbach

Matrikelnummer: 6124325

eingereicht am 22. Dezember 2011

Erstgutachter: Prof. Dr. Horst Oberquelle

Zweitgutachter: Dr. Guido Gryczan

Zusammenfassung

In dieser Masterarbeit wird ein Verfahren entwickelt, um Methoden zum Testen der Usability auf mobilen Anwendungen vergleichbar zu machen. Mit Hilfe dieses Verfahrens erfolgt anschließend eine Bewertung, um eine Auswahl an Usability-Testmethoden zu bewerten. Zu diesem Zweck werden die dafür relevanten Grundlagen des Usability-Engineerings erläutert und die Besonderheiten von mobilen Anwendungen durchleuchtet. Die Erarbeitung von Kriterien zur Bewertung der Usability-Testmethoden ergibt sich nachfolgend aus den zuvor genannten Grundlagen und Besonderheiten. Zusätzlich wird eine Möglichkeit beschrieben, die es erlaubt, diesen Kriterien Prioritäten zuzuordnen und daraus eine Gesamtbewertung zu errechnen. Im weiteren Verlauf werden sechs Usability-Testmethoden beschrieben und mit Hilfe der zuvor ermittelten Kriterien bewertet. Diese Methoden finden bereits Verwendung in der Analyse der Usability von Applikationen für immobile Computer. Autoren verschiedener wissenschaftlicher Arbeiten optimierten sie zum Testen von mobilen Anwendungen. Diese Studien bilden die Grundlage für die Bewertung der Methoden mit Hilfe der erarbeiteten Kriterien.

Abstract

This Master thesis develops an approach to make usability-testing methods for mobile applications comparable. The approach is subsequently used to evaluate a selection of those methods. For this purpose the relevant principles of usability-engineering and the characteristics of mobile applications and mobile use are specified. This information is then used to evolve criteria that allow the evaluation of mobile applications. In addition a technique to prioritize certain criteria and calculate an overall ranking is shown. Afterwards six usability-testing methods are explained and evaluated with the determined criteria. These methods were originally used to analyze non-mobile applications and were optimized by researchers to match the special characteristics of usability-testing for mobile applications. Their scientific studies are used to evaluate the selected usability-testing methods in this master thesis.

Inhaltsverzeichnis

Abbildungsverzeichnis	VI
Tabellenverzeichnis	VII
Abkürzungsverzeichnis	VIII
1 Einführung	1
1.1 Motivation	1
1.2 Zielsetzung	3
1.3 Aufbau	3
2 Grundlagen	4
2.1 Usability	4
2.1.1 Definition	4
2.1.2 Usability-Attribute	5
2.1.3 Usability-Engineering	6
2.1.4 Usability-Testing	11
2.1.5 Usability-Evaluation und Metriken	13
2.1.6 Normen und Standards	19
2.2 Mobile Anwendungen	20
2.2.1 Begriffserklärung: mobile Anwendungen	20
2.2.2 Usability-Testing für mobile Anwendungen	21
3 Methodik zur Evaluation der Usability-Testmethoden	33
3.1 Bewertungskriterien zur Evaluation der Usability-Testmethoden	33
3.1.1 Geräteunabhängige Kriterien	33
3.1.2 Geräteabhängige Kriterien	39
3.2 Verfahren zur Gesamtbewertung einer Usability-Testmethode	42
3.2.1 Absolute Gewichtung	42
3.2.2 Anwendungsbeispiel für die absolute Gewichtung	43
4 Usability-Testmethoden	45
4.1 Think Aloud Test (im Labor)	45
4.1.1 Beschreibung	45
4.1.2 Bewertung	46
4.2 Think Aloud Test(im Feld)	53
4.2.1 Beschreibung	54
4.2.2 Bewertung	55
4.3 Kognitiver Walkthrough	60
4.3.1 Beschreibung	62

Inhaltsverzeichnis

4.3.2	Bewertung	63
4.4	Heuristische Evaluation	70
4.4.1	Beschreibung	71
4.4.2	Bewertung	73
4.5	Interview	78
4.5.1	Beschreibung	79
4.5.2	Bewertung	80
4.6	Fragebogen	86
4.6.1	Beschreibung	86
4.6.2	Bewertung	87
4.7	Zusammenfassung	93
5	Fazit und Ausblick	95
5.1	Fazit	95
5.2	Ausblick	98
	Literaturverzeichnis	IX

Abbildungsverzeichnis

1.1	Relevanz der Usability für die Auswahl mobiler Endgeräte	2
2.1	Ebenen der Usability-Optimierung	7
2.2	Anwendungsbereich des Usability-Engineerings	8
2.3	Verhältnis zwischen gefundenen Problemen und eingesetzten Testnutzern . . .	12
2.4	Einordnung von Usabilityproblemen	16
2.5	Beispiel für die Veranschaulichung der Nutzerkommentare nach Anteilen . . .	18
2.6	Normen und Standards des Usability-Engineerings	19
2.7	Verteilung der Betriebssysteme auf Smartphones	22
2.8	Der Kontext von mobilen Anwendungen im Detail	27
2.9	Nutzungsorte des mobilen Internets	29
2.10	Video Kamera mit Tilt Pad	30
2.11	Video Kamera direkt auf dem mobilen Endgerät befestigt	30
2.12	Usability-Test mit vier Miniaturkameras	31
4.1	Beispiel eines Mock-Ups für mobile Anwendungen	48
4.2	Übersicht über die Bewertung des Think Aloud Tests (im Labor)	53
4.3	Ausrüstung des Think Aloud Test in natürlicher Nutzungsumgebung	55
4.4	Übersicht über die Bewertung des Think Aloud Tests (im Feld)	60
4.5	Übersicht über die Bewertung des Kognitiven Walkthroughs	70
4.6	Übersicht über die Bewertung der Heuristischen Evaluation	78
4.7	Übersicht über die Bewertung des Interviews	85
4.8	Fragebögen auf mobilen Endgeräten	87
4.9	Übersicht über die Bewertung des Fragebogens	92
4.10	Übersicht über die Bewertung aller Usability-Testmethoden	94

Tabellenverzeichnis

3.1	Bewertungstabelle für den Zeitaufwand	34
3.2	Bewertungstabelle für den Kostenaufwand	36
3.3	Bewertungstabelle für den Zeitpunkt der Durchführung	37
3.4	Bewertungstabelle über die Berücksichtigung der Usability-Attribute	37
3.5	Bewertungstabelle über die Berücksichtigung des emotionalen Zustands	39
3.6	Bewertungstabelle über die Berücksichtigung des physischen Zustands	39
3.7	Bewertungstabelle über die Berücksichtigung der natürlichen Nutzungsumgebung	40
3.8	Bewertungstabelle über die Berücksichtigung der mobilgerätespezifischen Ein- /Ausgabeprobleme	41
3.9	Bewertungstabelle über die Berücksichtigung der mobilgerätespezifischen Ein- /Ausgabeprobleme	41
3.10	Beispiel für einen Maßstab zur Gewichtung	43
3.11	Beispiel für eine mögliche Zuordnung von Kriterien, Bewertungen und Gewich- tungen	44
4.1	Übersicht der Methoden zur Usability Inspektion	61

Abkürzungsverzeichnis

\$	United States Dollar
%	Prozent
€	Euro
API	Application Programming Interface
bzw.	beziehungsweise
DIN	Deutsches Institut für Normung
HCI	Human-Computer-Interaction
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
MMS	Multimedia Messaging Service
OS	Operating System
PC	Personal Computer
RIM	Research In Motion
s.o.	siehe oben
SD	Standard Deviation (Standardabweichung)
SMS	Short Message Service
u.a.	unter anderem
UE	Usability-Engineering
UI	User Interface

Kapitel 1

Einführung

In dieser Einführung in die Masterarbeit über einen Vergleich von Verfahren zum Testen der Usability von mobile Anwendungen wird vorerst erläutert, wieso dieses Thema in der heutigen Zeit relevant ist und für wen es wichtig ist, sich mit diesem Thema auseinanderzusetzen. Darauf folgt die Zielsetzung dieser Arbeit und einer Schilderung darüber, welcher Nutzen aus ihren Ergebnissen gewonnen werden soll. Im letzten Teil dieses Kapitels dient Überblick über den Aufbau der Arbeit dazu, den Weg aufzuzeigen, der zu den vorher genannten Zielen führen soll.

1.1 Motivation

„It is far better to adapt the technology to the user than to force the user to adapt to the technology.“ - Larry Marine

Mit diesem Satz verdeutlicht Larry Marine, ein Experte für Usability und User Experience¹, worauf die Konzentration bei der Entwicklung neuer Technologien gerichtet sein sollte. Die Anpassung der Technologie an den Benutzer und somit eine Verbesserung der Usability sind essentiell für den Erfolg einer Technologie, mit der ein Mensch interagieren soll. Nach Jordan [Jor02] hat sich diese Wirkung der Usability auf die Akzeptanz einer Technologie im Laufe der Zeit entwickelt. Während die Usability zu Beginn der Mensch-Computer-Interaktion noch als „satisfier“, also als Bonus einer Technologie, galt, wird sie heutzutage vorausgesetzt. Die Benutzer einer Technologie erwarten, dass sie das Produkt effektiv, effizient und zu ihrer Zufriedenstellung verwenden können. Dadurch wächst der Druck auf die Entwickler solcher Technologien, die die Produkte nun zu mehr als nur zur Zufriedenstellung der Kunden entwickeln müssen [MS09].

Mobile Endgeräte gehören zu den Technologien, für die die Mensch-Computer-Interaktion und somit auch die Usability von besonderer Wichtigkeit sind. In den letzten Jahren wuchs der Markt dieser Endgeräte stark an. Nach einer Gartner Studie wurden im Jahr 2010 1,6 Milliarden mobile Endgeräte verkauft, was einem Zuwachs von 31,8 % zum Jahr 2009 bedeutet. Der Verkauf von Smartphones stieg sogar um 72,1 % [Gar11].

Dass die Usability ein wichtiges Kriterium für den Käufer von mobilen Technologien ist, fanden Mack & Sharples [MS09] in einer Studie über die Relevanz der Usability in der Produktauswahl heraus. In Abbildung 1.1 ist ein Ergebnis ihrer Datenerhebung als Diagramm dargestellt, welches angibt, wie viele der Benutzer ein bestimmtes Attribut als Grund für ihre Auswahl eines

¹ „Disziplin, die sich mit der Gestaltung der gesamten mit einem Produkt oder einer Dienstleistung verbundenen Kommunikation zum Benutzer auseinandersetzt.“ [kom10]

Mobiltelefons nannten. Darin wird deutlich, dass über 50 % die Usability als ein entscheidendes Kriterium für den Kauf eines Mobiltelefons angaben. Daraus ist zu schließen, dass sich über die Hälfte der Nutzer von mobilen Endgeräten der Existenz von Usability bewusst sind und einen gesteigerten Wert darauf legen. Für Anwendung, die auf mobilen Endgeräten bedient werden

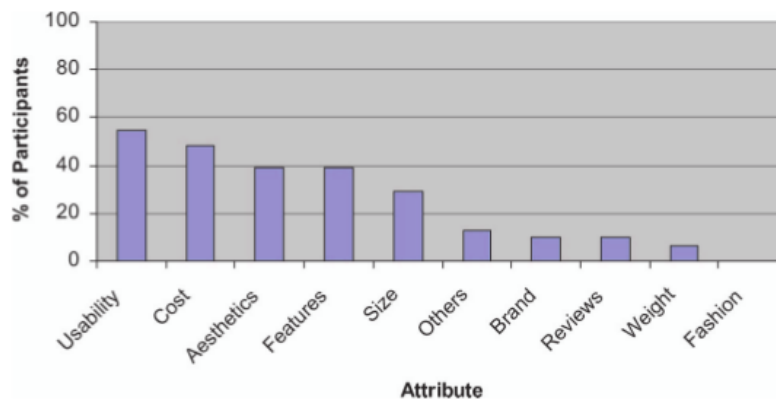


Abbildung 1.1: Anteile der Teilnehmer, die ein jeweiliges Attribut als Relevant für ihre Auswahl eines Mobiltelefons angaben [MS09]

sollen, gilt demnach, den Ansprüchen dieser Nutzer gerecht zu werden. Ein „gut benutzbares Telefon“ ist nur sinnvoll, wenn auch die darauf installierten Anwendungen benutzerfreundlich zu bedienen sind. Andernfalls würde die Usability des Endgerätes unter der Software leiden.

Eine Marktanalyse von Flurry Analytics & Estimates [Far11] ergab, dass, während im Januar 2010 etwa 300 Millionen mobile Anwendungen für die Betriebssysteme Android und iOS heruntergeladen wurden, im Oktober 2011 mit 2,6 Milliarden Downloads zu rechnen ist. Nach einer Studie von ABI Research sollen bis zum Jahr 2016 insgesamt 44 Milliarden mobile Anwendung heruntergeladen worden sein. Ähnlich schnell wächst auch die Anzahl der mobilen Applikationen, von denen im März 2009 noch insgesamt etwa 30.000, im Oktober 2011 aber schon ungefähr 850.000 verfügbar waren [Far11, Res11].

Auf diesem rapide anwachsenden Markt gilt es für die Anbieter mobiler Anwendungen, sich durchzusetzen und von der Konkurrenz hervorzuheben. Um dies mit Hilfe verbesserter Usability der eigenen Anwendung möglich zu machen, ist eine Analyse ihrer Schwachpunkte notwendig. Zu diesem Zweck sind bereits Methoden entwickelt worden. Zum einen gibt es Verfahren, die sich bereits für das Testen der Usability von Programm für andere Softwarearten bewährt haben. Zum anderen gibt es Verfahren, die speziell für das Testen von mobilen Anwendungen konzipiert wurden. Für ein Unternehmen oder eine Privatperson, die eine solche Applikation erfolgreich vertreiben möchte, gilt es dementsprechend herauszufinden, welche dieser Methoden für sie am ehesten geeignet ist.

1.2 Zielsetzung

Die Vielzahl der unterschiedlichen Verfahren, die sich zum Teil nur durch wenige Aspekte voneinander abgrenzen, die Heterogenität der Nutzungssituationen und Zwecke der mobilen Anwendungen sowie die ungleichen vorhandenen Ressourcen bei den Vertreibern der Anwendungen machen es unmöglich, eine Usability-Testmethode zu finden, die für alle denkbaren Anwendungen und deren Anbieter am besten geeignet ist. Ziel dieser Arbeit soll deshalb vielmehr die Erarbeitung einer Basis sein, auf der die Methoden miteinander verglichen werden können. Diese Basis soll anschließend verwendet werden, um ausgewählte Methoden, die ein möglichst großes Spektrum der Verfahrensvielfalt abdecken, zu evaluieren.

Auf diese Weise ist es möglich, aus den ausgewählten und evaluierten Methoden unter Berücksichtigung von gegebenen Prioritäten die beste für eine Anwendung auszuwählen. Dadurch wird eine Möglichkeit aufgezeigt, die Anwendung hinsichtlich ihrer Usability zu testen. Bei Bedarf kann anschließend weiter nach Abwandlungen dieser Usability-Testmethode geforscht werden, die unter Umständen noch besser für den Test geeignet sind. Jedwede Methode zum Testen der Usability von mobilen Anwendungen soll wiederum mit Hilfe der in dieser Arbeit geschaffenen Basis evaluiert werden können.

1.3 Aufbau

Damit die Evaluation der Usability-Testmethoden möglich ist, werden zunächst die grundlegenden Faktoren zur Messung der Usability einer Anwendung umrissen. Darauf folgt eine Darlegung der Besonderheiten, die durch die Mobilität der Endgeräte entstehen. Aus der Analyse dieser Eigenschaften der Usability und der mobilen Anwendungen sowie durch die Berücksichtigung wirtschaftlicher Aspekte werden die Kriterien erarbeitet, auf deren Grundlage die Evaluierung der Usability-Testmethoden durchgeführt werden kann. Die anschließende Vorstellung eines mathematischen Verfahrens schildert ein mögliches Vorgehen zur Priorisierung einzelner Kriterien und Evaluierung der Methoden. In Kapitel 4 werden daraufhin die in dieser Arbeit berücksichtigten Testverfahren beschrieben und mit Hilfe der erarbeiteten Kriterien bewertet.

Kapitel 2

Grundlagen

In diesem Kapitel werden die grundlegenden Aspekte beschrieben, aus denen sich Kriterien zur Bewertung der Usability-Testmethoden ableiten lassen. Zu diesem Zweck erfolgt zunächst eine Definition des Begriffes Usability und eine Beschreibung dessen, was für das Testen und Evaluieren von Usability von Relevanz ist. In Abschnitt 2.2 wird daraufhin erklärt, was unter mobilen Anwendungen zu verstehen ist, und die Faktoren erläutert, die für das Testen der Usability von mobilen Anwendungen von Belang sind.

2.1 Usability

Dieser Abschnitt beginnt mit der Definition von Usability und der Schilderung der dazugehörigen Attribute. Darauf folgt eine Zusammenfassung der für diese Arbeit relevanten Themen des Usability-Engineerings, in der unter anderem erklärt wird, wie sich die Verbesserung der Usability in den Softwareentwicklungsprozess eingliedert und welche Verfahren des Usability-Engineerings dem Usability-Testing zugeordnet werden. Abschnitt 2.1.5 zeigt daraufhin Möglichkeiten auf, Metriken zur Evaluation der Usability anzuwenden. Abschließend wird eine kurze Übersicht über die bestehenden Normen und Standards zum Thema Usability gegeben.

2.1.1 Definition

Der Begriff *Usability* wird im Allgemeinen durch Nutzbarkeit, Gebrauchstauglichkeit oder Benutzerfreundlichkeit übersetzt. Wird die folgende Definition von Usability in der Norm ISO/IEC 9126 zur Sicherstellung von Softwarequalität betrachtet, kommt „Benutzerfreundlichkeit“ dieser Definition wohl am nächsten.

„Software Usability is the capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions.“ [ISO91]

In dieser Norm ist Usability ein Teil des Software Product Quality Models und umfasst die Eigenschaften Verständlichkeit, Erlernbarkeit, Bedienbarkeit und Attraktivität [ISO91].

In der ISO Norm 9241² ist die Usability eines Produktes das Ausmaß, in dem das Produkt von einem bestimmten Benutzer verwendet werden kann, um bestimmte Ziele in einem bestimmten Kontext effektiv, effizient und zufriedenstellend zu erreichen.

²Standard, der Richtlinien der Interaktion zwischen Mensch und Computer beschreibt.

„The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.“ [ISO99]

In dieser Definition werden bereits die Eigenschaften Effektivität, Effizienz und Befriedigung als für die Usability relevant genannt. Der anerkannte Usability-Experte Jakob Nielsen definiert Usability durch 5 solcher Attribute: „Learnability“, „Efficiency“, „Memorability“, „Errors“ und „Satisfaction“. Die Usability selbst bezeichnet Nielsen wiederum als ein Attribut zur Messung von Qualität.

„Usability is a quality attribute that assesses how easy user interfaces are to use. The word „usability“ also refers to methods for improving ease-of-use during the design process. Usability is defined by 5 quality components[...]“ [Nie03b]

Im weiteren Verlauf dieser Arbeit versteht sich Usability durch die sich ähnelnden Definitionen von Nielsen und der ISO Norm 9241. Hierbei werden die „5 quality components“ von Nielsen übernommen und dafür der Begriff „Usability-Attribute“ eingeführt. Eine Beschreibung der in dieser Arbeit verwendeten Usability-Attribute findet sich im folgenden Abschnitt.

2.1.2 Usability-Attribute

Es sei an dieser Stelle gesagt, dass die folgenden Usability-Attribute nicht die einzige Zusammenstellung von Kriterien, die für die Definition von Usability von Relevanz sind, sind. Ähnliche Attribute finden sich zum Beispiel in „A practical guide to usability testing“ [DR99], „Don’t Make Me Think. A Common Sense Approach to Web Usability“ [Kru06] sowie in „Handbook of Usability Testing: How to plan, design and conduct effective tests“ [Rub95]. Auch in der ISO Norm 9241 findet sich seit 2006 im Teil 110 [ISO06] eine Liste von Usabilitykriterien. Die sieben dort aufgeführten Kriterien sind Aufgabenangemessenheit, Selbstbeschreibungsfähigkeit, Steuerbarkeit, Erwartungskonformität, Fehlerrobustheit, Individualisierbarkeit und Lernförderlichkeit. Diese Kriterien der ISO Norm finden sich zum Teil in den folgenden fünf „Quality Components“ von Nielsen wieder [Nie03b, NHH⁺08]:

Learnability / Erlernbarkeit: Wie einfach oder schwer ist es für einen Benutzer, grundlegende Aufgaben beim erstmaligen Bedienen der Anwendung zu erledigen?

Die Kriterien Selbstbeschreibungsfähigkeit und Lernförderlichkeit, aus der ISO Norm 9241, lassen sich als Bestandteile der Erlernbarkeit verstehen.

Efficiency / Effizienz: Wie schnell kann eine bestimmte Aufgabe durchgeführt werden, wenn der Benutzer mit der Anwendung vertraut ist?

Ein ähnliches Kriterium in der ISO Norm 9241 ist die Aufgabenangemessenheit, das sich darauf bezieht, ob ein Nutzer sein Ziel auf direktem Weg und vollständig erreicht. Die Effizienz verbessert oder verschlechtert sich also abhängig von der Aufgabenangemessenheit.

Memorability / Einprägsamkeit: Wie einfach ist es für einen Benutzer, die Anwendung zu verwenden, nachdem er sie längere Zeit nicht bedient hat? Wie leicht kann er seine Kenntnisse über die Anwendung erneut abrufen?

Errors / Fehler: Wie viele Fehler macht ein Benutzer, wie schwerwiegend sind diese und wie einfach kann der Benutzer nach einem Fehler weiterarbeiten?

Das Kriterium Fehlerrobustheit der ISO Norm 9241 ist diesem Attribut von Nielsen Definition sehr ähnlich. Die Fehlerrobustheit einer Anwendung wird dadurch bestimmt, dass ein Benutzer möglichst wenig Fehler macht, wobei Fehler Aktionen sind, die nicht zum gewünschten Ergebnis führen. Es sind demnach keine Fehler gemeint, die durch fehlerhaftes Programmieren entstanden sind (Bugs), sondern Fehler, die der Benutzer durch irreführende Menüführungen, falsche Platzierungen von Buttons oder ähnlichem macht.

Satisfaction / Zufriedenheit: Wie empfindet ein Benutzer beim Bedienen der Anwendung? Hat er Freude bei der Verwendung?

Durch diese Attribute wird also die Qualität der Usability einer Anwendung definiert. Wie die Messung dieser Usability-Attribute möglich ist, wird in Abschnitt 2.1.5 näher erläutert.

2.1.3 Usability-Engineering

Der Begriff Usability-Engineering (UE) wurde u.a. von Gould [GBU88] und Whiteside et al. [WBH88] Mitte der achtziger Jahre geprägt. Zu dieser Zeit betrachteten Entwickler es als ihre Hauptaufgabe, die Technologien schneller, kleiner und leistungstärker zu machen. Die HCI Experten legten ihren Fokus jedoch auf die Usabilitykomponenten: Effektivität, Effizienz und Benutzerfreundlichkeit. Der Benutzer sollte so schnell und angenehm wie möglich sein Ziel mit Hilfe des Anwendungssystems erreichen [Kar02].

Die vor dreißig Jahren geschaffenen Grundsätze haben sich bis heute nicht viel verändert. Unter dem heutigen Verständnis des UE befasst sich der Entwickler bzw. das Entwicklerteam einer Anwendung damit, die in Abschnitt 2.1.2 genannten Kriterien und somit die Usability einer Anwendung insgesamt zu optimieren. Diese Optimierung kann überall dort stattfinden, wo Benutzer mit Systemen interagieren und somit eine Form von Benutzerschnittstelle vorhanden ist. Bereits bei Betrachtung der Usability-Attribute lässt sich jedoch feststellen, dass sich die Gebrauchstauglichkeit einer Software nicht alleine durch die Veränderung der Benutzeroberfläche verbessern lässt. Vielmehr ist auch eine niedrige Fehlerrate und eine gute Performance wichtig,

um die Fehlerrate zu mindern bzw. die Zufriedenheit des Benutzers zu erhöhen und die Anwendung effizienter zu gestalten [Ric10]. Folgendes Zitat von Albert Einstein fasst das Ziel des UE treffend zusammen:

Alles sollte so einfach wie möglich gemacht werden, aber nicht einfacher.

Die Ebenen, auf denen das Usability-Engineering stattfinden kann, werden in der Abbildung 2.1 aufgezeigt. Für die Optimierung der Usability einer Anwendung ist es demnach wichtig, die Benutzer der Anwendung sowie den Anwendungskontext und die von der Anwendung zu erledigenden Aufgaben genau zu kennen und zu analysieren. Außerdem muss der Funktionsumfang der Applikation und die dafür benötigten Informationen ermittelt werden, um anschließend die effizientesten Abläufe und Prozesse erarbeiten zu können. Diese Abläufe müssen aus Benutzersicht und möglichst intuitiv gestaltet werden. Eine enge Einbeziehung der Nutzer scheint somit für das Usability-Engineering unumgänglich [Ric10].

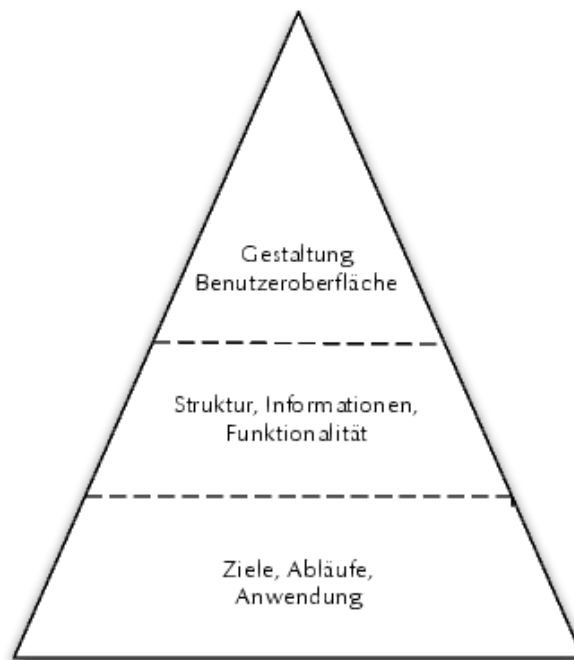


Abbildung 2.1: Ebenen der Usability-Optimierung [Ric10]

Dabei ist es essentiell, den richtigen Zeitpunkt zu finden, an dem die Verbesserung der Usability stattfinden sollte. Oder genauer: Es ist essentiell *die richtigen Zeitpunkte* zu finden. Das UE sollte nämlich nicht nur stattfinden, wenn es um die Gestaltung der Benutzeroberfläche geht, sondern kontinuierlich während des Entwicklungsprozesses und darüber hinaus [Nie94c]. Letzteres ist für diese Arbeit besonders relevant, da es bei vielen mobilen Applikationen zu häufigen Updates kommt und Entwickler diesem Trend folgen sollten, um nicht von der Konkurrenz überholt zu werden. Auf diese und weitere Besonderheiten der mobilen Anwendungen und ihren Distributionskanälen wird in Abschnitt 2.2 näher eingegangen.

Abbildung 2.2 zeigt, wie sich das UE in den Software Engineering Prozess eingliedert. Die mittlere Achse stellt dabei die Stufen dar, die in jedem Entwicklungsprozess vorhanden sein sollten [SM08]. Mit Hilfe dieses Schaubildes lässt sich erkennen, dass die Entwicklung des User Interfaces (User Interface Development) sowie die Entwicklung der Interaktionsmechanismen (User Interaction Development) bereits vor dem Design der Anwendung stattfinden sollten.

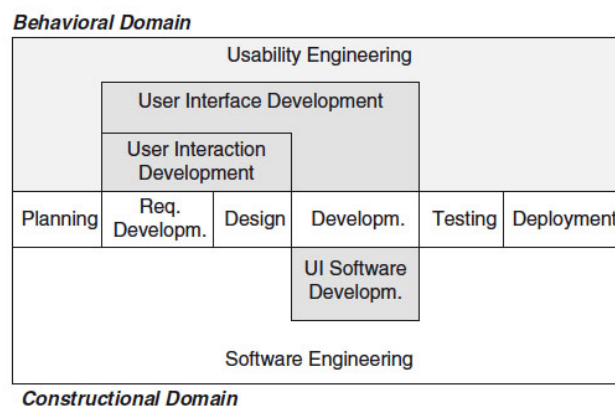


Abbildung 2.2: Anwendungsbereich des Usability-Engineerings [SM08]

Während sich in der Abbildung 2.2 das UE noch als eine Art „Black Box“ darstellt, deren einzelnen Bestandteile nicht näher definiert sind, gehen Morgan Kaufmann und Jakob Nielsen in ihrem Buch „Usability Engineering“ [Nie94c] einen Schritt weiter und beschreiben die Stufen eines *Usability Engineering Lifecycle Models*. Dieses Modell besteht aus folgenden Stufen:

1. Know the user
 - a) Individual user characteristics
 - b) The user's current and desired tasks
 - c) Functional analysis
 - d) The evolution of the user and the job
2. Competitive analysis
3. Setting usability goals
 - a) Financial impact analysis
4. Parallel design
5. Participatory design

6. Coordinated design of the total interface
7. Apply guidelines and heuristic analysis
8. Prototyping
9. Empirical testing
10. Iterative testing
 - a) Capture design rationale
11. Collect feedback from field use

Zu diesem Modell sei gesagt, dass es nicht notwendig ist, jeden Schritt durchzuführen, um beim UE ein positives Ergebnis zu erzielen. Bei begrenzten finanziellen Mitteln ist jedoch zu empfehlen, möglichst viele Schritte vor dem Design durchzuführen. Somit ist es später nicht mehr nötig, das Design wegen neuer Erkenntnisse bezüglich der Usability zu verändern. Eine Diskussion darüber, welche Schritte mit beschränkten Ressourcen und unter variierenden Zuständen zu priorisieren sind, findet sich in [Nie94c].

In der Literatur kommt es jedoch vor, dass Methoden des Usability-Engineerings mit Usability-Testmethoden oder Usability-Evaluations Methoden gleichgesetzt werden und unter Usability-Testing lediglich Tests verstanden werden, die unter Einbeziehung von einer Auswahl an *echten* Endnutzern stattfinden und diese die Anwendung während der Entwicklungsphase testen. Unter echten Endnutzern ist dabei eine Gruppe repräsentativer Nutzer zu verstehen, die ansonsten nicht direkt an der Entwicklung der Anwendung beteiligt sind und über keine oder wenige Vorkenntnisse verfügen.

In dieser Arbeit werden jedoch auch andere Methoden dem Begriff Usability-Testing untergeordnet und in dem Vergleich berücksichtigt. Zu diesen Methoden gehören jedwede Inspektionsverfahren, in dem ein Usabilityexperte eine Anwendung hinsichtlich der Gebrauchstauglichkeit analysiert und Befragungsmethoden, in denen Nutzer über ihre Meinung zu einer Anwendung, beispielsweise mit Hilfe eines Fragebogens, interviewt werden. Diese Ansätze werden in der Liste der beliebtesten Usability-Engineering Methoden von Metker et al. in Adoption-Centric Usability Engineering [SM08] jedoch als eigenständige Kategorie der UE Methoden verstanden.

Zu den beliebtesten UE Methoden gehören demnach:

1. die Befragung (Inquiry), bei der die Bewertung der Usability stattfindet, indem Benutzer gefragt werden, was ihnen an einer Anwendung gefällt bzw. nicht gefällt. Des Weiteren können hier auch Fragen gestellt werden, mit deren Hilfe herausgefunden werden kann, wie leicht die Applikation und deren Funktionen zu verstehen sind. Die Befragung kann in schriftlicher oder mündlicher Form stattfinden.

2. die Inspektion (Inspection), bei der ein Usability-Experte, in einigen Fällen auch Softwareentwickler oder andere Spezialisten, die Usability der Anwendung inspizieren.
3. das Testen (Testing), bei der eine repräsentative Gruppe von Benutzern einige typische und vorher festgelegte Anzahl an Aufgaben mit der Anwendung selbstständig erledigen soll. Dabei werden die Aktionen der Benutzer aufgezeichnet, protokolliert und/oder beobachtet. Diese Ergebnisse werden anschließend genutzt, um die Usability zu analysieren.
4. das Prototyping, bei dem möglichst schnell ein Modell (Prototyp) entwickelt wird. So können verschiedene Designs getestet werden, Ideen und Funktionen schnell, prototypisch umgesetzt und gezeigt werden, um frühe Resonanz von potentiellen Nutzern einzuholen.
5. das kognitive Modellierung (Cognitive Modelling), bei dem ein Berechnungsmodell entsteht, das darstellt, wie Aufgaben erledigt und Probleme gelöst werden. Es basiert auf psychologischen und kognitiven Grundsätzen. Mit Hilfe solcher Modelle ist es möglich, die Zeit zum Vollenden einer Aufgabe vorauszusagen, welche Fehler ein Benutzer dabei macht, welche Entscheidungen er trifft oder welche Buttons oder Menüeinträge er verwendet. Dabei kann es sich sowohl um ein Modell handeln, das schemenhaft auf dem Papier existiert, als auch um ein vollständiges Computerprogramm.
6. die Benutzeranforderungsanalyse (User Requirements Analysis), bei der Informationen über die Benutzeroberfläche, die Nutzer, Aufgaben und den Nutzungskontext gesammelt werden, um anschließend festzulegen, welche Anforderungen formalisiert werden müssen.
7. die analytischen und prädiktiven Methoden (Analytical and Predictive Methods), bei der der Entwickler der Anwendungsoberfläche voraussagen muss, wie ein Benutzer reagiert, wenn er mit einer UI oder einer UI-Komponente konfrontiert wird. Dabei versucht der Entwickler alles so einfach und intuitiv benutzbar wie möglich zu gestalten. Diese Arbeit kann häufig von Design Richtlinien unterstützt werden und findet vor allem zu Beginn der UI-Gestaltung Verwendung.

Die ersten drei dieser Methoden sind dabei die Methoden, die direkt das Testen oder zumindest Ansehen der Applikation voraussetzen und den Rahmen der in dieser Arbeit untersuchten Methoden vorgeben. Das Prototyping stellt hingegen eine Methode zur Softwareentwicklung dar, die mit Methoden des Usability-Testings kombiniert werden kann.

Eine genaue Beschreibung aller untersuchten Methoden findet sich in Kapitel 4.

2.1.4 Usability-Testing

Unter Usability-Testing ist im Allgemeinen das Testen einer Anwendung durch einen Benutzer zu verstehen. Der Benutzer muss dabei der Anwendung entsprechend sinnvolle, vorher festgelegte Aufgaben durchführen. Häufig wird als Definition von Usability-Testing auf Rubins Handbook of Usability Testing [Rub95] verwiesen, der Usability-Testing wie folgt definiert:

„[...]a process that employs people as testing participants who are representative of the target audience to evaluate the degree to which a product meets specific usability criteria.“ [Rub95, S. 29]

Dies schließt somit jegliche Methoden aus, die keine repräsentativen Testbenutzer einbinden, sondern von Experten durchgeführt werden. In dieser Arbeit versteht sich Usability-Testing jedoch so, dass Usability-Testing auch durch qualifizierte Usability-Experten durchgeführt werden kann, die die Usabilityprobleme mit Hilfe der *Heuristischen Evaluation*³ [Nie94c] aufdecken.

Bei den Usability-Tests, bei denen die testende Person kein Experte für Usability ist, ist es ratsam, den selben Test von mehreren Personen unabhängig voneinander durchführen zu lassen. Dadurch wird die Zuverlässigkeit (Reliability) des Tests erhöht und die Wahrscheinlichkeit, dass eine Wiederholung des Tests zu einem ähnlichen Ergebnis führt, wird gesteigert. Da es nicht unüblich ist, dass die schnellste Testperson ihre Aufgabe zehnmal so schnell durchführt wie die langsamste und die schnellsten 25% etwa doppelt so schnell wie die langsamsten 25% sind [Ega88], ist es wichtig, eine angemessen große Anzahl an Testnutzern auszuwählen. Da mehr Testpersonen aber auch mehr Zeit- und somit Kostenaufwand für die Durchführung und Analyse der Tests bedeutet, sollten jedoch auch nicht mehr Personen als nötig einbezogen werden [Nie94c]. Zu dieser Problematik erarbeiteten Nielsen und Landauer die Formel:

$$\text{GefundeneUsabilityProbleme} = N(1 - (1 - L)^n)$$

Wobei n die Anzahl der Testnutzer ist, N die Anzahl aller Usability-Probleme ist und L der Anteil an Fehlern, die ein einzelner Testnutzer *typischerweise*⁴ findet [NL93]. In Abbildung 2.3 wird dieses Verhältnis mit einem Wert $L = 31\%$ dargestellt. Es wird deutlich, dass eine zweite Testperson zwar neue Probleme entdeckt, allerdings nicht im selben Ausmaß wie die erste im Vergleich zu null Testpersonen. Dies liegt daran, dass der zweite Nutzer einige Usability-Probleme aufdeckt, die bereits durch den ersten erkannt wurden. Mit Hilfe einer dritten Testperson werden weitere Erkenntnisse gewonnen, jedoch werden hierbei noch weniger *neue* Daten

³Eine Erklärung der heuristischen Evaluation findet sich in Abschnitt 4)

⁴Nielsen et al. fanden durch Studien heraus, dass der typische Wert für $L = 31$ ist. Sie empfehlen jedoch einen für das jeweilige Projekt geschätzten oder berechneten Wert für L zu verwenden [Nie00].

gewonnen. Dies gilt für jede weitere Testperson, da immer mehr Probleme mehrfach identifiziert werden und der Anteil der neu gewonnenen Daten somit immer weiter sinkt.

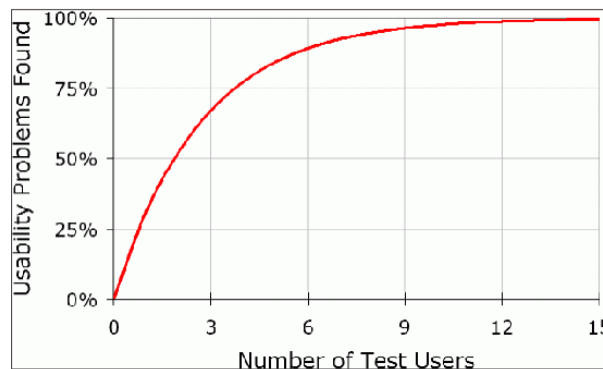


Abbildung 2.3: Verhältnis zwischen gefundenen Problemen und eingesetzten Testnutzern [Nie00]

Auf Grund dieses Verlaufs empfiehlt Nielsen fünf Testpersonen pro Test einzusetzen, die etwa 85% der Probleme aufdecken, selbst wenn das Budget beispielsweise für 15 Testpersonen ausreichen würde. Es sei sinnvoller, das restliche Budget in weitere Tests zu investieren, die durchgeführt werden, wenn die Probleme, die im ersten Test erkannt wurden, behoben wurden. Dadurch werde festgestellt, ob die Probleme tatsächlich behoben wurden und/oder neue dazu kamen. Außerdem würden die neuen fünf Testpersonen einen Großteil (98%) der im ersten Test nicht erkannten Probleme entdecken⁵ [Nie00]. Dieser Ansatz unterscheidet sich maßgeblich von früheren Usability-Tests, die noch bis in die neunziger Jahre praktiziert wurden. Zu dieser Zeit wurden die Tests noch als Forschungsexperimente angesehen und von Kognitionswissenschaftlern, experimentellen Psychologen oder Human Factors Engineers⁶ durchgeführt. Dabei wurden meistens zwischen 30 und 50 Testpersonen eingesetzt, wodurch die Kosten für einen solchen Usability-Test entsprechend hoch wurden [Bar10].

Abgesehen von der Zuverlässigkeit eines Tests, die wie eben beschrieben durch statistische Modelle gewährleistet werden kann, spielt die Validität im Usability-Testing eine wichtige Rolle. Validität bedeutet in diesem Zusammenhang, dass der Test etwas misst, was in der realen Welt bei der Benutzung der Anwendung relevant ist. Die typischen Probleme, die dabei auftauchen können, sind [Nie94c]:

- Die Testpersonen werden nicht *sinnvoll* ausgewählt. So sollten möglichst Menschen ausgewählt werden, die zur Zielgruppe der zu testenden Applikation zählen.
- Die durch die Testpersonen zu erledigenden Aufgaben spiegeln nicht die normale Verwendung des Programmes wider.

⁵Die letzten 2% werden nach Niensens Modell bei dem dritten Test identifiziert.

⁶„Human Factors Engineering is the practice and process of evaluating the way humans interact with machinery and optimizing the machinery and processes to make it wasier on humans.“ [Bam]

- Zeitrahmen und soziale Einflüsse, die in der realen Welt existieren würden, werden im Test nicht berücksichtigt.

Um einen Test so zu gestalten, dass das Ergebnis möglichst aussagekräftig ist, sollte vor dem Test ein Plan entworfen werden, in dem der Zweck des Usability-Tests genau bestimmt wird. Einige⁷ wichtige Fragen, die in diesem Plan geklärt werden müssen, sind [Nie94c]:

- Was ist das Ziel des Test? Was soll erreicht werden?
- Wo und wann findet der Test statt?
- Wie viele Testpersonen sind erforderlich?
- Welche Aufgaben sollen die Testpersonen erledigen?

Außerdem muss die Art des Tests und somit der Evaluation festgelegt werden. Hierbei wird zwischen der summativen und der formativen Evaluation unterschieden. Welche Evaluation angewandt werden soll, lässt sich in der Regel bereits durch den Zeitpunkt, wann der Usability-Test stattfinden soll, ermitteln. Die summative Evaluation eignet sich, um einen Gesamteindruck über die Usability einer Anwendung zu gewinnen. Sie kann eingesetzt werden, um einen Vergleich mit Konkurrenzprodukten durchzuführen oder sicherzustellen, dass Anforderungen eingehalten wurden, und findet im Normalfall nach der Fertigstellung der Anwendung in einem größeren Rahmen statt. Das Ergebnis dieser Evaluation sind in der Regel Messgrößen wie Erfolg oder Misserfolg bei einer Aufgaben, durchschnittliche Bearbeitungszeit, Fehlerraten usw. Diese Ergebnisse stellen häufig eine Basis dar, die bei neueren Versionen des Produktes übertroffen oder mindestens gehalten werden sollte. Die summative Evaluation kann sich dabei dieselben Aufgaben und Szenarien verwenden wie die formative Evaluation. Möglich ist aber auch, dass spezielle Funktionen oder Prozesse evaluiert werden, die fertig entwickelt und implementiert sind, um die Usability dieser Teile des Produktes zu validieren.

Die formative Evaluation hingegen begleitet den Entwicklungsprozess einer Anwendung und wird häufiger durchgeführt. Ihr Ziel ist es, Usabilityprobleme zu erkennen sowie Stärken und Schwächen des Designs aufzudecken. Ihr Wert liegt also darin, dass sie die Entwickler noch in der Entwicklungsphase des Produktes über Probleme der Anwendung informieren, die dann behoben werden können. Anschließend wird eine neue Studie zur formativen Evaluation durchgeführt, mit der überprüft werden kann, ob die Probleme behoben wurden [Nie94c, Bar10].

2.1.5 Usability-Evaluation und Metriken

In diesem Abschnitt wird ein Überblick darüber gegeben, auf welche Weise bei Usability-Tests Daten messbar und vergleichbar gemacht werden können. Ausführliche Informationen über

⁷Eine vollständige Liste der in einem solchen Testplan zu beantwortenden Fragen findet sich in [Nie94c].

geeignetes Analysieren und Präsentieren der so erlangten Daten werden in [TA08] gegeben. Der Anspruch dieses Unterkapitels ist es nicht, die unterschiedlichen Methoden, um die Daten zu gewinnen, im Detail zu erläutern. Die Beschreibung von sechs möglichen Usability-Testmethoden, die einige der folgenden Metriken verwenden, findet in Kapitel 4 statt.

Tullis & Albert gliedern die Metriken unter fünf verschiedene Gruppen. Die „*Performance Metrics*“ sind dabei ein mächtiges Werkzeug, um die Usability eines Produkts / einer Anwendung zu messen. Mit ihrer Hilfe ist es möglich, Aussagen darüber zu treffen, ob die Anwendung zur Veröffentlichung bereit ist [TA08]. Sie machen die in Abschnitt 2.1.2 beschriebenen Usability-Attribute direkt messbar. Es folgt eine Liste der Performance-Metriken. Hinter den Namen der Metriken sind die Usability-Attribute angegeben, die durch die jeweilige Metrik gemessen werden.

- *Task Success* (Learnability), um zu messen, ob die Teilnehmer eines Usability-Tests im Stande waren, eine Aufgabe abzuschließen. Hierbei wird zwischen „binärem Erfolg“ und verschiedenen Graden des Erfolgs unterschieden. Bei dem binären Erfolg wird lediglich zwischen Erfolg und Misserfolg unterschieden.
- *Time-on-task* (Efficiency, Learnability), um zu messen, wie lange ein Benutzer braucht, um bestimmte Aufgaben mit der Anwendung zu erfüllen. Dabei kann zum Beispiel die Gesamtzeit aller Nutzer, eine Gruppe von Benutzern oder der Anteil der Nutzer, die es geschafft haben, in einem bestimmten Zeitrahmen eine Aufgabe zu erfüllen, relevant sein.
- *Errors* (Errors) messen die Fehler, die gemacht wurden, während eine bestimmte Aufgabe mit der Anwendung bearbeitet wurde. Eine Aufgabe kann dabei eine oder mehrere Möglichkeiten bieten, um Fehler zu machen, wobei die Fehler von unterschiedlicher Wichtigkeit sein können.
- *Efficiency* (Efficiency) bewertet den kognitiven und physischen Aufwand, der erforderlich ist, um eine Aufgabe zu erfüllen. Sie wird häufig mit Hilfe der Anzahl an Schritten oder Aktionen, die ein Nutzer benötigt, gemessen. Eine alternative Messmethode ist das Verhältnis der Erfüllungsquote einer Aufgabe zur durchschnittlichen Zeit pro Aufgabe.
- *Learnability* (Memorability) misst, wie sich die Effizienz (Efficiency) im Laufe der Zeit verbessert. Sie erfordert meist mehrere Tests mit den gleichen Testpersonen. Die Learnability wird dazu gebraucht zu erfahren, wie und wann ein Testnutzer eine Anwendung zu einem bestimmten Grad beherrscht [TA08].

Bei der Betrachtung dieser Metriken fällt auf, dass die Usability-Attribute Effizienz (Efficiency) und Erlernbarkeit (Learnability) mit Hilfe von zwei Werten gemessen werden. Außerdem besteht eine Diskrepanz in dem Verständnis von Learnability zwischen Tullis & Albert und

Nielsen. Während letzterer damit die Eigenschaft einer Anwendung bezeichnet, einen Benutzer bei der erstmaligen Verwendung das Erlernen der Bedienung einfach zu machen, definieren Tullis & Albert die Learnability als Attribut dafür, wie sich die Effizienz der Anwendung nach mehrmaliger Nutzung durch das Lernverhalten der Nutzer verbessert. Somit ist dieses Verständnis indirekt die Einprägsamkeit (Memorability) nach Nielsen. Liegen die verschiedenen Tests weit genug auseinander, trat gegebenenfalls nur ein schwacher Lerneffekt auf und die Effizienz verschlechterte sich, was Aufschluss über die Einprägsamkeit des Programms geben würde. Während eine Verbesserung der Effizienz bedeuten würde, dass sich der Benutzer die erforderlichen Schritte zur Erfüllung einer Aufgabe gemerkt hat, was somit eine gute Einprägsamkeit impliziert [TA08, Nie03b].

Die Zufriedenheit (Satisfaction) eines Benutzers hingegen wird durch keine der Performance-Metriken unterstützt. Grund dafür ist die Problematik, die Zufriedenheit einer Person zu messen und zu beziffern. Wie die Zufriedenheit dennoch gemessen und in einem Usability-Test berücksichtigt werden kann, wird im weiteren Verlauf dieses Kapitels geschildert.

Zunächst wird jedoch die zweite Gruppe der von Tullis & Albert zusammengetragenen Metriken, die „*Issues-Based Metrics*“, vorgestellt. Dabei wird der Ansatz verfolgt, zu erkennen, was genau das Problem (Issue) während der Bedienung der Anwendung ist, und zu bewerten, wie hoch die Dringlichkeit zur Behebung dieses Problems ist. Ein Usability-Problem kann hierbei beispielsweise bestimmter Teil des Inhalts oder der Navigation sein. Aber auch ein Teil der Anwendung, der aus einem bestimmten Grund nicht wahrgenommen wird, obwohl er wahrgenommen werden sollte, ist ein mögliches Usability-Problem. Folgende Liste gibt ein besseres Verständnis darüber, was Usability-Probleme sind [TA08]:

- alles, was das Erfüllen der Aufgabe verhindert
- alles, was vom Erfüllen der Aufgabe ablenkt
- alles, was Verwirrung erzeugt
- alles, was einen Fehler zur Folge hat
- wenn etwas nicht gesehen wird, was beachtet werden sollte
- wenn etwas als korrekt angenommen wird, was nicht korrekt ist
- wenn angenommen wird, dass eine Aufgabe erfüllt wurde, wenn es nicht der Fall ist
- wenn die falsche Aktion durchgeführt wird
- wenn ein Teil des Inhalts fehlinterpretiert wird
- wenn die Navigation nicht verstanden wird

Als eine Möglichkeit, Usability-Probleme zu entdecken, nennen Tullis & Albert die Laborstudie⁸, bei dem mehrere Personen die Testnutzer beobachten. Wobei es dabei noch gilt herauszufinden, ob ein entdecktes Usability-Problem tatsächlich ein solches ist oder durch eine ungewöhnliche Aktion eines Benutzers hervorgerufen wird. Ungewöhnliche Aktionen sind dabei solche Aktionen, für die es keine rationale Erklärung gibt und bei denen es unwahrscheinlich ist, dass das Problem bei anderen Nutzern ebenfalls auftritt. Anschließend können die Usability-Probleme verschieden gewichtet werden. So wird festgelegt, wie dringend ein Problem behoben werden muss [TA08]. Es gibt zwei weit verbreitete Arten, die Gewichtung vorzunehmen. Bei der ersten findet eine Einordnung des „Problems“ auf einer Skala von null bis vier statt, wobei null dabei bedeutet, dass eigentlich kein Problem vorliegt. Eine Einstufung auf 1 besagt, dass ein kosmetisches Problem vorliegt, was nicht behoben werden muss, falls nicht genügend Zeit verfügbar ist. Eine 2 bedeutet, dass ein Problem mit geringer Priorität vorliegt, eine 3 ist hingegen ein bedeutenderes Problem und eine 4 ist eine „Usability-Katastrophe“. Ein solches Problem muss auf jeden Fall gelöst werden, bevor die Anwendung veröffentlicht werden kann [Nie94c].

Die zweite Art, eine Gewichtung vorzunehmen, wird in Abbildung 2.4 dargestellt. Dabei wird ein Portfolio erstellt, in dem die Probleme in vier Felder eingeteilt werden können. Der Unterschied zu der zuvor vorgestellten Gewichtungsmethode ist, dass hierbei nur zwischen niedriger, mittlerer und hoher Schwere des Problems unterschieden wird. Diese ergeben sich aus dem Verhältnis daraus, wie viele Benutzer ein Problem entdeckt haben, zu dem Einfluss, den das Problem auf das „Benutzungserlebnis“ (User Experience) hat.

		Proportion of users experiencing the problem	
		<i>Few</i>	<i>Many</i>
Impact of problem on the users who experience it	<i>Small</i>	Low severity	Medium severity
	<i>Large</i>	Medium severity	High severity

Abbildung 2.4: Einordnung von Usabilityproblemen nach dem Einfluss auf die User Experience im Verhältnis zu der Anzahl an Benutzern, die das Problem feststellten [Nie94c].

Die dritte Gruppe der Metriken aus [TA08], sind die „*Self-Reported Metrics*“. Die Art der dort gesammelten Daten werden auch „subjective data“ oder „preference data“ genannt. Sie werden dadurch gewonnen, dass die User nach seinen Erlebnissen mit einer Anwendung gefragt werden.

⁸Tatsächlich besteht die Aufgabe jeder Usability-Testmethode darin, die Usability-Probleme einer Anwendung aufzudecken.

Somit werden auch Informationen über die Zufriedenheit der Testpersonen, die durch die Performance-Metriken nicht erlangt werden konnten, gewonnen. Dabei gibt es viele verschiedene Arten, einen User nach seiner Meinung über die Anwendung zu fragen. Zu den effizientesten zählen laut Tullis & Albert Fragebögen, deren Fragen mit einer Art von Bemessungsskala beantwortet werden können. Solche Skalen sind zum Beispiel die „Likert-Skala“ oder Osgoods „Semantic Differential“-Technik⁹. Es gibt drei verschiedene Möglichkeiten, einen Benutzer mit Hilfe eines solchen Fragebogens über die Usability eines Systems auszufragen. Für den Testnutzer ist es gewöhnlicherweise am einfachsten, wenn er seine Bewertung verbal abgeben kann. Dieses Verfahren eignet sich zum schnellen Bewerten einzelner Aspekte nach jeder zu erledigenden Aufgabe. Ein Fragebogen in Papierform sowie ein Online-Fragebogen eignen sich sowohl für längere Befragungen als auch für ein schnelles Rating, wobei die Online-Variante an Popularität gewinnt. So kann ein Befragter beispielsweise direkt zu einer Folgefrage, die sich aus einer Antwort ergibt, weitergeleitet werden, während er sie bei einem Fragebogen in Papierform erst suchen müsste. Es hat sich nach Tullis & Albert bewährt, einen Testnutzer den Fragebogen online ausfüllen zu lassen und ihm gleichzeitig die Möglichkeit zu geben, auf die Anwendung zurückzugreifen [TA08]. Durch die Anwendung der Rating Skalen in den Fragebögen lassen sich die Bewertungen messbar und visualisierbar machen. Denkbar sind hier zum Beispiel Balkendiagramme, die bei größeren Umfragen den Mittelwert und die Standardabweichung zu den einzelnen abgefragten Bewertungen darstellen. Aber auch ein Radar-Chart kann sinnvoll sein, um die Schwachpunkte einer Applikation zu erkennen.

Bei der vierten Gruppe der Metriken handelt es sich um die „*Behavioral and Physiological Metrics*“. Mit ihnen soll das Verhalten der Testpersonen, während sie eine Aufgabe erfüllen oder einen Fragebogen ausfüllen, messbar gemacht werden. Die verbalen Äußerungen sowie Mimik und Gestik des jeweiligen Nutzers kann dabei entweder direkt durch den Testleiter oder eine Person, die extra dafür den Test verfolgt, wahrgenommen und notiert werden, oder durch eine Kamera aufgezeichnet werden. Aber nicht nur das direkt Sichtbare kann Aufschluss darüber geben, was eine Person beim Benutzen einer Anwendung empfindet. Auch die Herzfrequenz, Pupillenerweiterungen oder leichter Schweißausbruch können beispielsweise wichtige Informationsquellen für die kognitive Belastung sein.

Verbale Äußerungen, die ein Testteilnehmer während der Benutzung der Anwendung von sich gibt, lassen sich in drei Kategorien unterteilen: positiv, neutral oder negativ. Neutrale Äußerungen sind dabei beispielsweise Sätze wie „Das ist aber interessant“. Je nach Bedarf kann man allerdings auch feinere Granularitäten wählen, um zum Beispiel zwischen sehr positiven Kommentaren und anderen positiven Kommentaren zu unterscheiden. So lassen sich die verbalen Reaktionen auf eine Anwendung kumulieren und in ein Verhältnis bringen. In Abbildung 2.5 werden die Anteile der positiven, neutralen und negativen Resonanz beispielhaft dargestellt.

⁹Für Erläuterungen zu den Skalierungen sei hiermit auf Likerts „A technique for the measurement of attitudes“ [Lik32] und Osgoods „The measurement of meaning“ [OST71] verwiesen.

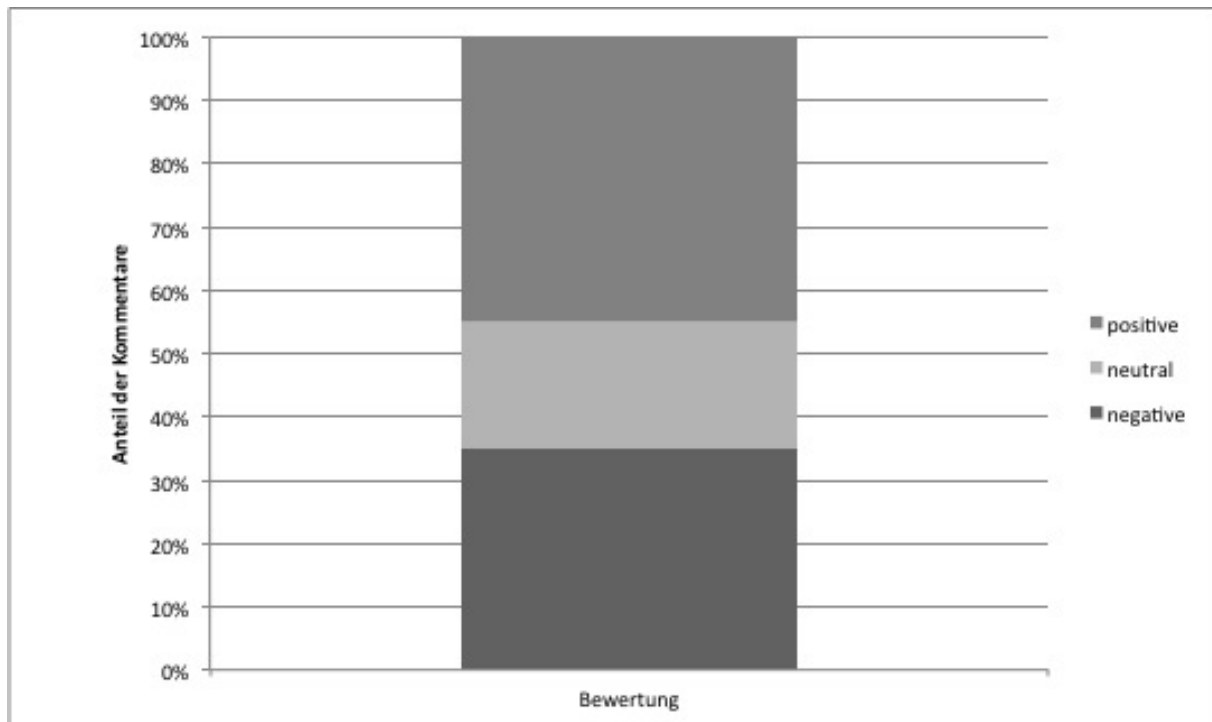


Abbildung 2.5: Beispiel für die Veranschaulichung der Anteile von positiven, neutralen und negativen Kommentaren für eine Anwendung. In Anlehnung an [TA08]

Nonverbale Äußerungen wie zum Beispiel Lächeln, Hochziehen der Augenbrauen oder am Kopf kratzen können auch Hinweise auf die Gedanken und den Gemütszustand eines Testnutzers geben. Demnach können weitere Daten zur Messung der Zufriedenheit einer Testperson während der Bedienung der Anwendung gewonnen werden. Für die Datensammlung sind jedoch auch die Augen beziehungsweise die Blickrichtung des Benutzers von Interesse. Indem Eye-Tracking Soft- und Hardware verwendet wird, kann bestimmt werden, auf welchen Teil eines Designs der Fokus der Testperson liegt. So kann zum Beispiel gemessen werden, wie lange die Person auf einem bestimmten visuellen Abschnitt der Anwendung verharret, ob der Nutzer nach etwas sucht (zum Beispiel einer Funktion oder einem Button) oder ob er mit Lesen beschäftigt ist. Die auf diese Weise erlangten Werte können anschließend in ein Verhältnis zueinander gebracht werden, um Aufschluss darüber zu erlangen, womit ein Benutzer die meiste Zeit beim Erledigen einer Aufgabe verbringt und ob die Dauer angemessen ist. Auch die Größenänderung der Pupillen kann ein relevanter Wert für die Usability-Evaluation werden, da die Pupillen von Menschen dazu neigen, sich zu vergrößern, wenn der mentale Anspruch an eine Person wächst [TA08].

Anshu Agarwal und Andrew Meyer untersuchten in einer Studie, ob es Unterschiede in den durch „*Behavioral and Physiological Metrics*“ aus Emotionen gewonnen Daten und denen aus traditionellen Usability-Metriken (Performance Metrics) gibt. Zu diesem Zweck entwickelten sie eine Methode, in der verbale und nonverbale Äußerungen in *einer* Skala kombiniert

werden können. Anschließend ließen sie Testnutzer zwei verschiedene Designs einer Customer Relationship Management-Software anwenden und prüften, ob durch die Behavioral and Psychological Metrics Unterschiede zwischen den beiden Designs aufzudecken sind, die die Performance Metrics nicht entdeckt wurden. Das Ergebnis dieser Forschung war, dass, obwohl mit Hilfe der traditionellen Metriken keine signifikanten Unterschiede zwischen den beiden Interfaces festgestellt werden konnte, durch die aus den Emotionen der Nutzer gewonnen Daten ein Unterschied der beiden Oberflächen ausgemacht werden konnte. Die „Behavioral and Physiological Metrics“ machen es demnach möglich, Einblicke in die Wahrnehmung der Usability durch seinen Benutzer zu erlangen. Diese Informationen könnten dann eingesetzt werden, um die Usability der getesteten Anwendung zu verbessern [AM09].

2.1.6 Normen und Standards

Die International Standardization Organization (ISO) veröffentlichte bereits mehrere Normen, die den Herstellern von Anwendungen und Endgeräten dabei helfen sollen, benutzerfreundliche Systeme zu entwickeln. Die Hersteller können die Normen als Leitfäden für die Entwicklung verwenden. Im Idealfall wird durch die Einhaltung der Normen eine höhere Akzeptanz bei den Benutzern erzielt, da die Systeme einen gewissen Wiedererkennungswert besitzen und benutzerfreundlich gestaltet sind. Der Wiedererkennungswert wiederum beeinflusst das Empfinden eines Benutzers über die Usability eines mobilen Endgerätes [MS09] und Menschen tendieren dazu zu präferieren, was ihnen bekannt vorkommt [Kha01].

Abbildung 2.6 zeigt die ISO Normen, die Richtlinien für das Usability-Engineering bereitstellen. Hierbei handelt es sich jedoch nicht um Normen, die explizit für die Entwicklung von mobilen Anwendungen konzipiert wurden, sondern vielmehr entstanden sind, als die Mensch-Computer-Interaktion noch hauptsächlich auf stationären Geräten stattfand.

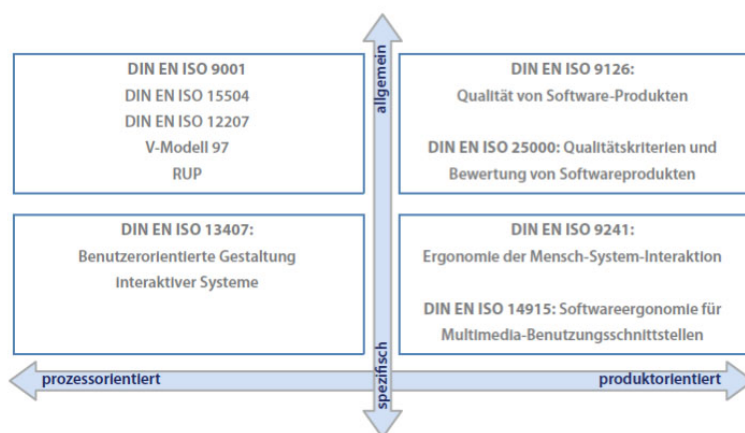


Abbildung 2.6: Normen und Standards des Usability-Engineerings [Kra, S. 136].

Krannich analysiert in seiner Dissertation über Mobile Usability Testing [Kra] die spezifisch für die Benutzerinteraktion ausgearbeiteten Normen, die in der unteren Spalte der Matrix (Abbildung 2.6) dargestellt sind. Er kommt zu dem Schluss, dass sich die Normen wegen der besonderen (physikalischen) Eigenschaften mobiler Endgeräte (vgl. Abschnitt 2.2.2) nur bedingt anwenden lassen. Des Weiteren stellt Krannich fest, dass Usability entsprechend der Normen möglichst früh, also bereits in der Planungsphase, eine hohe Aufmerksamkeit verdienen sollte. Daraus folgt, dass auch die Benutzer schon früh an dem Entwicklungsprozess beteiligt sein sollten. So können die Ansprüche der Benutzer, der Benutzungskontext, die Arbeitsziele und die Anwendungsfunktionen als sich gegenseitig beeinflussende Faktoren während der Entwicklung berücksichtigt werden [Kra, S. 136 ff.]. Eine frühe Einbeziehung der möglichen Nutzer und ein Berücksichtigen des Nutzungskontexts werden somit zu wichtigen Kriterien für Bewertung von Usability-Testmethoden.

2.2 Mobile Anwendungen

In diesem Abschnitt werden die Besonderheiten der mobilen Anwendungen und deren Usability-Testing vorgestellt. Dazu werden mobile Anwendungen definiert und die in dieser Arbeit berücksichtigten Typen mobiler Endgeräte vorgestellt. Darauf folgt eine Übersicht über die Hardware dieser Geräte und der darauf installierten Betriebssysteme sowie eine Beschreibung des Nutzungskontexts und der Nutzungssituation. Dabei wird erklärt, wieso diese Faktoren für das Usability-Testing relevant sind. Anschließend werden Hilfsmittel vorgestellt, die bei dem Testen von Anwendungen zum Einsatz kommen können.

2.2.1 Begriffserklärung: mobile Anwendungen

Da der Begriff „mobile Anwendungen“ keinesfalls selbsterklärend ist, erfolgt hier eine kurze Definition dessen, was im Kontext dieser Arbeit unter mobilen Anwendungen verstanden wird. Eine „Anwendung“ ist in diesem Zusammenhang im Sinne der Informatik als Programm oder Anwendungssystem zu verstehen und sollte den Lesern dieser Arbeit bekannt sein. Dabei ist es für die Usability-Tests von geringer Bedeutung, in welcher Programmiersprache die Anwendung geschrieben wurde. Lediglich die Implementation von zusätzlichen Methoden, um bestimmte Messwerte aufzeichnen zu können, kann für einige Usability-Testmethoden von Bedeutung sein. Für stationäre Systeme gibt es zum Beispiel eine Usability-Testmethode, die die automatische Protokollierung der Anwenderaktionen (Logging) zur Auswertung benutzt [Nie94c, S. 217 ff.]. Eine Umsetzung dieser Methode für mobile Endgeräte müsste also entsprechende Funktionen implementieren können.

„Mobil“ hingegen kann unterschiedlich rezipiert werden. So kann beispielsweise ein Laptop

oder Netbook als mobiles Gerät bezeichnet werden, da es an verschiedenen Orten gleichermaßen benutzt und transportiert werden kann. Im Fokus dieser Arbeit stehen jedoch Geräte, die in Bewegung benutzt werden können und sich dabei mit verschiedenen Netzwerkzugangspunkten verbinden. Dabei sind die Geräte darauf ausgelegt, möglichst wenig Strom zu verbrauchen. Außerdem sind die Benutzungsschnittstellen für den mobilen Einsatz entworfen. Dies trifft immer noch auf eine große Anzahl heterogener Endgerädetypen zu, die große Unterschiede in Hard- und Software aufweisen. Um die Methoden zum Testen der Usability sinnvoll analysieren zu können, ist es deshalb notwendig, die in Betracht gezogenen Geräte weiter einzugrenzen. In dieser Arbeit werden mobile Anwendungen dementsprechend Anwendungen sein, die speziell für mobile Endgeräte konzipiert wurden, und unter mobilen Endgeräten werden vorrangig Mobiltelefone, Smartphones und PDAs zusammengefasst. Einige der untersuchten Usability-Testmethoden lassen sich jedoch genauso oder ähnlich wie in Kapitel 4 vorgestellt auf mobile Endgeräte, wie mobilen Spielkonsolen (Nintendo Game Boy, Playstation Portal u.ä.) oder mobile Diagnose-/ Datenerfassungsgeräte, übertragen. Ob eine der analysierten Methoden für eins der in dieser Arbeit nicht berücksichtigten Endgeräte anwendbar ist, ist bei Bedarf also gesondert zu überprüfen. Dies trifft vor allem auf solche Methoden zu, die das Installieren einer zusätzlichen Software auf dem jeweiligen Endgerät erfordern.

2.2.2 Usability-Testing für mobile Anwendungen

In diesem Abschnitt werden nicht explizit einzelnen Methoden zum Usability-Testing beschrieben. Es soll viel mehr Aufschluss darüber gegeben werden, was das Besondere daran ist, Anwendungen für mobile Endgeräte zu testen. Dabei wird unter anderem auf den Unterschied zum Testen für stationäre Systeme wie Desktop Computer hingewiesen. Dieser Unterschied ergibt sich aus zwei Kernpunkten. Zum einen ist die Situation, in der die mobilen Geräte genutzt werden, eine andere als bei stationären. Diese Nutzungssituation wird im zweiten Teil dieses Abschnittes erläutert. Zum anderen ist der Hard- und Softwareunterschied zwischen mobilen und stationären Geräten ein zu berücksichtigender Gesichtspunkt. Zudem sind auch die Differenzen zwischen den Spezifikationen der mobilen Endgeräte (Mobile Device Fragmentation) für das Usability-Testing von Bedeutung und werden erläutert. Abschließend werden mögliche Verfahren für die Aufzeichnung von Usability Tests vorgestellt. Diese Verfahren sind unabhängig von den Usability-Testmethoden und dienen als Hilfsmittel für die spätere Analyse der Tests.

Hardware und Betriebssystem

Nach einer Studie von Nielsen aus dem Mai 2011 sind 55% der verkauften Mobiltelefone von März 2011 bis Mai 2011 Smartphones¹⁰ [Com11]. Es ist zwar auch möglich, Anwendungen für weniger leistungsstarke Mobiltelefone zu entwickeln. Der Trend der verkauften Mobilgeräte scheint jedoch deutlich in Richtung Smartphones zu gehen und somit sind die Betriebssysteme dieser Endgeräte für Unternehmen und Anwendungsentwickler deutlich interessanter. Hinzu kommt, dass die Möglichkeiten der Anwendungen mit der Hardware wachsen und sich spezielle vorinstallierte Marktplätze für den Vertrieb der Applikationen etabliert haben, was die Attraktivität der Smartphones für kommerzielle Zwecke noch erhöht. In Abbildung 2.7 wird dargestellt, wie sich die Betriebssysteme auf die in der erwähnten Nielsen Studie verkauften Smartphones verteilen. Dabei fällt auf, dass die drei häufigsten Betriebssysteme bereits auf über 85% dieser verkauften Endgeräte installiert sind. Soll eine zu entwickelnde Anwendung also eine breite Masse ansprechen, ist die Auswahl des zu unterstützenden Betriebssystems demnach stark eingeschränkt und spielt somit für die Usability-Tests eine geringfügige Rolle. Doch sogar un-

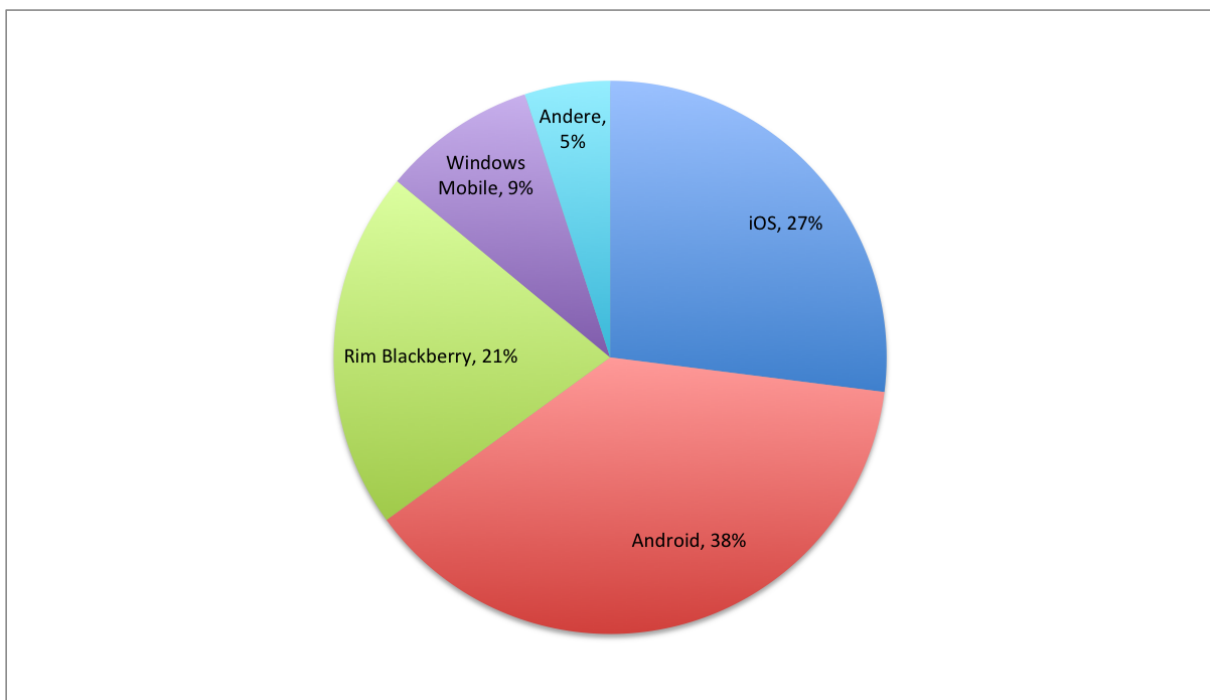


Abbildung 2.7: Verteilung der Betriebssysteme auf zwischen März 2011 und Mai 2011 in den USA akquirierten Smartphones. In Anlehnung an [Com11].

ter diesen stark eingeschränkten Voraussetzungen, nur für Smartphones, die eines der drei am meisten verbreiteten Betriebssystemen verwenden. zu entwickeln und zu testen, gibt es entscheidende Unterschiede zwischen den Endgeräten, die für die Usability von Relevanz sind. Am offensichtlichsten sind die Unterschiede im Bereich der Ein- und Ausgabe. Während die

¹⁰Geräte, die einem Computer in Leistungsfähigkeit und Konnektivität ähnlich sind.

Eingabe bei einem Computer normalerweise über eine Tastatur und / oder Maus bzw. Touchpad erfolgt, gibt es bei Mobiltelefonen eine Reihe von Möglichkeiten. Dabei haben sich bisher drei Eingabemethoden durchgesetzt:

- Ein Ziffernblock mit den Zahlen von 0 bis 9 sowie zwei Extratasten, so dass eine Reihe von 4 mal 3 Knöpfen entsteht, die jeweils mit einer Zahl und mehreren Buchstaben belegt sind.
- Ein QWERTY- bzw. QWERTZ Keyboard¹¹ das ein Hardwarebestandteil des Gerätes ist.
- Ein Keyboard oder andere Buttons, die auf einem Touchscreen erscheinen und deren Layout sich individuell gestalten lässt.

Bei Geräten, die Apples iOS verwenden, wurde bisher immer die reine Touchscreen-Eingabe verwendet, so dass Entwickler für dieses Betriebssystem zumindest keine Rücksicht auf verschiedene, durch das Endgerät bestimmte Eingabemethoden nehmen müssen. Bei RIMs BlackBerry Endgeräten hingegen besitzen die meisten Modelle ein Hardware-Keyboard, allerdings gibt es auch Modelle wie das BlackBerry Storm, bei dem ein Touchscreen zur Eingabe dient. Die größten Unterschiede weisen jedoch die Modelle, die Google's Android OS verwenden, auf, da dieses Open Source Betriebssystem von vielen verschiedenen Herstellern eingesetzt wird. Doch die Eingabe von Zeichen oder das Drücken von Knöpfen sind nicht die einzigen möglichen Eingabemethoden dieser Art mobiler Endgeräte. Einige Anwendungen, wie beispielsweise Barcode Scanner, verwenden die integrierten Kameras der Mobiltelefone zum Auslesen von Informationen. Neuere Geräte, die die Near Field Communication (NFC) Technologie verwenden, können damit Daten ohne Berührung erfassen. Außerdem ist auch eine Spracheingabe auf vielen Smartphones verfügbar.

Die Performance der mobilen Anwendungen wird (neben der Programmiersprache und Qualität des Quellcodes) vor allem durch den Prozessor und den Arbeitsspeicher bestimmt¹². Mit diesen beiden Faktoren sowie dem internen Speicher verhält es sich wie bei stationären Endgeräten. Die Geschwindigkeit der Endgeräte schwankt durch die Vielzahl unterschiedlich schneller Speichermedien und CPUs erheblich. Die Entwickler und Tester stoßen somit auf die gleichen Herausforderungen, nämlich ihre Anwendung unter möglichst vielen verschiedenen Hardwarebedingungen zu testen und bestmöglich benutzbar zu machen.

Auf den ersten Blick scheint es sich bei den Ausgabemedien ähnlich zu verhalten. Schließlich gibt es auch bei PCs keinen einheitlichen Bildschirm mit derselben Displaygröße, Auflösung

¹¹In manchen Ländern sind andere Tastaturlayouts verbreitet. So ist zum Beispiel in Frankreich ein AZERTY-Keyboard üblich.

¹²Die Geschwindigkeit der Internet- beziehungsweise Netzwerkverbindung ist bei mobilen Anwendungen, die auf eine bestehende Internetverbindung angewiesen sind, ebenfalls von Relevanz. Dabei ist gesondert darauf hinzuweisen, dass die Verbindungsqualität bei der Benutzung in Bewegung variieren kann, was auch im Kontrast zu der gleichbleibenden Verbindung eines stationären Gerätes steht.

(und die sich daraus ergebenden Pixeldichte), Kontrast, Farbtiefe oder Reaktionszeit. Wird jedoch der Nutzungskontext und die Tatsache, dass bei Touchscreens die Ausgabe gleichzeitig die Eingabe ist, berücksichtigt, werden die Differenzen zwischen den Modellen der mobilen Endgeräten relevanter.

Die Bedienung einer mobilen Applikation auf einem 3,2 Zoll großen Touchscreen, bei einer Auflösung von 320x480 Pixeln, wird sich in den meisten Fällen stark von der auf einem 4,5 Zoll Touchscreen mit einer Auflösung von 1280x720 Pixeln unterscheiden. In der Regel ist es einfacher, auf großen Bildschirmen spezielle Bereiche des Touchscreens zu berühren. Daraus lässt sich schließen, dass die Usability auf möglichst unterschiedlichen Endgeräten, sofern diese zur Zielgruppe gehören, zu testen ist. Außerdem ist es bei den verhältnismäßig kleinen Displays der mobilen Endgeräte von besonderer Bedeutung, dass die relevanten Informationen und Eingabeoptionen an der richtigen Stelle zur richtigen Zeit sichtbar sind [BCD⁺09].

Aber auch die anderen Attribute der Bildschirme nehmen an Bedeutung zu, wenn man die Nutzungssituation der Endgeräte berücksichtigt. Dadurch, dass mobile Endgeräte auch unter freiem Himmel und in Bewegung genutzt werden, können die Lichtverhältnisse und Sonneneinstrahlung erheblich variieren. Attribute der Bildschirme wie Kontrast und maximale Helligkeit besitzen somit mehr Relevanz als bei einem Bildschirm, der zu einem Desktop Computer in einem Büro gehört. Beim Testen einer mobilen Anwendung sind somit auch die mobilen Endgeräte zu berücksichtigen, deren Bildschirme nicht zu den leistungsstärksten auf den Markt verfügbaren gehören.

Je nach Anwendung kann auch die Ein- und Ausgabe durch Sprache beziehungsweise Sounds von Relevanz sein. Dadurch, dass viele mobile Endgeräte ein Mikrofon zur Spracheingabe und zum Telefonieren integriert haben, können für diese Geräte auch Anwendungen entwickelt werden, die mit Sprachsteuerung funktionieren. Die eingebauten Lautsprecher können außerdem als zusätzliches Ausgabemedium für die mobile Anwendung fungieren.

Die Energieversorgung der mobile Endgeräte ist ein weiteres Differenzierungsmerkmal zu den stationären Geräten. Während diese meist mit einer Steckdose verbunden sind und somit über nahezu unendlichen Strom verfügen, müssen die mobilen Geräte mit einem Akku auskommen. Die Qualität der Akkus entscheidet darüber, wie lang der Benutzer das Gerät verwenden kann. Je nach Anwendung wird die Hardware der Endgeräte mehr oder weniger beansprucht und führt zu entsprechendem Ladungsverlust des Akkus.

Allgemein lässt sich festhalten, dass bei den Usability-Tests nur die Endgeräte berücksichtigt werden müssen, die mit der Anwendung auch angesprochen werden sollen. In der Regel ist es jedoch so, dass Entwickler, die ihre Anwendung kommerziell vertreiben, möglichst viele Endgeräte ansprechen wollen, um sich einen großen potenziellen Kundenkreis zu wahren. Dieses Problem macht der Journalist Ben Charny in seinem Artikel „Write once, run anywhere not working for phones“ deutlich:

„[...] hardware makers were busy producing cell phones that were like snowflakes:

No two were alike. Some had huge screens and tiny dial pads, others just the opposite. Application makers have had to account for the nuances or risk severely limiting the reach of their products.“ [Cha11]

Seit dem Erscheinen dieses Artikels im Jahr 2005 hat sich an dieser Tatsache nicht viel geändert. Aus den bisherigen Ausführungen könnte jedoch die Vermutung angestellt werden, dass es ausreichend ist, jeweils ein Endgerät mit der schlechtesten Performance, der niedrigsten Auflösung und dem kleinsten Bildschirm von einem Typ der möglichen Eingabemedien zu wählen und diese als Grundlage für die Tests zu verwenden. Es liegt jedoch auch im Sinne der Entwickler, das Nutzungserlebnis auf allen Endgeräten zu optimieren. Die Betriebssysteme der mobilen Geräte unterstützen die Entwickler dabei, indem sie die Möglichkeit bieten, das Layout entsprechend der Pixeldichte und Bildschirmgröße anzupassen [Dev11, Bla11, App11]. Die APIs der Betriebssysteme unterstützen auf diese Weise allerdings lediglich die Programmierung für differierende Hardwarekomponenten. Den Nutzungskontext und die Umgebungsvariablen, also Wetter, Lichtverhältnisse, Lautstärke, Position des Gerätes und ähnliches, zu berücksichtigen, bleibt vollständig den Entwicklern überlassen. Im Folgenden werden die Besonderheiten der Nutzungskontexte und das Nutzungsverhalten der Endgeräte und ihrer Anwender näher beschrieben.

Nutzungskontext und Nutzungsverhalten

Der Nutzungskontext mobiler Anwendungen lässt sich in zwei große Kontextgruppen unterteilen. Zum einen die Variablen, die sich direkt auf die Person, die die Anwendung verwendet, beziehen. Zum anderen die Variablen, die die Person von außen beeinflussen, die Umgebungsvariablen bzw. äußeren Umstände. Der persönliche Kontext kann dabei noch in emotionale Zustände und physische Zustände¹³ gegliedert werden. Emotionale Zustände können beispielsweise freudige Erregung oder Trauer sein. Während physische Zustände unter anderem gehen oder stehenbleiben sind [KKL⁺02]. Diese emotionalen Zustände sind für die Usability-Testmethoden in zweierlei Hinsicht interessant. Ein Grund dafür wurde in Abschnitt 2.1.5 genannt: Aus der Interaktion mit der Anwendung kann sich ein neuer emotionaler Zustand ergeben und der persönliche Kontext wird somit direkt durch die Benutzung beeinflusst. Für das Usability-Testing ist jedoch auch interessant, wie der emotionale Zustand der Person ist, bevor er mit der Applikation interagiert. So kann bei Testmethoden, die die Resonanz der Testnutzer für die Bewertung der Usability einer Anwendung verwenden, untersucht werden, ob der emotionale Zustand der Testperson sein Empfinden der Usability beeinflusst hat.

¹³Unter physischen Zuständen wird im Verlauf dieser Arbeit die Bewegung und Haltung eines Menschen und nicht der Aggregatzustand zusammengefasst. Demnach gehören zum Beispiel rennen, gehen, stehen, sitzen und liegen zu möglichen physischen Zuständen.

Aber auch bei den physischen Zuständen der Endbenutzer kann es wertvoll sein, einen Vergleich vor und während der Benutzung der Anwendung anstellen zu können. Befindet sich die Testperson zum Beispiel gerade aktiv¹⁴ in Bewegung und muss dann für eine bestimmte Aufgabe, die er während des Tests mit der Anwendung erfüllen muss, stehenbleiben, so kann dies eine wichtige Information für die Bewertung der Usability sein. Ähnlich verhält es sich mit den Händen und Augen der Testnutzer. Für eine Auswertung eines Usabilitytests kann es ebenso interessant sein, ob die Person beide Hände zur Verfügung hat und ob er diese auch benutzt hat bzw. ob die Augen des Testnutzers die ganze Zeit auf dem Gerät verweilen oder er seinen Blick während des Erledigens einer Aufgabe vom Gerät abwendete. All diese Informationen geben Informationen darüber, wie gut eine mobile Anwendung in den verschiedenen Nutzungskontexten benutzbar sind und wie viel Aufmerksamkeit auf die Benutzung der Anwendung gelegt werden muss. Dies ist je nach Applikation mehr oder weniger kritisch einzuschätzen. Eine Navigationsanwendung für ein Auto sollte dementsprechend möglichst wenig kognitive Kapazität zur Bedienung verbrauchen, während es bei Anwendungen, die der Unterhaltung dienen, wie zum Beispiel Spiele, vermutlich weniger relevant ist. Dies ist damit zu begründen, dass Spiele, um dem Beispiel zu folgen, in der Regel eingesetzt werden, wenn der Nutzer wenig Aufmerksamkeit auf andere Dinge legen muss und sich ablenken möchte.

Die zweite große Gruppe des Nutzungskontextes, die Umwelt des Nutzers, ist sehr vielseitig und kann verschieden klassifiziert werden. Kim et al. [KKL⁺02] nehmen die Unterteilung des Umgebungskontextes in physischen und sozialen Kontext vor. Der physische Kontext ist hierbei bestehend aus visuellen und auditiven Elementen, die den Benutzer der Anwendung beeinflussen. Unter sozialem Kontext wird verstanden, wie viele Menschen um die Testperson herum sind und mit wem während der Benutzung interagiert wird [KKL⁺02]. Für das Usability-Testing sind jedoch noch weitere Variablen der Umgebung der Nutzer relevant. Lee und Grice [LG04] und Gorlenko und Merrick [GM03] nennen mit den Wetterbedingungen und den Lichtverhältnissen weitere interessante zu berücksichtigende Faktoren. Die Wetterbedingungen sind zum einen deshalb interessant, weil sie oft die Lichtverhältnisse durch Sonneneinstrahlung beeinflussen, zum anderen, weil die Temperatur Einfluss auf die Usability eines Endgerätes haben kann. Während man bei stationären Systemen wie Arbeitsplatzcomputern noch von etwa gleichbleibender Zimmertemperatur ausgehen kann, wodurch die Usability nicht beeinflusst wird, können die Temperaturen, bei der eine mobile Anwendung verwendet wird, erheblich variieren. Stellen Sie sich zum Beispiel eine Applikation vor, die über die Abfahrtszeiten von Bussen informiert. Jemand, der die Applikation im Sommer benutzt, um unterwegs schnell herauszufinden, ob er rennen muss, damit er den nächsten Bus noch erreicht, nimmt die Usability dieser Anwendung anders wahr als jemand, der bei Minusgraden im Winter mit Handschuhen versucht dieselbe Information zu erlangen.

¹⁴Aktiv meint hier, dass die Person Konzentration dafür aufbringen muss, sich zu bewegen. Ein Fußgänger oder Autofahrer hat in diesem Sinne weniger kognitive Kapazität für die Bedienung einer mobilen Anwendung als eine Person, der sich in einem Zug oder Bus befindet, obwohl sich all diese Menschen in Bewegung befinden.

Abbildung 2.8 aus „Context as a Necessity in Mobile Applications“ von Eleni Christopoulou zeigt den Nutzungskontext gegliedert in fünf kleinere Bereiche, Benutzer (User), Ort (Place), Zeit (Time), Endgerät (Artefact) und physische Umgebung (Physical Environment). Die Abbildung dient der Übersicht über alle zu berücksichtigenden Faktoren des Nutzungskontextes. Einige mobile Endgeräte verfügen bereits über verschiedene Sensoren zur Wahrnehmung der

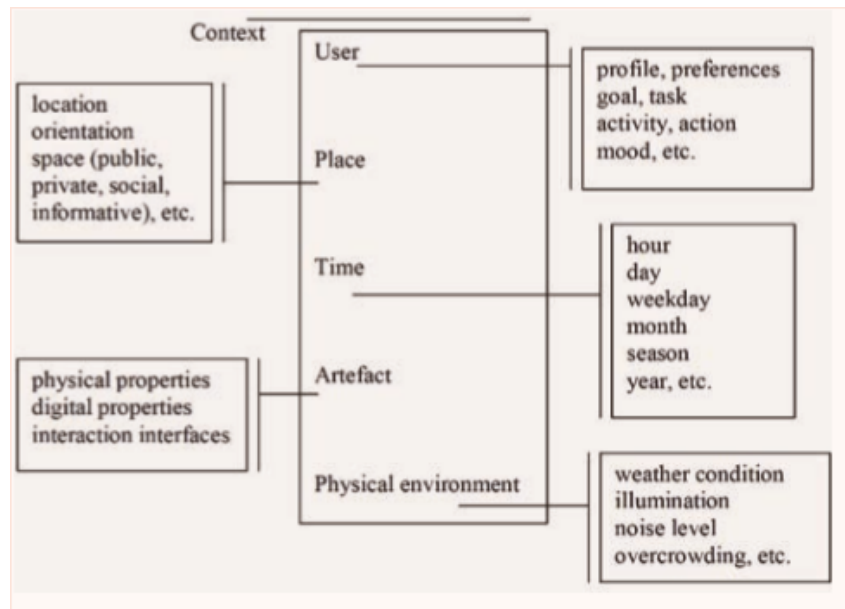


Abbildung 2.8: Der Kontext von mobilen Anwendungen im Detail [Chr08].

Umgebung, wie zum Beispiel einem Lichtsensor. Mit Hilfe dieses Sensors kann die Helligkeit des Bildschirms angepasst werden, um bestmögliche Lesbarkeit der Elemente auf dem Bildschirm gewährleisten zu können. Diese Wahrnehmung der äußeren Umstände wird Contextual Awareness genannt. Unter diesem Begriff werden die Location Awareness (Wo befindet sich das Gerät?), Environmental Awareness (z.B.: Wie laut ist die Umgebung? Befindet sich das Gerät in einem geschlossenen Raum?), Mobility Awareness (Bewegt sich der Nutzer? Wie bewegt er sich?), Health Awareness (Wie geht es dem Benutzer? Wie ist sein Blutdruck / seine Herzfrequenz / seine Körpertemperatur?) und Activity Awareness (Was tut der Benutzer gerade?) zusammengefasst [GM03]. Vielseitig ausgestattete Endgeräte besitzen verschiedene Sensoren, um möglichst große Contextual Awareness zu bieten. Weitere häufig eingebaute Sensoren sind das Global Positioning System (GPS), der Beschleunigungssensor (Accelerometer)¹⁵, das Magnetometer und der Näherungssensor. Diese Sensoren werden üblicherweise bei Bedarf von den mobilen Anwendungen angesprochen. Zum Zeitpunkt des Verfassens dieser Masterarbeit sind keine Usability-Testmethoden veröffentlicht worden, die die Sensoren der Endgeräte zur Gewinnung relevanter Daten verwenden.

¹⁵Der Beschleunigungssensor wird auch 3-Achsen-Beschleunigungssensor genannt. Er dient dazu herauszufinden, in welcher Lage sich das mobile Endgerät befindet.

Kim et al. fanden in ihrer empirischen Studie des Nutzungskontextes und der Usability Probleme im mobilen Internet weitere interessante Zusammenhänge heraus. Da die Studie jedoch im Jahr 2002 veröffentlicht wurde, ist die Aussagekraft für heutige Verhältnisse begrenzt. Für das Thema dieser Arbeit ist besonders hervorzuheben, dass die Testpersonen der Studie abhängig von ihrem Nutzungskontext verschiedene Usabilityprobleme aufdeckten und sich die Nutzung des mobilen Internets auf verschiedene Schlüsselsituationen beschränkt haben [KKL⁺02]. Da sich die Studie ausschließlich mit dem mobilen Internet befasste, repräsentieren diese Schlüsselsituationen allerdings nicht die Nutzungskontexte aller mobilen Anwendungen zum jetzigen Zeitpunkt. Hinzukommt, dass die technischen Möglichkeiten in den letzten neun Jahren fortgeschritten sind. Zum Beispiel sind die Übertragungsraten gestiegen, was zu weniger Wartezeiten bei der Benutzung führt. Daraus folgt, dass die Anwendungen auch benutzt werden können, wenn dem Benutzer weniger Zeit zur Verfügung steht. Die Kernaussage, dass verschiedene Nutzungskontexte verschiedene Usability-Probleme aufdecken, bleibt jedoch bestehen.

Die TOMORROW FOCUS Media GmbH veröffentlichte im Januar 2011 eine ähnliche Studie, in der unter anderem gemessen wurde, wo das mobile Internet vorwiegend genutzt wird. Die Ergebnisse dieser Messung sind in Abbildung 2.9 dargestellt. Da das mobile Internet eine spezielle mobile Anwendung ist, nämlich der Browser des Endgerätes, lassen sich die Ergebnisse der Studie nicht anwenden, um die Nutzungsorte aller mobilen Anwendungen statistisch darzustellen. Sie dienen jedoch als aktuelles Beispiel für eine Klasse von Anwendungen und deren charakteristische Nutzungssituation. Bei anderen Anwendungen gilt es dementsprechend typische Nutzungsorte zu ermitteln, damit die Usability gezielt für diese Situationen verbessert werden kann. Die Nutzungsdauer der Applikationen variiert mit dem Grund, warum ein Benutzer die Anwendung aufgerufen hat. Während zum Beispiel ein Preisvergleich in einem Supermarkt im Normalfall nur weniger Minuten dauert, kann das Lesen von mehreren Artikeln mit der Anwendung einer Zeitschrift eine unbestimmte Zeit dauern. Diese auf rationalem Denken beruhende Vermutung wird durch eine Studie von Falaki et al. gestützt. Sie untersuchten das Nutzungsverhalten von 255 Smartphonennutzern. Zu der Beziehung zwischen Nutzungsdauer und Anwendungstyp fanden sie folgendes heraus:

„Different application types have different session lengths, [...],for the categories defined earlier. Interactions with maps and games tend to be the longest and those related to productivity and system tend to be the shortest.“ [FMK⁺10]

Zu den Ergebnissen ihrer Arbeit gehörte außerdem, dass Mittelwerte über Nutzungsdauer oder Nutzungshäufigkeit und ähnlichen Werten keine große Bedeutung beigemessen werden sollte, da sie zu abhängig von Applikation und Nutzer sind. Für die Entwicklung mobiler Anwendungen bedeutet dies, dass eine möglichst große Spanne sinnvoller Nutzungszeiten getestet werden sollte.

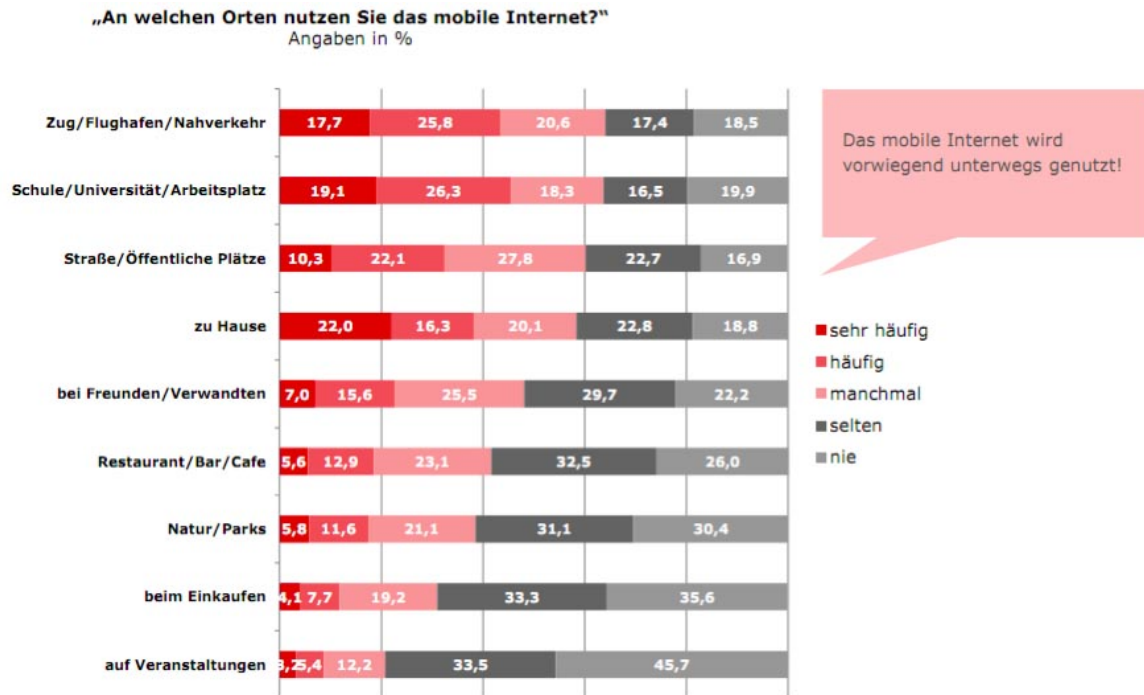


Abbildung 2.9: Nutzungsorte des mobilen Internets [Rot11].

Hilfsmittel zur Aufzeichnung von Kontext und Verhalten während des Usability-Tests

Um den Nutzungskontext und die Testperson während des Verwenden der mobilen Applikation aufnehmen zu können, gibt es verschiedene Hilfsmittel, die im Folgenden vorgestellt werden. Eine der Methoden zum Aufnehmen der Aktionen auf dem Endgerät ist, das mobile Gerät auf einem Stativ zu befestigen und eine Kamera darauf zu richten. Mit Hilfe der Kamera wird demnach der Bildschirm und die Bedienung des Endgerätes aufgezeichnet. Da das Gerät durch das Stativ unbeweglich wird, geht der mobile Charakter verloren. Die Kamera kann demnach nur im Laborumfeld benutzt werden und nicht bei einem Usability-Test in freier Umgebung. Das Stativ ist notwendig, da das Endgerät vom Benutzer sonst möglicherweise außerhalb des Aufnahmebereiches der Kamera bedient oder durch schnelle Bewegungen das Bild unscharf und somit unbrauchbar wird. Eine Alternative zu dem Stativ ist ein „Tilted Pad“, ein schräg angebrachtes Brett, was die Aufnahmefläche der Kamera kennzeichnet. Ein solches „Tilted Pad“ ist in Abbildung 2.10 zu sehen. Dadurch, dass die Kamera hinter der Testperson in einem schrägen Winkel aufnimmt, ist es dem Benutzer des Endgerätes möglich, die mobile Anwendung in einer natürlicheren Position zu bedienen, als es der Fall wäre, wenn die Kamera von oben aufnehmen würde und der Benutzer das Endgerät in einer waagerechten Position halten müsste [Sch10, Kra]. Eine andere Möglichkeit ist es, die Kamera direkt auf dem mobilen Endgerät zu befestigen. So ist es weiterhin im begrenzten Rahmen möglich, das Endgerät zu bewegen. Ein Beispiel für eine solche Vorrichtung wird in Abbildung 2.11 gezeigt.

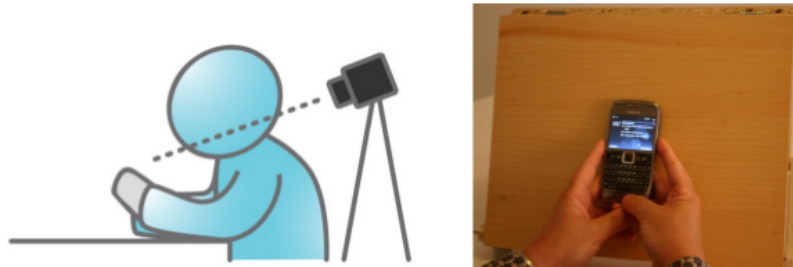


Abbildung 2.10: Video Kamera mit Tilt Pad [Sch10].



Abbildung 2.11: Video Kamera direkt auf dem mobilen Endgerät befestigt [SWL07].

Durch die Vorrichtung treten jedoch mehrere Probleme auf. Wird die Vorrichtung in Bewegung genutzt, können durch die sich ändernden Lichtverhältnisse Spiegelungen auftreten, so dass die Aufnahmen der Kamera unbrauchbar werden können. Außerdem ist der Blick auf das Display durch Kamera und Halterung in bestimmten Blickwinkeln unmöglich, was die Natürlichkeit der Anwendung ebenso verfälscht wie die Tatsache, dass das Endgerät an Gewicht und Größe zunimmt und zum Beispiel nicht mehr einfach in die Tasche gesteckt werden kann. Da die Vorrichtung deutlich sichtbar ist und Aufmerksamkeit auf die Testperson zieht, kann es bei der Verwendung in der Öffentlichkeit zu psychologischen Auswirkungen auf diese Person kommen [Kra].

Zur Wahrnehmung der Umwelt des Testnutzers kann eine Kamera direkt an der Testperson befestigt werden. Diese muss dann zum Beispiel einen Helm oder einen Rucksack mit einer entsprechenden Vorrichtung tragen. Hierbei ist es ebenfalls möglich, dass die Aufnahmen durch die Bewegungen und daraus entstehende Unschärfe oder Lichtverhältnisse, wie direkte Sonneneinstrahlung oder Dunkelheit, unbrauchbar werden [Kra]. Eine Kamera, die ebenfalls am Endgerät befestigt ist, aber auf den Benutzer anstatt auf das Endgerät ausgerichtet ist, ist eine Möglichkeit, den Benutzer direkt zu filmen, um seine Reaktionen analysieren zu können. Ein anderer Ansatz ist es, die Kamera nicht zu befestigen, sondern den Test von einem Kameramann filmen zu lassen. Roto et al. [ROH⁺04] entwickelten eine Methode, die mehrere von den genannten Kameras einsetzt. Bei ihrem Ansatz werden vier Kameras verwendet, um Nutzer und Kontext aufzuzeichnen. Eine Kamera ist auf der Schulter des Testnutzers installiert und erfasst den Kontext vor der Person, ein Kameramann filmt den Testnutzer im Kontext und zwei weitere Kameras sind auf dem Endgerät installiert. Sie erfassen das Gesicht der Testperson und das Display des Endgerätes. Eine Foto ihres Versuchsaufbaus ist in Abbildung 2.12 zu sehen. Der Test mit Hilfe dieser Apparaturen hat jedoch einige Nachteile. Zum einen ist der Testnutzer



Abbildung 2.12: Usability-Test mit vier Miniaturkameras [ROH⁺04].

in seiner Bewegungsfreiheit eingeschränkt, was die Natürlichkeit seines Verhaltens beeinträchtigt.

tigen kann. Zum anderen ist der Aufwand zur Vorbereitung des Tests relativ hoch. Außerdem könnte auch hier die Testperson durch die erhöhte Aufmerksamkeit, die ihr vermutlich durch die auffällige Apparatur von der Außenwelt zugeteilt wird, psychisch beeinflusst werden [Kra, S. 152 ff.].

Beim Usability-Testing von Anwendungen für Desktop Computer ist das Eye-Tracking mit Hilfe spezieller Kameras eine beliebte Methode, um herauszufinden, auf welche Bereiche des Bildschirms der Blick der Testperson fokussiert ist. Die auf das Eye-Tracking spezialisierte schwedische Firma *Tobii* hat ein Verfahren entwickelt, mit dem diese Methode auch auf mobilen Endgeräten möglich ist. Die mobilen Endgeräte müssen dazu auf einem von *Tobii* entworfenen Pult fixiert werden, während eine spezielle Kamera die Augen der Testperson aufnimmt. Eine Analysesoftware bestimmt mit Hilfe der Daten von der Kamera, auf welchen Bereich des Gerätes der Blick der Testperson zu einem bestimmten Zeitpunkt gerichtet war [tob11]. Ein großer Nachteil dieser Technologie liegt darin, dass das mobile Endgerät nicht von dem Pult gelöst werden kann und somit nicht natürlich zu bedienen ist. Das Verfahren kann lediglich an einem Ort stattfinden und die Benutzung während einer Feldstudie wird somit unmöglich.

Kapitel 3

Methodik zur Evaluation der Usability-Testmethoden

In diesem Kapitel werden zunächst die Kriterien vorgestellt, die zur Bewertung der Usability-Testmethoden relevant sind. Diese Kriterien ergeben sich aus den zuvor vorgestellten Eigenschaften der Usability, des Usability-Engineering und den Charakteristika der mobilen Anwendungen und deren Nutzungskontext. Für diese Kriterien werden jeweils Bewertungsskalen aufgestellt, die die Ergebnisse der Evaluation einer Usability-Testmethode mit der einer anderen vergleichbar machen. Im Anschluss an die Vorstellung der Bewertungskriterien wird eine Methodik erklärt, die es erlaubt, eine Gesamtbewertung für eine der analysierten Methoden zu errechnen und dabei Schwerpunkte auf bestimmte Kriterien zu legen.

3.1 Bewertungskriterien zur Evaluation der Usability-Testmethoden

In diesem Abschnitt werden unter Berücksichtigung der zuvor beschriebenen Charakteristika der Usability und der Besonderheiten bei der Nutzung von mobilen Endgeräten Kriterien aufgestellt, nach denen die Usability-Testmethoden bewertet werden können. Diese Kriterien sind unterteilt in solche, die für alle Usability-Tests gelten können und nicht abhängig von den Besonderheiten der mobilen Anwendungen sind, sowie in Kriterien, die die beschriebenen Besonderheiten berücksichtigen. Die Evaluation der Kriterien findet auf einer Skala von null bis fünf statt.

3.1.1 Geräteunabhängige Kriterien

Die geräteunabhängigen Kriterien sind die Kriterien, die sich aus den Eigenschaften des Usability-Engineerings und -Testings für Endgeräte und Anwendungen aller Art ergeben. Hinzu kommen Bewertungskriterien, die wirtschaftliche Aspekte bei der Evaluation der Methoden in Betracht ziehen.

Zeitaufwand

Ein geringer Zeitaufwand ist, insbesondere unter Berücksichtigung der geringeren Kosten, von Vorteil, sofern die Qualität darunter nicht leidet. Daraus folgt, dass Entwicklungszyklen nicht unnötig in die Länge gezogen werden. Besonders interessant sind kurze Zeitperioden für das Testen der Anwendung unter der Berücksichtigung von Nielsens Rat, mehrere Usability-Tests innerhalb eines Entwicklungszyklus durchzuführen (vgl. Abschnitt 2.1.4). Je kürzer die Tests

sind, desto häufiger können sie letztendlich durchgeführt werden und desto mehr Usabilityprobleme werden aufgedeckt. Die Zeit ist jedoch ein Faktor, der niemals ohne Berücksichtigung der Auswirkungen auf die anderen Faktoren optimiert werden sollte. So leidet vermutlich in den meisten Fällen die Qualität der Testergebnisse, wenn man den Testnutzer weniger Aufgaben erledigen lässt, um dadurch Zeit zu sparen.

Die Akquise der geeigneten Testpersonen und die Analyse der Testergebnisse nach dem eigentlichen Testdurchlauf werden bei diesem Kriterium nicht berücksichtigt. Zum einen bleibt dies unberücksichtigt, weil die Anwerbung geeigneter Testnutzer bereits während der Entwicklung der Anwendung geschehen kann und somit keine direkte Verzögerung des Tests bedeutet, zum anderen, weil bis zum Zeitpunkt des Schreibens dieser Arbeit keine Statistiken über die Dauer der Analyse von Protokollen und Aufzeichnungen veröffentlicht wurden. Eine Schätzung dieses Wertes wäre vermutlich sehr ungenau und würde die Ergebnisse verfälschen. Der Zeitaufwand setzt sich demnach aus der Dauer für den eigentlichen Testdurchlauf zuzüglich der Dauer für mögliche Briefings mit Testteilnehmern zusammen.

Tabelle 3.1 beschreibt, welche Zeiträume zu welcher Bewertung in der Evaluation der Usability-Testmethoden führen.

Punkte	Bedeutung
5	weniger als 2 Stunden
4	2-4 Stunden
3	4-6 Stunden
2	6-8 Stunden
1	1-2 Tage ¹
0	länger als zwei Tage

¹ Ein Tag entspricht einem Arbeitstag von acht Stunden.

Tabelle 3.1: Bewertungstabelle für den Zeitaufwand

Kosten

Für die Kosten gilt ähnliches wie für den benötigten Zeitraum. Ein Unternehmen wird sich nicht über geringere Kosten beschweren. Geringere Kosten können jedoch zum Beispiel einen Qualitätsverlust der Testergebnisse oder eine Verlängerung der benötigten Zeit bedeuten, wenn Einsparungen beim Personal vorgenommen werden. Die Kosten der Methoden ergeben sich aus mehreren Faktoren. Dazu gehören vor allem die benötigten technischen Mittel, die Vergütung der Testpersonen und die Bezahlung der Fachkräfte. Als typische Kostenpunkte für das Usability-Testing nennt Nielsen [Nie94c, S. 171 ff.]:

- Usability-Spezialisten zur Planung, Durchführung und Analyse des Tests,

- Assistenten für administrative Aufgaben, wie Terminvereinbarungen mit Testnutzern und Dateneintragungen,
- Softwareentwickler, um Code zur Datensammlung oder andere Anpassungen an der Software vornehmen zu können,
- Zeitaufwandsentschädigungen für die Testnutzer,
- Computer und Hardware, die während des Testens benutzt wird,
- die Räumlichkeiten für die Tests sowie
- Kosten für Aufnahmemedien und weitere Verbrauchsmaterialien.

Diese Kostenpunkte treffen nicht alle für jede Usability-Testmethode zu, stellen jedoch einen Anhaltspunkt für die zu berücksichtigenden Kosten dar. Die Kosten für die Usability-Testmethoden können in jedem Fall nur eine Schätzung sein, da Dauer und zur Verfügung stehende Kapazitäten (Räume, Personal und Hardware) je nach zu testender Anwendung, Erfahrung der Experten/Moderatoren und Unternehmen erheblich variieren. Zur Vereinfachung werden deshalb die Gesamtkosten lediglich unter Berücksichtigung der Kosten für Experten, Testnutzer und *spezieller Hardware* ermittelt. Unter spezieller Hardware werden jegliche Kameras, Sensoren und besondere Ein- oder Ausgabegeräte zusammengefasst. Desktop Computer und mobile Endgeräte werden dabei nicht berücksichtigt, da diese im Normalfall vorhanden sind und für verschiedene Zwecke wiederverwendet werden können. Die Kosten für einen Usability-Experten ergeben sich aus den jährlichen Gehältern der Usability-Experten, zusammengetragen in „Salary Trends for Usability Professionals“ [Nie11]. Dabei wird der Mittelwert der Gehälter von erfahrenen Experten (\$ 84000) und Berufseinsteigern (\$ 60000) durch 1600 Stunden (200 Arbeitstage mal 8 Stunden pro Arbeitsag) dividiert. So ergibt sich ein Stundenlohn von \$ 45 pro Stunde oder 32,69€¹⁶. Es ist damit zu rechnen, dass weitere Zeit des Experten in Anspruch genommen werden muss, wenn bei der Usability-Testmethode Aufzeichnungen stattgefunden haben, die anschließend analysiert werden müssen. Da über die Dauer dieser Analyse keine ausreichenden Informationen verfügbar sind (vgl. Abschnitt 3.1.1), können diese Kosten dafür hier nicht berücksichtigt werden. Des Weiteren wird mit durchschnittlichen Kosten von \$ 171 beziehungsweise 123,4€ pro Testnutzer gerechnet [Nie03a].

Tabelle 3.2 zeigt, welche Kostenwerte in der Evaluation der Methoden zu welcher Bewertung führen. Die Bewertung wird mit geschätzten Grenzen zur Vergabe der Punkte vorgenommen, damit es möglich ist, weitere Usability-Testmethoden, die in dieser Arbeit keine Berücksichtigung finden, auf die gleiche Art evaluieren zu können. Es wurde also bewusst darauf verzichtet, die Bewertungen relativ zueinander vorzunehmen. Da jedoch bei allen Methoden die gleichen

¹⁶Entspricht dem Wechselkurs vom 23.10.2011 (Quelle: Google Finance).

Annahmen getroffen werden, genügen die Schätzungen, um eine akzeptable Bewertung zum Vergleich der Kosten vornehmen zu können. Die tatsächlichen Kosten, die aus genannten Gründen nicht ermittelt werden können, sind jedoch wesentlich höher einzuschätzen.

Punkte	Kosten
5	0€ bis 1000€
4	1.001€ bis 2.000€
3	2.001€ bis 3.000€
2	3.001€ bis 4.000€
1	4.001€ bis 5.000€
0	über 5.000€

Tabelle 3.2: Bewertungstabelle für den Kostenaufwand

Zeitpunkt der Durchführung

Nach den Erkenntnissen aus Abschnitt 2.1.6 ist ein früher möglicher Einsatzzeitpunkt von vorhandenen Normen und Standards zur Benutzerinteraktion empfehlenswert. Werden Usabilityprobleme früh erkannt, können sie meist schneller und einfacher behoben werden. In der Bewertung wird ein früher möglicher Zeitpunkt zum Einsatz der Usability-Testmethode demzufolge besser bewertet als ein späterer. Dabei wird von einem einfachen Softwareentwicklungsmodell ausgegangen¹⁷, dessen Phasen aus Analyse, Entwurf, Implementierung, Test und Wartung bestehen. Ein Test, der während der Entwurfsphase stattfindet, benötigt dementsprechend keine Zeile Programmiercode zum Testen der Usability. Während der Implementierungsphase können zwei verschiedene Testarten realisiert werden. Zum einen können solche Tests durchgeführt werden, die keine Prototypen auf mobilen Endgeräten benötigen. Diese Tests sind beispielsweise mit Hilfe von Emulatoren auf Desktop Computern umsetzbar. Die Emulatoren simulieren dabei ein mobiles Endgerät auf dem die Anwendung ausführbar ist. Zum anderen gibt es Tests, die einen Prototypen auf einem mobilen Endgerät testen. Den Prototyp zeichnet dabei aus, dass nur einige Funktionen der mobilen Anwendung funktionieren. In der Testphase sind bereits alle Funktionen implementiert. Sie dient in der Anwendungsentwicklung dazu, vor der Veröffentlichung zu behebbende Fehler zu finden. In der Wartungsphase ist die Anwendung bereits auf dem Markt/im Einsatz und noch vorhandene Fehler werden nachträglich durch Updates behoben. In Tabelle 3.3 werden die Einsatzzeitpunkte der Usability-Testmethoden den Punkten für die Bewertung zugeordnet.

¹⁷Die Entwicklungsphasen entsprechen denen des Wasserfall- und Spiralmodells.

Punkte	Zeitpunkt der Durchführung
5	Der Usability-Test kann zu jedem Zeitpunkt durchgeführt werden.
4	Der Usability-Test findet in der Entwurfsphase statt.
3	Der Usability-Test findet mit Hilfe von Emulatoren statt.
2	Der Usability-Test wird auf einem Prototypen durchgeführt.
1	Der Usability-Test findet während der Testphase der Entwicklung statt.
0	Der Test kann erst nach Veröffentlichung der Anwendung durchgeführt werden.

Tabelle 3.3: Bewertungstabelle für den Zeitpunkt der Durchführung

Berücksichtigung der Usability-Attribute

Für die Bewertung der Usability einer Anwendung sind die Usability-Attribute (vgl. Abschnitt 2.1.2) essentiell. Es ist darum wichtig, dass die Usability-Testmethoden möglichst viele oder am besten alle diese Attribute in ihrem Test berücksichtigen. Nur so kann gewährleistet werden, dass alle Aspekte der Usability einer Anwendung untersucht werden können. Dies kann auf den ersten Blick als selbstverständlich erscheinen. Wird jedoch das Attribut *Einprägsamkeit* betrachtet, bei der der Nutzer die Anwendung einige Zeit nicht nutzen darf, ist schnell ersichtlich, dass eine Berücksichtigung aller Attribute sehr komplex und zeitaufwändig werden kann. Die Messung dieser Attribute kann auf verschiedene Art und Weise durchgeführt werden und wird in Abschnitt 2.1.5 näher erläutert. Bei Usability-Testmethoden, die sowohl das Testen mit Zeichnungen oder Mock-Ups¹⁸ als auch das Testen mit teilweise oder vollständig funktionierenden Anwendungen unterstützen, wird die Berücksichtigung der Usability-Attribute unter Verwendung einer funktionierenden Anwendung bewertet. Die folgende Tabelle 3.4 veranschaulicht das Bewertungsprinzip dieses Kriteriums für die Evaluation der Usability-Testmethoden.

Punkte	untersuchte Attribute
5	Alle Usability-Attribute werden untersucht.
4	Vier Usability-Attribute werden untersucht.
3	Drei Usability-Attribute werden untersucht.
2	Zwei Usability-Attribute werden untersucht.
1	Ein Usability-Attribute wird untersucht.
0	Keines der Usability-Attribute wird untersucht.

Tabelle 3.4: Bewertungstabelle über die Berücksichtigung der Usability-Attribute

¹⁸Definition: „Eine Attrappe der Benutzerschnittstelle, die zur Evaluation mit Benutzern und anderen Interessensvertretern verwendet werden kann. [...]“ [kom10]

Berücksichtigung des emotionalen Zustands

Der emotionale Zustand ist ein Teil des persönlichen Nutzungskontextes, der in Abschnitt 2.2.2 beschrieben wurde. Da die Emotionen sowohl bei stationären als auch bei mobilen Endgeräten für die Bewertung der Usability relevant sind, ist das Kriterium den geräteunabhängigen Kriterien zugeordnet. Der emotionale Zustand ähnelt dem Usability-Attribut *Satisfaction / Zufriedenheit*. Allerdings beschränkt sich dieses Usability-Attribut auf das Befinden des Nutzers während der Anwendung und damit, ob sich dieser Grad der Zufriedenheit während der Bedienung ändert. Der emotionale Zustand im persönlichen Benutzungskontext umfasst jedoch auch die Stimmung, bevor die Anwendung bedient wird, um möglicherweise Zusammenhänge zwischen der vom Nutzer empfundenen Qualität der Usability der Anwendung und seines Gemütszustands herstellen zu können. Dadurch kann das von der Testperson gegebenen Feedback besser eingeschätzt werden.

Ein weiterer Grund, weshalb der emotionale Zustand hier erneut als Kriterium aufgeführt wird, ist, dass er einen oftmals sehr wichtigen Faktor repräsentiert: Was empfindet der Benutzer der Anwendung und macht es ihm Spaß, die Anwendung zu verwenden?

Mit Hilfe des emotionalen Zustands lässt sich also feststellen, ob ein Benutzer die Anwendung benutzen *möchte*, während die anderen Usability-Attribute feststellen, ob er die Anwendung benutzen *kann*. In der Gesamtbewertung einer Usability-Testmethode sollte es also möglich sein, dieses spezielle erweiterte Usability-Attribut gesondert gewichten zu können.

Der emotionale Zustand kann zum einen durch den Dialog (schriftlich oder mündlich) mit der Testperson, entsprechend der *Self-Reported Metrics*, ermittelt werden. Zum anderen ist das nonverbale Nutzerfeedback eine geeignete Quelle, um etwas über den Gemütszustand des Menschen zu erfahren. Unter nonverbalem Feedback werden erkenn- oder messbare Veränderungen der Stimmung oder des Zustands des Testnutzers zusammengefasst. Unterstützt eine Usability-Testmethode das Betrachten und/oder Aufzeichnen des Nutzers, können die in Abschnitt 2.1.5 beschriebenen *Behaviorial and Physiological Metrics* angewandt werden, um weitere Informationen über die Usability zu gewinnen. Das Erfassen der Nutzerzustände kann dabei durch Befragungen, direktes Beobachten, Filmen sowie mit Hilfe unterschiedlicher Sensoren (z.B. für die Herzfrequenz oder Körpertemperatur) stattfinden. Tabelle 3.5 bildet ab, welcher Grad der Berücksichtigung des emotionalen Zustandes der Testperson zu welcher Punktzahl in der Evaluation führt.

Punkte	Grad der Berücksichtigung
5	sehr starke Berücksichtigung
4	starke Berücksichtigung
3	mittlere Berücksichtigung
2	schwache Berücksichtigung
1	sehr schwache Berücksichtigung
0	keine Berücksichtigung

Tabelle 3.5: Bewertungstabelle über die Berücksichtigung des emotionalen Zustands

3.1.2 Geräteabhängige Kriterien

Die geräteabhängigen Kriterien sind solche, die sich durch den mobilen Charakter der Endgeräte ergeben. Sie entstehen aus dem besonderen Nutzungskontext und den Hardwareeigenschaften der Geräte. Der Nutzungskontext wird hierbei entsprechend der Erklärung in Abschnitt 2.2.2 in zwei Themengebiete unterteilt: die äußeren Umstände, die die Testpersonen umgeben, sowie der persönlichen Kontexte beziehungsweise der physische Zustand der Tester.

Berücksichtigung des physischen Zustands

Die Usability-Testmethoden sollten die Anwendung bei möglichst allen Bewegungsarten testen, die für die Anwendung üblich sind. Dazu gehören je nach Anwendung aktive und passive Bewegungen, sich ändernde Geschwindigkeiten und sich ändernde Nutzungsumgebungen. Das alles führt zu unterschiedlicher kognitiver Belastung der Testpersonen durch den Nutzungskontext und zur Aufdeckung unterschiedlicher Usability-Probleme (vgl. Abschnitt 2.2.2).

In Tabelle 3.6 sind die Grade der Berücksichtigung des physischen Zustands im Verhältnis zu den Punkten der Evaluation gesetzt.

Punkte	Grad der Berücksichtigung
5	sehr starke Berücksichtigung
4	starke Berücksichtigung
3	mittlere Berücksichtigung
2	schwache Berücksichtigung
1	sehr schwache Berücksichtigung
0	keine Berücksichtigung

Tabelle 3.6: Bewertungstabelle über die Berücksichtigung des physischen Zustands

Berücksichtigung der natürlichen Nutzungsumgebung

Bei der Berücksichtigung der natürlichen Nutzungsumgebung wird bewertet, wie gut die Usability-Testmethode die verschiedenen möglichen Umgebungsvariablen, in denen die Applikation genutzt werden kann, berücksichtigt.

Die Nutzungsumgebung kann auch Einflüsse auf das verbale sowie nonverbale Nutzerfeedback und den persönlichen Nutzungskontext haben. Die Testperson kann zum Beispiel durch ihre Umgebung gestresst sein, wodurch das Nutzerfeedback dementsprechend ausfiele. Ein anderes Beispiel wäre, dass der Benutzer durch eine rote Ampel zum Anhalten gezwungen würde, was den physischen Zustand und somit persönlichen Nutzungskontext beeinflusste.

Damit solche Koinzidenzen das Ergebnis des Usability-Tests nicht ungewollt beeinträchtigen, ist es wichtig, dass der persönliche Nutzungskontext und das Nutzerfeedback mit der Nutzungsumgebung in Zusammenhang gebracht werden können. Dieser Aspekt fließt ebenfalls, sofern die natürliche Nutzungsumgebung überhaupt berücksichtigt wird, in die Bewertung ein.

In Tabelle 3.7 ist die Zuordnung der Punkte der Evaluation zu dem Grad der Berücksichtigung der natürlichen Nutzungsumgebung dargestellt.

Punkte	Grad der Berücksichtigung
5	sehr starke Berücksichtigung
4	starke Berücksichtigung
3	mittlere Berücksichtigung
2	schwache Berücksichtigung
1	sehr schwache Berücksichtigung
0	keine Berücksichtigung

Tabelle 3.7: Bewertungstabelle über die Berücksichtigung der natürlichen Nutzungsumgebung

Berücksichtigung der mobilgerätespezifischen Ein- und Ausgabeprobleme

Mobile Endgeräte sind für die Heterogenität ihrer Ein- und Ausgabemedien bekannt. Sofern eine Anwendung also nicht speziell für ein einziges Gerät entwickelt wird, ist es relevant, die Anwendung auf möglichst vielen unterschiedlichen Geräten ohne überproportional wachsendem Aufwand¹⁹ testen zu können. Die folgende Tabelle 3.8 zeigt, wie die Berücksichtigung der Ein- und Ausgabeprobleme bewertet wird.

¹⁹Überproportionaler Aufwand meint hier, dass der Aufwand für einen weiteren Testdurchlauf mit einem anderen Gerät nicht erheblich mehr Zeit- und Kostenaufwand als ein weiterer Test mit demselben Gerät haben sollte.

Punkte	Grad der Berücksichtigung
5	sehr starke Berücksichtigung
4	starke Berücksichtigung
3	mittlere Berücksichtigung
2	schwache Berücksichtigung
1	sehr schwache Berücksichtigung
0	keine Berücksichtigung

Tabelle 3.8: Bewertungstabelle über die Berücksichtigung der mobilgerätespezifischen Ein-/Ausgabeprobleme

Beeinflussung des natürlichen Nutzungsverhaltens

Dieses Kriterium soll sicherstellen, dass das Testen der Usability möglichst repräsentativ ist. Je weniger der Nutzer in seiner Arbeit mit der Anwendung durch Tätigkeiten oder Gegenstände, die nur wegen des Tests bestehen, abgelenkt wird, desto natürlicher ist das Nutzungserlebnis. Zu diesen Ablenkungen können zum Beispiel Gespräche mit Betreuern des Tests, zeitlich beschränkte Nutzungsdauern oder auf den Endgeräten installierte Kameras gehören. Im Idealfall würde der Benutzer also die mobile Anwendung ohne durch den Test bedingte Beeinträchtigungen durch Gegenstände oder Personen benutzen. Wie sich diese Beeinflussung auf die Punkte der Evaluation auswirkt, wird in Tabelle 3.9 dargestellt.

Punkte	Grad der Berücksichtigung
5	keine Beeinflussung
4	sehr schwache Beeinflussung
3	schwache Beeinflussung
2	mittlere Beeinflussung
1	starke Beeinflussung
0	sehr starke Beeinflussung

Tabelle 3.9: Bewertungstabelle über die Berücksichtigung der mobilgerätespezifischen Ein-/Ausgabeprobleme

Gefundene Usability-Probleme

Für einige Usability-Testmethoden wurden Forschungen durchgeführt, in denen unter anderem getestet wurde, wie groß der Anteil der gefundenen Usability-Probleme im Vergleich zu den mit anderen Methoden gefundenen Problemen ist. Sofern solche Forschungen vorliegen, werden sie unter diesem Kriterium erwähnt. Da jedoch nicht für alle Methoden Forschungsarbeiten, die das Usability-Testing für mobile Endgeräte mit der entsprechenden Methode analysieren, durchgeführt wurden, wird dieses Kriterium nicht in der Gesamtwertung berücksichtigt, sondern dient

nur zur weiteren Information.

Die durchgeführten Studien unterscheiden sich außerdem in der Anzahl der eingesetzten Testpersonen, der Erfahrung der Usability-Experten, der untersuchten Anwendungen und der verfügbaren Mittel. Ein Vergleich anhand der Ergebnisse dieser Arbeiten wäre demnach nicht repräsentativ.

Die gefundenen Usability-Probleme fließen jedoch indirekt in die Bewertung ein. So sollte eine Usability-Testmethode, die mehr Usability-Attribute berücksichtigt, in der Regel auch mehr Usability-Probleme aufdecken. Ebenso sollte eine Methode, die das natürliche Nutzungsverhalten wenig beeinflusst und/oder die endgerätespezifischen Ein- und Ausgabemedien stärker berücksichtigt, eine größere Chance haben, Usability-Probleme zu entdecken.

3.2 Verfahren zur Gesamtbewertung einer Usability-Testmethode

Im Folgenden wird ein Gewichtungsverfahren vorgestellt, mit welchen jedes Usability-Testing Verfahren in seiner Gesamtheit bewertet werden kann. Es dient zur Vereinfachung der Vergleichbarkeit der untersuchten Methoden unter der gleichzeitigen Berücksichtigung von individuell bestimmbar Prioritäten. Das vorgestellte Verfahren zur Gewichtung der Kriterien wurde gewählt, da sich mit ihm Kriterien schnell und eindeutig priorisieren lassen und nachträgliche Änderungen zum Variieren der Ergebnisse leicht durchzuführen sind.

Ob die Notwendigkeit besteht, die verschiedenen Kriterien zu gewichten, und wie diese zu gewichten sind, ist abhängig von der zu testenden mobilen Anwendung, der wirtschaftlichen Situation der Person oder des Unternehmens, die die Anwendung entwickelt, sowie der Dringlichkeit des Tests. Aus diesem Grund ist es nicht möglich, eine generelle Gewichtung der Kriterien festzulegen. Es ist zum Beispiel denkbar, dass die Kosten der Methode für ein Unternehmen eine eher untergeordnete Rolle spielen, während ein anderes Unternehmen bereits am Limit ihres Budgets angekommen ist und so der Kostenfaktor entscheidend ist.

3.2.1 Absolute Gewichtung

Um die absolute Gewichtung durchzuführen, wird im ersten Schritt jedes Kriterium mit Hilfe eines Maßstabs gewichtet. Die Anzahl der Stufen dieses Maßstabs kann frei gewählt werden. Sie wirkt sich darauf aus, wie differenziert die Gewichtungen vorgenommen werden können. Bei einem Maßstab mit drei Stufen ist zum Beispiel nur eine Einteilung in „weniger wichtig“, „wichtig“ und „sehr wichtig“ möglich, wobei „sehr wichtig“ dann einer Gewichtung der Stufe 3 bedeuten würde. Bei einem Maßstab mit zehn Stufen ist die Gewichtung wiederum wesentlich feiner durchführbar. Als Grundregel gilt jedoch immer, dass die größte Zahl der höchsten

Priorität entspricht. Außerdem ist es möglich, eine Stufe 0 einzuführen, mit deren Hilfe Kriterien komplett aus der Gesamtbewertung fallen würden. Bei der absoluten Gewichtung muss jedoch nicht jede Stufe der Gewichtung auch tatsächlich vergeben werden und jede Stufe kann mehr als ein Mal vergeben werden [Bec78]. Tabelle 3.10 zeigt einen fünfstufigen Maßstab zur Durchführung der Gewichtungen.

Priorität	Gewichtungsstufe
sehr wichtig	4
wichtig	3
weniger wichtig	2
unwichtig	1
irrelevant	0

Tabelle 3.10: Beispiel für einen Maßstab zur Gewichtung

Nachdem für jedes Kriterium eine Gewichtung bestimmt wurde, müssen im zweiten Schritt alle vergebenen Gewichtungen addiert werden. Die Einzelgewichtungen dividiert durch die Summe der Gesamtgewichtungen ergibt einen Prozentsatz für jedes Kriterium. Anschließend wird die Summe über die Produkte der Prozentsätze und den Kriteriumsbewertungen gebildet, um eine gewichtete Gesamtbewertung zu erhalten [Bec78]. Das Ergebnis ist in diesem Fall immer eine Zahl zwischen der niedrigsten möglichen Bewertung²⁰ und der höchsten möglichen Bewertung. Bei einer Anzahl von n , den Kriterien i , den Gewichtungen g und Bewertungen b ergibt sich folgende Formel:

$$\text{Gesamtbewertung} = \sum_{i=1}^n \frac{g_i}{\sum_{j=1}^n g_j} * b_i$$

3.2.2 Anwendungsbeispiel für die absolute Gewichtung

In diesem Abschnitt wird ein kurzes Anwendungs- und Rechenbeispiel für Ermittlung der Gesamtbewertung gegeben. Es soll jedoch kein realistisches Szenario abgebildet werden. Dieser Abschnitt dient lediglich dazu, die Berechnung der Gesamtbewertung zu verdeutlichen.

Zu diesem Zweck wird von einer imaginären Usability-Testmethode ausgegangen, bei der die Kriterien bereits bewertet wurden. Außerdem wurden bereits Prioritäten hinsichtlich der Kriterien festgelegt und daraus die Gewichtung bestimmt. Für die Gewichtung wurde ein Maßstab von null bis vier gewählt. Die Bewertungen b der Kriterien der Beispielmethode, die festgelegten Gewichtungen g und das Produkt dieser Werte sind in Tabelle 3.11 dargestellt.

²⁰In dem Sonderfall, dass jedes Kriterium als irrelevant und somit mit Null gewichtet wird, ist auch eine Gesamtbewertung von null Punkten möglich.

Kriterium	<i>b</i>	<i>g</i>	<i>b * g</i>
Zeitaufwand	4	4	16
Kosten	3	1	3
Zeitpunkt der Durchführung	2	4	8
Berücksichtigung der Usability-Attribute	4	2	8
Berücksichtigung des emotionalen Zustands	5	1	5
Berücksichtigung des physischen Zustands	1	3	3
Berücksichtigung der natürlichen Nutzungsumgebung	4	4	16
Berücksichtigung der mobilgerätespezifischen Ein-/Ausgabeprobleme	3	4	12
Beeinflussung des natürlichen Nutzungsverhaltens	2	4	8

Tabelle 3.11: Beispiel für eine mögliche Zuordnung von Kriterien, Bewertungen und Gewichtungen

Die Summe der Gewichtungen ist demnach:

$$\sum_{j=1}^9 g_j = 27.$$

Daraus ergibt sich nach Einsetzen in die oben genannte Formel:

$$\text{Gesamtbewertung} = \sum_{i=1}^9 \frac{g_i * b_i}{27}$$

Für das gegebene Beispiel ergibt sich also eine Gesamtbewertung von etwa 2,926. Diese Berechnung müsste nun für alle in Frage kommenden Usability-Testmethoden durchgeführt werden. Dabei ist es für die Vergleichbarkeit der Ergebnisse relevant, dass dieselben Gewichtungen für alle Methoden gelten.

Kapitel 4

Usability-Testmethoden

In diesem Kapitel werden ausgewählte Usability-Testmethoden beschrieben und mit Hilfe der zuvor aufgestellten Kriterien und Bewertungsskalen evaluiert. Die Bewertungen ergeben sich ausschließlich aus Literaturstudien, da Zeit und Mittel dieser Masterarbeit nicht ausreichen, um die Usability-Testmethoden in der Praxis zu analysieren. Daraus folgt, dass zur Bewertung der Zeit und Kosten der einzelnen Methoden nicht ausreichend Informationen verfügbar waren. Bei diesen Kriterien müssen deshalb teilweise Schätzungen vorgenommen werden.

Die untersuchten Methoden stellen sich aus etablierten Usability-Methoden zusammen, die auch für stationäre Endgeräte verwendet werden, sowie Methoden, die speziell für mobile Endgeräte beziehungsweise mobile Anwendungen entwickelt wurden. Die Auswahl der Methoden wurde getroffen, indem jeweils zwei gebräuchliche Methoden aus den möglichen Varianten der Usability-Tests ausgewählt wurden. Der Think Aloud Test im Labor und im Feld vertreten in diesem Sinne die Variante des formative Usability-Testings unter Einbeziehung von Testpersonen ohne erweiterte Kenntnisse von Usability. Der Kognitive Walkthrough und die Heuristische Evaluation sind die am weitesten verbreiteten Methoden der Usability-Inspektionen und das Interview und der Fragebogen gehören zu der Gruppe der Usability-Befragungen.

Auf die Beschreibung und Bewertung dieser Methoden folgt ein kurzer Vergleich der Evaluationsergebnisse ohne Gewichtung der einzelnen Kriterien.

4.1 Think Aloud Test (im Labor)

Der *Think Aloud Test*²¹ wurde ursprünglich als psychologische Forschungsmethode verwendet. Mittlerweile wird das Verfahren jedoch ebenfalls für die Evaluation in Bereichen der Mensch-Computer-Interaktion verwendet [Nie94c, S. 195 ff.] und gehört in diesem Bereich zu den bekanntesten Methoden des Usability-Testing [Sch10, S. 82 ff.]. In diesem Abschnitt wird die Durchführung des Tests in einer kontrollierten, laborähnlichen Umgebung betrachtet.

4.1.1 Beschreibung

Der Think Aloud Test benötigt eine Testperson sowie einen Moderator und möglichst einen Beisitzer, der den Ablauf des Tests zur späteren Auswertung protokolliert. Der Test besteht hauptsächlich daraus, dass der Testperson eine Aufgabe gestellt wird, die sie mit der Anwendung erledigen soll. Während sie dazu die Anwendung bedient, soll sie verbalisieren, was sie

²¹Der *Think Aloud Test* wird in der Literatur auch *Thinking Aloud Test*, *Thinking Aloud Protokoll* oder *Cooperative Evaluation* genannt.

tut und warum sie es tut. Das Ziel der Methode ist es herauszufinden, wie ein potenzieller Nutzer, der die Anwendung noch nicht kennt, diese versteht. Hierbei liegt der Fokus auf der Positionierung und Bedienbarkeit der Buttons, Menüpunkte und anderer Bestandteile der Benutzeroberfläche. Dabei wird unter anderem ermittelt, welche Missverständnisse durch Abbildungen oder Beschriftungen auftauchen oder welche Menüführungen schwer nachzuvollziehen sind. Der Nutzer kann jedoch auch Kommentare über Performance oder Lesbarkeit der Anwendung machen, wodurch ebenfalls Usability-Probleme aufgedeckt werden können. Da das durchgehende Erzählen über die Elemente, die eine Person anklickt, und warum sie dies tut, für die meisten Menschen unnatürlich ist, muss der Testmoderator entsprechend reagieren. Dies kann zum Beispiel geschehen, indem er die Testperson nach den Gründen für ihr Handeln fragt. Aufgrund der geringen Größe der Displays von mobilen Endgeräten müssen der Protokollant und der Moderator dicht bei der Testperson sitzen, um sehen zu können, was auf dem Bildschirm geschieht. Durch diese Tatsache und dadurch, dass die Testperson unter andauernder Beobachtung ist, kann es dazu kommen, dass sie sich unwohl fühlt, ähnlich wie in einer Prüfungssituation [Nie94c, Sch10].

Bei der Analyse der Kommentare des Testnutzers gilt es abzuwägen, welche Anmerkungen von Relevanz sind und welche vernachlässigt werden können. Die Benutzer neigen dazu, eigene Theorien darüber aufzustellen, warum etwas problematisch und wie dieses Problem zu beheben ist. Wenn der Benutzer zum Beispiel einen bestimmten Bedienungsbereich der Anwendung nicht gesehen hat und behauptet, er würde ihn an einer bestimmten anderen Stelle besser wahrnehmen, ist es wichtiger, Aufzeichnungen darüber zu haben, was der Benutzer tat, als er den Bereich übersehen hat [Nie94c, S. 195 ff.].

Zur Dokumentation des Tests können während des Think Aloud Tests die in Abschnitt 2.2.2 beschriebenen Aufzeichnungsmethoden verwendet werden. Da es sich bei dem Think Aloud Test im Allgemeinen um einen im Labor beziehungsweise in geschlossenen Räumen stattfindenden Test handelt, können auch Kameras benutzt werden, die jedoch die Mobilität des Nutzers einschränken. Zu beachten ist dabei zudem auch, dass die natürliche Handhabung des Gerätes gegebenenfalls darunter leidet (vgl. Abschnitt 2.2.2). Durch technische Hilfsmittel, wie kleine Videokameras, ist es auch möglich, den Think Aloud Test im Feld stattfinden zu lassen. Da dies jedoch unter anderem mehr Zeit und Aufwand für Testnutzer und Testmoderator bedeutet und somit auch die Kosten steigen [KK05], wird diese Variante des Think Aloud Tests separat beschrieben und bewertet.

4.1.2 Bewertung

Nachfolgend wird der Think Aloud Test anhand der in Abschnitt 3.1 erarbeiteten Kriterien bewertet.

Zeitaufwand

Die Dauer des Think Aloud Test ist stark abhängig von den Aufgaben, die ein Testnutzer zu erfüllen hat. Eine typische Sitzungsdauer (inklusive Briefings) bei einem Usability-Test mit einem Moderator und einem Teilnehmer dauert in der Regel zwischen ein und zwei Stunden [DL08, S. 2]. Ausgehend von einer Normalverteilung dieser Testdauer²² ergibt sich also eine durchschnittliche Dauer von 90 Minuten. Nach Jakob Nielsen sind fünf Testpersonen (vgl. Abschnitt 2.1.4) für einen solchen Usability-Test ausreichend. Diese Werte führen zu einer Gesamttestdauer von 450 Minuten beziehungsweise 7,5 Stunden. Daraus folgt eine Bewertung des Zeitaufwands der Think Aloud Test-Methode mit zwei Punkten.

Kosten

Unter Berücksichtigung der in Abschnitt 3.1.1 aufgeführten Kosten für eine Testperson ergeben sich:

$$\text{Kosten für die Testnutzer} = 5 * 123,4\text{€} = 617\text{€}$$

und für 7,5 Stunden Arbeitszeit eines Usability-Experten:

$$\text{Kosten für den Usability-Experten} = 7,5h * 32,69\text{€/h} = 245,175\text{€}.$$

Aus der Beschreibung der Methode ist zu entnehmen, dass eine Vielzahl von Aufzeichnungsmöglichkeiten besteht. Die Kosten für entsprechende Hardware sind demnach nicht genau zu bestimmen. Zur Kalkulation der Gesamtkosten wird aus diesem Grund mit der Anschaffung einer hochauflösenden Kamera und entsprechender Halterung für einen Gesamtwert von 500€ gerechnet. So belaufen sich die geschätzten Gesamtkosten der Think Aloud Methode auf circa 1360€, was einer Bewertung von vier Punkten entspricht. Es ist jedoch zu beachten, dass weitere Kosten durch die Analyse der Aufzeichnungen entstehen.

Zeitpunkt der Durchführung

Die Think Aloud Methode ist auf jeder Entwicklungsstufe praktisch durchführbar. Es ist in diesem Sinne durchaus möglich, einer Testperson eine Skizze oder ein Mock-Up vorzulegen und ihn darum zu bitten, zu erklären, was er erwartet, wenn er einen bestimmten Button drückt oder einen ausgewählten Teil des Bildschirms berührt. Abbildung 4.1 zeigt ein Beispiel für ein solches Mock-Up. Der Usability-Test mit Hilfe der Think Aloud Methode kann jedoch

²²Über die tatsächliche Art der Verteilung sind keine Informationen verfügbar.



Abbildung 4.1: Beispiel eines Mock-Ups für mobile Anwendungen [Hur09]

auch mit Hilfe eines Emulators auf einem stationären Computer, mit einem Prototypen der Anwendung oder aber mit der fertigen Anwendung durchgeführt werden, woraus fünf Punkte in der Bewertung dieses Kriteriums folgen.

Berücksichtigung der Usability-Attribute

Im Folgenden wird erläutert, ob und wie die Think Aloud Methode das Testen der in Abschnitt 2.1.2 vorgestellten Usability-Attribute unterstützt.

Learnability / Erlernbarkeit:

Da der Kern der Think Aloud Methode darin liegt zu verbalisieren, welche Gedanken ein Benutzer während der Bedienung der Anwendung hat, und da der Benutzer unter Beobachtung steht, ist es möglich zu analysieren, wie einfach oder schwer es für die Testperson ist, die grundlegenden Aufgaben während der erstmaligen Bedienung zu erledigen. Verschiedene Grade des Erfolgs beim Absolvieren einer Aufgabe mit der Anwendung helfen dabei die Erlernbarkeit zu klassifizieren (vgl. Task Success in Abschnitt 2.1.5). Die Berücksichtigung des Usability-Attributs Erlernbarkeit ist somit gegeben.

Efficiency / Effizienz:

Eine Methode zur Bemessung der Effizienz einer Anwendung beinhaltet die Time-on-task, also die Zeit zum Erledigen einer Aufgabe, zu stoppen, um so zum Beispiel Vergleiche zu ähnlichen Produkten ziehen zu können. Dabei gilt es außerdem herauszufinden, ob die benötigte Zeit für den Benutzer akzeptabel ist. Durch das laute Denken während der Durchführung des Tests wird die Testperson jedoch zusätzlich belastet und die Zeit zum Vollenden einer Aufgabe möglicherweise länger. Die Usability-Expertin Dana Chisnell²³ schreibt dementsprechend über

²³Chisnell ist die Co-Autorin des Buches „Handbook of Usability Testing - How to Plan, Design, and Conduct Effective Tests“ [RC08].

die Messung der Effizienz während des Think Aloud Tests:

„If you want to measure how much time it takes people to complete a task because you are particularly concerned with efficiency, introducing think-aloud is probably a bad idea. Talking about what you’re thinking slows you down while you choose words to convey your ideas about what’s happening and why.“ [Chi07]

Zur Messung der Effizienz einer Anwendung gibt es jedoch die Metrik nach Tullis & Albert (vgl. Performance Metriken in Abschnitt 2.1.5), nach welcher die Effizienz unter anderem durch die Anzahl der Schritte oder Aktionen, die ein Nutzer benötigt, bewertet werden kann. Auch dieses Verfahren ist mit der Think Aloud Methode nicht optimal durchführbar. Die Testpersonen könnten durch die zusätzliche kognitive Belastung oder die Präsenz des Moderators und Protokollanten in eine Stresssituation geraten. Stress hat unterschiedliche Auswirkungen auf die Testpersonen beziehungsweise Personen im Allgemeinen und deren Leistungsfähigkeit [Sta04, S. 96]. Der Testnutzer wählt somit möglicherweise nicht denselben Weg zur Lösung des Aufgabe, den er unter normalen Umständen gewählt hätte. Die Effizienz einer Anwendung ist somit nur unzureichend mit der Think Aloud Methode überprüfbar.

Memorability / Einprägsamkeit:

Es ist durchaus möglich, einen Testnutzer zu einem späterem Zeitpunkt erneut einzuladen, um zu analysieren, wie leicht ihm die Benutzung der Anwendung nach einem bestimmten Zeitraum fällt. Dieser Ablauf gehört jedoch nicht zu der üblichen Verfahrensweise der Think Aloud Methode (s.o.) und würde zusätzliche Arbeitsstunden und Kosten bedeuten, die in der Evaluation der Kriterien nicht berücksichtigt wurden. Des Weiteren ergibt sich die Einprägsamkeit aus der Verbesserung oder Verschlechterung der Effizienz, die jedoch wie zuvor beschrieben, nicht in ausreichendem Maße berücksichtigt werden kann. Die Einprägsamkeit einer mobilen Anwendung wird demzufolge unter normalen Bedingungen nur unzureichend durch die Think Aloud Methode getestet.

Errors / Fehler:

Durch die Beobachtung der Testperson im Umgang mit der Anwendung durch den Testmoderator sowie gegebenenfalls die Aufzeichnung des Displays des mobilen Endgerätes und den darauf stattfindenden Aktionen können die Fehler, die der Benutzer während der Bedienung macht, festgestellt werden. Durch die direkte Kommunikation kann außerdem ermittelt werden, wie es zu diesen Fehlern während der Anwendung kommt. Mit Hilfe der Kommentare der Testperson kann daraufhin analysiert werden, als wie schwerwiegend und störend dieser Fehler empfunden wurde und ob der Fehler möglicherweise durch die Stresssituation bedingt war. Die Überprüfung des Usability-Attributes Fehler wird demnach von der Think Aloud Methode unterstützt.

Satisfaction / Zufriedenheit:

Bei der Think Aloud Methode gibt es mehrere Möglichkeiten zu erkennen, was der Testnutzer bei der Bedienung der mobilen Applikation empfindet. Zum einen können bereits durch die ver-

balen Äußerungen Rückschlüsse auf die Laune der Testpersonen gewonnen werden. Geschickt gestellte Fragen oder Andeutungen können dabei helfen herauszufinden, ob den Benutzer etwas frustriert. Aber auch Gestik und Mimik sind unter Umständen Informationsquellen, die Auskunft über den Gemütszustand geben. Die Think Aloud Methode eignet sich also gut, um die *Self-Reported Metrics* und *Behaviorial and Physiological Metrics* (vgl. Abschnitt 2.1.5) einzusetzen und die Zufriedenheit der Anwender während der Bedienung zu analysieren.

Zusammenfassung:

Mit der Think Aloud Methode lassen sich folglich die drei Usability-Attribute Erlernbarkeit, Fehler und Zufriedenheit überprüfen, während die Effizienz und Einprägsamkeit nur bedingt beziehungsweise gar nicht getestet wird. Dies führt zu einer Bewertung von drei von fünf Punkten für die Berücksichtigung der Usability-Attribute durch die Think Aloud Methode.

Berücksichtigung des emotionalen Zustands

Der emotionale Zustand des Testnutzers kann bevor und während des Tests ermittelt werden. Dazu ist es zum einen möglich, die *Self-Reported Metrics* anzuwenden und die Testperson vor und während des Tests nach ihrem Befinden zu fragen. Zum anderen können auch *Behaviorial and Physiological Metrics* verwendet werden, um etwas über den emotionalen Zustand des Testnutzers zu erfahren. Dadurch, dass der Test im Allgemeinen in einer kontrollierten Umgebung stattfindet, kann der Test durch Sensoren ergänzt werden, die dabei helfen herauszufinden, wie die Testperson auf die Anwendung reagiert. Außerdem ist der direkte Sichtkontakt während der gesamten Testdauer gegeben, so dass erfahrene Testmoderatoren (oder gegebenenfalls spätere Betrachter der Aufzeichnung) die Gestik und Mimik des Testteilnehmers jederzeit beobachten und deuten können. Diese Aspekte und die Tatsache, dass die Aussprache der Gedankengänge der Kern des Think Aloud Tests ist, führen dazu, dass die Berücksichtigung des emotionalen Zustandes dieser Methode als *sehr stark* und somit mit fünf Punkten bewertet wird.

Berücksichtigung des physischen Zustands

Der Think Aloud Test wird in der Regel im Sitzen an einem Tisch durchgeführt, damit der Testmoderator und der Protokollant gute Einsicht in die Aktivitäten der Testperson auf dem mobilen Endgerät haben. Sollte es sich um ein Mock-Up der Anwendung handeln, ist durch die ruhende Position gegeben, dass detailliert über die Bestandteile der Benutzeroberfläche gesprochen und darauf gezeigt werden kann. Findet der Think Aloud Test mit Hilfe einer Emulation der Anwendung auf einem Desktop Computer statt, ist die Mobilität bereits durch Größe, Gewicht und Energieversorgung des Computers stark eingeschränkt/nicht vorhanden.

In begrenztem Rahmen ist es auch möglich die Testperson durch das Labor gehen zu lassen und

währenddessen die Anwendung zu bedienen. Dabei ist zu beachten, dass die zusätzliche Aufmerksamkeit, die der Nutzer möglicherweise aufbringen muss, um Hindernissen auszuweichen, die Kommunikationsfähigkeiten einschränkt. Außerdem wird es durch die Bewegung schwieriger für den Moderator, den Bildschirm des Endgerätes einzusehen.

Die Berücksichtigung des physischen Zustands ist demzufolge nur *sehr schwach* vorhanden und wird mit einem Punkt bewertet.

Berücksichtigung der natürlichen Nutzungsumgebung

Sofern die mobile Anwendung nicht dafür konzipiert wurde, ausschließlich in geschlossenen Räumen benutzt zu werden, findet die natürliche Nutzungsumgebung keine direkte Berücksichtigung in dem Think Aloud Test. Es ist jedoch möglich, Umgebungsvariablen, wie verschiedene Lichtverhältnisse oder Umgebungslautstärken, zu simulieren, um so die Lesbarkeit oder die Sprachsteuerung einer Anwendung unter verschiedenen Bedingungen zu testen. Aus diesem Grund wird die Berücksichtigung des natürlichen Nutzungsverhaltens noch mit zwei Punkten bewertet.

Berücksichtigung der mobilgerätespezifischen Ein- und Ausgabeprobleme

In einer späten Entwicklungsphase der Anwendung kann diese auf beliebig vielen verschiedenen Endgeräten, mit unterschiedlichen Ein- und Ausgabemedien installiert werden. Es ist kein Problem, den Test mit verschiedenen Endgeräten durchzuführen, da keine weiteren Hard- oder Softwareinstallationen nötig sind. Dem Testnutzer muss lediglich ein anderes Gerät zur Bedienung der Applikation vorgelegt werden. Die Kamera, die auf das Endgerät gerichtet ist, muss unter Umständen neu fokussiert werden. Das ist allerdings nur der Fall, wenn das weitere Endgerät wesentlich größer oder kleiner ist, und kostet im Normalfall nur einige Minuten Zeit. Wenn es für die Entwicklung der Anwendung notwendig ist, lässt der Think Aloud Test also eine *sehr starke* Berücksichtigung der mobilgerätespezifischen Ein- und Ausgabeprobleme zu.

Beeinflussung des natürlichen Nutzungsverhaltens

Die Testperson wird in seinem natürlichen Verhalten mit der Anwendung durch mehrere Faktoren gestört. Einer dieser Faktoren ist die Umgebung. Die geschlossenen Räume eines Usability-Labors entsprechen nicht der natürlichen Nutzungsumgebung der Anwendung. Hinzu kommt die Anwesenheit des Testmoderators und gegebenenfalls eines Protokollanten sowie der Dialog mit dem Testmoderator. Außerdem sind Uhrzeit und Dauer des Tests festgelegt und entsprechen somit in der Regel nicht der Realität. Ist eine Kamera auf das Endgerät gerichtet, ist es der

Testperson zudem nicht möglich, das Endgerät auf natürliche Art zu halten, da es sonst aus dem Blickfeld der Kamera geraten könnte (vgl. Abschnitt 2.2.2). Der Think Aloud Test beeinflusst das natürliche Nutzungsverhalten also *sehr stark* und wird dementsprechend mit null Punkten bewertet.

Gefundene Usability-Probleme

In einer Studie von Kallio et al. [KK05] wurde der Think Aloud Test für mobile Anwendungen im Feld mit derselben Methode im Labor verglichen. Dabei wurden mit beiden Verfahren 46 unterschiedliche Usability-Probleme gefunden, von denen 22 von mehr als einer Person aufgedeckt wurden. Weitere Testergebnisse über die gefundenen Probleme waren:

- Im Feld gab es lediglich bei drei Probleme einen statistisch signifikanten Unterschied in der Anzahl der gefundenen Probleme.
- Die im Feld gefundenen Probleme waren im Durchschnitt nicht schwerwiegender als die im Labor gefundenen.

Ein ähnlicher Versuch verglich ebenfalls das Usability-Testing im Feld mit dem Think Aloud Test im Labor. Das Ergebnis dieser Studie unterscheidet sich stark von der zuvor beschriebenen. In der Arbeit „Usability Evaluation for Mobile Device: A Comparison of Laboratory and Field Tests“ von Chen et al. [DTC06] wurden mit dem Labortest von zehn Testpersonen 92 Usability-Probleme aufgedeckt, während dieselbe Anzahl an Testpersonen im freien Feld 172 Probleme feststellte. Der größte Unterschied bestand bei der Anzahl der kritischen Probleme. Im Labor wurden 12 dieser kritischen Probleme ermittelt, während es im Feld 64 waren.

Die Testpersonen selbst nannten verschiedene Gründe für die große Differenz der gefundenen Probleme und führen es auf fünf Punkte zurück, die sich mit der Bewertung der Berücksichtigung der natürlichen Nutzungsumgebung des Think Aloud Tests im Labor decken. Diese fünf Punkte beziehen sich auf die im Feldtest gegebenen Eigenheiten [DTC06]:

- Geräuschpegel der Nutzungsumgebung
- Bewegung
- Mangel an Privatsphäre
- erhöhter Anspruch die Aufgaben zu erfüllen
- zusätzlicher Stress/Nervosität

Auch bei näherer Betrachtung der Studien auf getestete Anwendung und zum Testen benutzte Mittel wird nicht ersichtlich, wie der deutliche Unterschied der beiden Studien zustande kommt. Es lässt sich also schlussfolgern, dass die Anzahl der gefundenen Probleme im Labor vermutlich

selten höher ist als in einem Feldversuch. Außerdem lassen sich mutmaßlich mehr Usability-Probleme, die abhängig von der direkten Nutzungsumgebung sind, durch den Think Aloud Test im Feld ermitteln.

Gesamtbewertung

Zusammenfassend lässt sich feststellen, dass die größten Nachteile des Think Aloud Tests, der im Labor durchgeführt wird, darin liegen, dass der Benutzer nicht auf eine natürliche Art und Weise mit dem Endgerät und der Anwendung umgehen kann. Hinzu kommt, dass die Methode eine lange Dauer zur Durchführung benötigt²⁴, wenn nur ein Usability-Experte als Moderator zur Verfügung steht und die Testpersonen nacheinander befragt werden müssen. Diese Nachteile sind auch in der Übersicht über die Bewertung der einzelnen Kriterien in Abbildung 4.2 offensichtlich²⁵. Dort sind jedoch auch die Vorteile des Verfahrens, die, neben dem flexiblen Einsatzzeitpunkt und dem leichten Austausch der Endgeräte, vor allem in der Berücksichtigung der Nutzerresonanz liegen, eindeutig zu erkennen. Es wird deutlich, dass der Think Aloud Test großen Wert darauf legt, was die Testnutzer über die mobile Anwendung denken, und somit seinem Namen gerecht wird.

Zeitaufwand	●●○○○
Kosten	●●●●○
Zeitpunkt der Durchführung	●●●●●
Berücks. der Usability-Attribute	●●●○○
Berücks. des emotionalen Zustands	●●●●●
Berücks. des physischen Zustands	●○○○○
Berücks. der natürlichen Nutzungsumgebung	●●○○○
Berücks. der mobilgerätespez. Ein-/Ausgabeprobleme	●●●●●
Beeinflussung des natürlichen Nutzungsverhaltens	○○○○○

Abbildung 4.2: Übersicht über die Bewertung des Think Aloud Tests (im Labor)

4.2 Think Aloud Test(im Feld)

Im folgenden Abschnitt wird ein Verfahren beschrieben und bewertet, mit welchem sich der Think Aloud Test in natürlicher Umgebung durchführen lässt. Als Grundlage dienen dafür die Studie „*Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing*“ von Kallio et al. [KK05], die die Durchführung des Think Aloud Tests im Labor mit

²⁴Bei der Dauer der Methode ist außerdem noch zu berücksichtigen, dass die Analyse der Aufzeichnungen gegebenenfalls zusätzliche Zeit kostet.

²⁵Die ausgefüllten Punkte in der Abbildung stehen für die vergebenen Punkte in der Bewertung des jeweiligen Kriteriums. Die nicht ausgefüllten Punkte stellen die Punkte dar, die noch maximal zusätzlich hätten erreicht werden können.

der im Feld vergleicht. In einem Versuch von Duh et al. [DTC06] wird derselbe Vergleich unter der Verwendung von ähnlichen Mitteln, um die Testnutzer während des Benutzens einer mobilen Applikation im Feld aufnehmen zu können, angestellt. Der größte Unterschied zwischen diesen beiden wissenschaftlichen Arbeiten ist, dass Kallio et al. die Testpersonen von einem Moderator begleiten lassen. Dieser Moderator ist in der Lage, mit den Teilnehmern zu interagieren, wodurch diese Methode dem Think Aloud Test in kontrollierter Umgebung ähnlicher ist.

Des Weiteren befassen sich Beck et al. [BCK⁺03] mit dem Versuch, die Laborumgebung an die tatsächlichen Umstände einer natürlichen Benutzung anzupassen. Um Referenzwerte zu erhalten, wenden auch sie den Think Aloud Test im Feld an. Da der Fokus der Arbeit jedoch auf Methoden zur Simulation der natürlichen Nutzungsumgebung liegt, nimmt der Feldversuch nur eine untergeordnete Rolle ein. Einer der zwei dort durchgeführten Experimente im Feld findet sogar ohne Kamera zur Beobachtung der Testpersonen statt. Aus diesen Gründen werden die Studien von Duh et al. [DTC06] und Beck et al. [BCK⁺03] lediglich zur Ergänzung der von Kallio et al. [KK05] gewonnenen Erkenntnisse genutzt.

4.2.1 Beschreibung

Der Think Aloud Test in einer natürlichen Nutzungsumgebung verläuft ähnlich wie der Think Aloud Test im Labor (vgl. 4.1.1). Da der Moderator jedoch nicht die Möglichkeit hat, ständig dem Testteilnehmer über die Schulter zu blicken, um dessen Aktionen auf dem Bildschirm zu verfolgen, müssen entsprechende Anpassungen vorgenommen werden. Kalli et al. [KK05] lösen dies, indem sie den Teilnehmer und den Moderator mit folgenden Ausrüstungen ausstatten. An der Person, die die mobile Anwendung bedient, werden dazu drei Kameras befestigt: eine zur Aufzeichnung des mobilen Endgerätes, eine zur Aufzeichnung des Gesichts der Testperson und einer zur Aufnahme der Umgebung. Zu der Ausrüstung gehören außerdem ein Mikrofon, ein Audio- und Video-Aufnahmegerät, ein kabelloser Video Transceiver, ein Quad-Prozessor zur Verarbeitung der Videodaten und Batterien für die Energieversorgung der gesamten Ausrüstung.

Der Moderator hält einen sechs Zoll großen Bildschirm, auf dem er die Interaktion des Testnutzers mit dem Endgerät verfolgen kann. Dazu sind ebenfalls ein kabelloser Video Transceiver und eine Batterie notwendig. Die Umgebung aus der Sicht des Moderators wird mit einer weiteren Kamera aufgezeichnet [KK05]. Abbildung 4.3 zeigt Fotos der Ausrüstung eines solchen Versuchsaufbaus.

Der Moderator kann mit diesen Mitteln einige Schritte von der Testperson entfernt sein und das Geschehen weiterhin verfolgen. Wie bei dem Think Aloud Test im Labor sollen die Testnutzer nach Möglichkeit ihre Gedanken über die Anwendung frei äußern. Sie werden dazu angehalten zu erzählen, was sie anklicken, welches Resultat sie erwarten und ob etwas Unerwartetes



Abbildung 4.3: Ausrüstung des Think Aloud Test in natürlicher Nutzungsumgebung [KK05]

geschieht. Während des gesamten Tests sollen die Teilnehmer in ihren Bewegungen nicht eingeschränkt werden. Ihnen wird somit zum Beispiel nicht vorgeschrieben, dass sie fortwährend gehen oder sich zu einem bestimmten Zeitpunkt hinsetzen sollen. Auf diese Weise wird die Natürlichkeit der Aktionen gewährleistet [KK05].

4.2.2 Bewertung

Der Think Aloud Test im Feld basiert auf dem Prinzip des gleichen Tests im Labor: das Gewinnen von Informationen über die Usability einer Anwendung dadurch, dass die Testpersonen ihre Gedanken laut aussprechen. In der nachfolgenden Bewertung können deshalb die Kriterien, die nicht durch den Nutzungskontext und die damit verbundenen Anpassungen an das Testverfahren beeinflusst werden, auf dieselbe Weise evaluiert werden wie die Kriterien bei der Bewertung des Think Aloud Tests im Labor.

Zeitaufwand

Nach Kallio et al. [KK05] ist es angemessen, für einen Think Aloud Test, der in freier Umgebung stattfindet, die doppelte Zeit zu berechnen, die für denselben Test im Labor benötigt werden würde. Diese Zeit resultiert aus den Zeiten, in denen keine direkte Interaktion mit der Anwendung stattfindet, zum Beispiel die Fahrt oder der Fußweg zu einem Platz, an dem die Applikation für gewöhnlich auch benutzt werden würde. Weiterer Faktor, der die Dauer des Tests

verlängert, ist die Installation der Kameras zu Beginn des Tests mit einer neuen Testperson. Mit den in Abschnitt 4.1.2 vorgenommenen Annahmen, dass fünf Personen und ein Moderator für den Test zur Verfügung stehen, verdoppelt sich die dort festgestellte Zeit auf 15 Stunden. Dies entspricht einer Bewertung von einem Punkt.

Kosten

Die längere Dauer des Tests führt zu erhöhten Kosten des Usability-Experten. Bei der Dauer von 15 Stunden ergeben sich:

$$\text{Kosten für den Usability-Experten} = 15h * 32,69\text{€/h} = 490,35\text{€}.$$

Die durchschnittlichen Kosten für fünf Testnutzer mit einem Betrag von 617€ und die folgenden Kosten für die Ausrüstung kommen zu den Kosten für die Experten hinzu. Für einen Quad-Prozessor wird mit Anschaffungskosten von 200€, für die beiden Kameras zur Aufzeichnung der Umgebung mit jeweils 250€, für die Batterie 50€ und für die Video Transceiver mit insgesamt 150€ gerechnet²⁶. Für die Kamera, die auf dem mobilen Endgerät befestigt wird, werden 500€ berechnet. Dies entspricht dem Preis, der für dieselbe Kamera im Labor angenommen wurde. Bei der Kamera ist es wichtig, dass sie aus naher Entfernung und in Bewegung detailliert fokussiert bleibt und detailliert aufzeichnen kann. Außerdem sollte sie möglichst klein sein, um das Endgerät nicht unnötig schwerer und unhandlicher zu machen.

Mit diesen genannten Ausgaben für Personal und Ausrüstung ergeben sich Gesamtkosten von etwa 2507€. Daraus ergibt sich eine Bewertung von drei Punkten. Die Video- und Audioaufzeichnung muss anschließend noch analysiert werden, wodurch weitere Kosten entstehen.

Zeitpunkt der Durchführung

Im Gegensatz zu dem Think Aloud Test im Labor, bei dem die Testteilnehmer auch mit Hilfe von Mock-Ups oder Skizzen ihre Gedanken über eine Anwendung äußern können, ist es im Feld wenig sinnvoll, keine zumindest zum Teil funktionierende Anwendung zur Verfügung stellen zu können. Die Bedienung der Funktionen in der natürlichen Nutzungsumgebung ist der Aspekt, der dieser Methode zusätzlichen Nutzen erbringen soll [KK05, DTC06]. Funktioniert die Anwendung nicht, müssen sich die Kommentare der Testnutzer auf das Design der Benutzeroberfläche beschränken, was im Labor mit weniger Umstand verbunden ist. Da also zumindest ein Prototyp der Anwendung fertiggestellt sein muss, wird dieses Kriterium mit zwei Punkten bewertet.

²⁶Diese Preise ergaben sich aus einem Vergleich der Preise auf <http://shopping.google.com> am 7.12.2011.

Berücksichtigung der Usability-Attribute

Die Berücksichtigung Usability-Attribute kann von dem Think Aloud Test im Feld ähnlich wie im Labor stattfinden. Nachfolgend wird deshalb die Berücksichtigung einzelnen Attribute nur in kurzer Form erneut erläutert.

Learnability / Erlernbarkeit:

Die Audioaufzeichnungen der Gedanken des Testnutzers erlaubt es in der späteren Analyse, die Probleme des Nutzers beim Erlernen der Anwendung zu erkennen. Zusätzlich können die Videoaufzeichnungen des Gesichts und somit der Mimik der Testperson weitere Informationen liefern (vgl. Abschnitt 2.1.5).

Efficiency / Effizienz:

Auch bei dem Think Aloud Test im Feld verhindert das laute Denken, die ungewohnte Situation und zusätzliche kognitive Belastungen eine hinreichende Berücksichtigung der Effizienz einer Anwendung. Durch die genannten Faktoren hat der Nutzer nicht die Möglichkeit, ein natürliches Maß an Konzentration auf die Bedienung der Anwendung zu legen, so dass die Effizienz nicht durch Time-On-Task oder andere Metriken gemessen werden kann.

Memorability / Einprägsamkeit:

Dadurch, dass die Effizienz durch den Think Aloud Test nicht ausreichend berücksichtigt werden kann, gilt dasselbe für die Einprägsamkeit. Die Einprägsamkeit würde bemessen, wie gut ein Nutzer die Anwendung nach einiger Zeit des Nichtbenutzens noch bedienen kann. Ein wichtiges Merkmal dafür ist die Veränderung der Effizienz und somit zum Beispiel die Differenz der Dauer zum Erledigen einer Aufgabe zwischen den Versuchen.

Errors / Fehler:

Die Aufzeichnung des Bildschirms und die verbalen Äußerungen der Testnutzer helfen während der Analyse, die Fehler, die der Nutzer gemacht hat, zu erkennen und die Gründe dafür zu ermitteln. Außerdem kann der Moderator während des Tests gezielt Fragen stellen und damit ein besseres Verständnis über die Gründe für bestimmte Aktionen, die zu Fehlern führten, gewinnen.

Satisfaction / Zufriedenheit:

Die Zufriedenheit des Testnutzer kann mit Hilfe der Aufzeichnungen seiner Gesichtszüge und den ausgesprochenen Gedanken ermittelt werden. Es lassen sich also sowohl *Self-Reported Metrics* als auch *Behavioral and Physiological Metrics*(vgl. Abschnitt 2.1.5) anwenden.

Zusammenfassung:

Insgesamt lassen sich also drei der fünf Usability-Attribute durch den Think Aloud Test im Feld ausreichend berücksichtigen.

Berücksichtigung des emotionalen Zustands

Der emotionale Zustand der Testpersonen kann vor und während dem Test ermittelt werden. In einem einleitenden Gespräch kann zum Beispiel herausgefunden werden, ob den Teilnehmer etwas belastet, was sich womöglich auf seine Stimmung auswirkt. So können spätere Fehlinterpretationen vermieden werden. Durch die Aufzeichnungen des Gesichts während des Tests und die direkte Nähe des Moderators lassen sich Gestik, Mimik und verbale Äußerungen dazu verwenden, etwas über den emotionalen Zustand des Testnutzer herauszufinden. Die Grundlage des Think Aloud Tests, den Nutzer seine Meinung über die Anwendung frei äußern zu lassen, ist auch durch den Wechsel des Orts in eine unkontrollierten Umgebung nur leicht beeinträchtigt. Der Moderator ist möglicherweise nicht immer im Stande, direkten Sichtkontakt zu der Testperson zu halten. Dennoch reichen die durchgängige Kommunikation mit der Testperson und die Videoaufzeichnungen dazu aus, dass der emotionale Zustand der Nutzer *sehr stark* berücksichtigt werden kann.

Berücksichtigung des physischen Zustands

Den Testpersonen ist es möglich sich, frei zu bewegen und die Applikation dort zu bedienen, wo es üblich ist. Dadurch, dass das „Labor“ in Form der Ausrüstung und der Moderator sie begleitet, kann die Usability während sämtlicher Bewegungsarten getestet werden. Dazu gehören demnach aktive und passive Bewegungsarten und damit verbundene sich ändernde Geschwindigkeiten und Umgebungen. Der physische Zustand kann also *sehr stark* berücksichtigt werden.

Berücksichtigung der natürlichen Nutzungsumgebung

Die Berücksichtigung der natürlichen Nutzungsumgebung ist ebenso möglich wie die des physischen Zustands. Die Testpersonen können dazu aufgefordert werden, die mobile Anwendung an vorher ausgesuchten Orten zu benutzen oder selbst Orte zu wählen, die sie für angemessen halten. In der Studie von Kallio et al. [KK05] wurde die Testnutzer zum Beispiel dazu aufgefordert, die Anwendung auf einer Rolltreppe, in einem Zug oder während des Gehens durch eine Einkaufspassage zu benutzen. Die Berücksichtigung der natürlichen Nutzungsumgebung wird dementsprechend als *sehr stark* möglich bewertet.

Berücksichtigung der mobilgerätespezifischen Ein- und Ausgabeprobleme

Es ist grundsätzlich möglich, dass mobile Endgerät zu tauschen, um die verschiedenen Ein- und Ausgabemedien zu testen. Dies ist bei einem Feldversuch jedoch mit mehr Aufwand ver-

bunden, als es bei einem Test im Labor der Fall ist. Die Ausrüstung muss eventuell angepasst werden, weil die Kamera, die auf das Endgerät gerichtet ist, neu justiert werden muss. Im Labor kann eine Kamera eingesetzt werden, die nicht direkt an dem Endgerät befestigt ist, was den Austausch der Geräte vereinfacht. Während des Feldtests muss die Kamera jedoch an dem mobilen Endgerät befestigt werden. Dies führt unter Umständen dazu, dass Teile des Bildschirms von der Halterung verdeckt werden, wenn die Einfassung nicht genügend Fläche bietet. Aus diesen Gründen ist insgesamt nur eine *mittlere* Berücksichtigung der mobilgerätespezifischen Ein- und Ausgabeprobleme möglich.

Beeinflussung des natürlichen Nutzungsverhaltens

Vorteilhaft gegenüber des Think Aloud Tests im Labor ist, dass der Moderator nicht mehr in unmittelbarer Nähe zu der Testperson stehen muss, um mit auf den Bildschirm blicken zu können. Die Testperson wird im Normalfall durch diese also nur wenig in seinem natürlichen Nutzungsverhalten gestört. Die Befestigung einer Kamera auf dem mobilen Endgerät schränkt dessen natürliche Handhabung sehr stark ein. Gewicht und Größe werden durch die zusätzliche Hardware erhöht, die Balance des Gerätes verändert und die Mobilität eingeschränkt. Es kann somit nicht mehr in die Hosentasche gesteckt werden, sondern muss dauerhaft von der Testperson gehalten werden. Das zusätzliche Gewicht auf der Oberseite des Endgeräts führt möglicherweise dazu, dass die Testperson es auf eine andere Art und Weise festhält, als es normalerweise der Fall wäre. Dadurch könnte wiederum die Bedienung der Anwendung unüblich ausfallen, weil die Finger nicht mehr dieselben Stellen der Benutzungsoberfläche erreichen. Das natürliche Nutzungsverhalten der Testpersonen wird während des Think Aloud Tests im Feld also *stark* beeinflusst.

Gefundene Usability-Probleme

Ein ausführlichere Zusammenfassung der Vergleiche der gefundenen Usability-Probleme von dem Think Aloud Test im Feld mit dem im Labor ist in Abschnitt 4.1.2 zu finden. Kallio et al. [KK05] sowie [BCK⁺03] haben dabei keine bedeutenden Unterschiede zwischen den Verfahren feststellen können. Chen et al. [DTC06] hingegen haben mit dem Think Aloud Test im Feld fast doppelt so viele Probleme wie im Labor aufdecken können. Diese großen Unterschiede in den Ergebnissen sind vermutlich auf die Verschiedenartigkeit der getesteten Anwendung und möglicherweise auf die Expertise der eingesetzten Usability-Experten zurückzuführen. Um einen repräsentativen Vergleich der beiden Methoden zu haben, ist das Testen von weiteren Anwendungen nötig. Diese Stichprobe würde somit eine, statistisch betrachtet, größere Aussagekraft besitzen.

Gesamtbewertung

Abbildung 4.4 zeigt, dass zuvor festgestellte Nachteile des im Labor durchgeführten Think Aloud Tests (vgl. 4.1.2) durch den gleichen Test im Nutzungskontext behoben werden. Während mit Hilfe dieser Variante des Think Aloud Test weiterhin Erkenntnisse aus dem Feedback der Nutzer gewonnen wird, ist nun zusätzlich eine bessere Berücksichtigung der natürlichen Nutzungsumgebung und der damit verbundenen Bewegungsarten der Nutzer möglich. Daraus resultieren jedoch höhere Kosten und eine lange Gesamtdauer. Die während des Tests aufgezeichneten Aktionen und Umgebungen der Nutzer müssen außerdem noch analysiert und ausgewertet werden. Hinzu kommt, dass der Test nicht zu jeder Phase der Entwicklung stattfinden kann und gegenüber dem Test im Labor an Flexibilität verliert.

Zeitaufwand	●○○○○
Kosten	●●●○○
Zeitpunkt der Durchführung	●●○○○
Berücks. der Usability-Attribute	●●●○○
Berücks. des emotionalen Zustands	●●●●●
Berücks. des physischen Zustands	●●●●●
Berücks. der natürlichen Nutzungsumgebung	●●●●●
Berücks. der mobilgerätespez. Ein-/Ausgabeprobleme	●●●○○
Beeinflussung des natürlichen Nutzungsverhaltens	●○○○○

Abbildung 4.4: Übersicht über die Bewertung des Think Aloud Tests (im Feld)

4.3 Kognitiver Walkthrough

Der Kognitive Walkthrough gehört zu der Gruppe der Usability-Inspektionen. Eine Usability-Inspektion zeichnet sich dadurch aus, dass ein oder mehrere Usability Experten ein Interface hinsichtlich der Usability inspizieren. Sie wird auch als *discount usability engineering solution* bezeichnet, wodurch die Kosteneffizienz dieser Methode ausgedrückt wird [Nie94b]. Es gibt zahlreiche verschiedene Methoden der Usability Inspektion, die sich im Kern jedoch gleichen. Sie gelten alle als schnelle, kostengünstige Evaluationsmethoden, die bereits in frühen Entwicklungsphasen eingesetzt werden können, und sie werden hauptsächlich von einem oder mehreren Usability Experten durchgeführt [Kra, Nie94b]. Tabelle 4.1 zeigt die etablierten Methoden der Usability-Inspektion in einer Übersicht.

Methode	Beschreibung
Heuristische Evaluation	Der Usability-Experte untersucht, ob die Anwendung etablierten Prinzipien (Heuristiken) folgt.
Heuristische Estimation	Der Experte vergleicht und bewertet die Usability mehrerer Systeme anhand der Heuristiken.

Methode	Beschreibung
Kognitiver Walkthrough	Der Evaluator simuliert einen typischen Benutzer und dessen Ziele mit der Anwendung. Er überprüft dabei, ob diese simulierten Ziele mit dem Wissen des Beispielanwenders (der Persona) erreicht werden kann.
Heuristischer Walkthrough	Die Heuristiken werden mit den Anwendungsszenarien des Walkthroughs kombiniert.
Kontextueller Walkthrough	Der Heuristische Walkthrough wird im Nutzungskontext durchgeführt.
Pluralistischer Walkthrough	Der Usability-Experte, ein Entwickler und ein Testanwender verwenden die Applikation für ein festgelegtes Ziel und diskutieren ihre Vorgehensweise.
Formale Usability Inspektion	Zuvor festgelegte Arbeitsaufgaben werden ausgeführt und aufgetretene Fehler notiert und im Quellcode markiert. Nielsen bezeichnet diese Inspektion als Kombination von heuristischer Evaluation und vereinfachtem Kognitiven Walkthrough.
Feature-Inspektion	Die typischen Funktionen der Anwendungen werden überprüft. Dabei wird festgestellt, ob die Funktionen zu unnatürlich zu bedienen sind, zu viele Schritte benötigen oder zu viel Wissen voraussetzen.
Konsistenz-Inspektion	Die Anwendung und ihre Elemente werden auf Konsistenz überprüft. Dies ist besonders relevant, wenn mehrere Anwendungen entwickelt werden, deren Design anwendungsübergreifend konsistent sein soll.
Standards-Inspektion	Der Usability-Experte überprüft, ob die Anwendung festgelegte Standards einhält.
Guideline-Checkliste	Es wird überprüft, ob die Anwendung zuvor festgelegte Richtlinien einhält.

Tabelle 4.1: Übersicht der Methoden zur Usability Inspektion. In Anlehnung an [Kra, Nie94b]

Die am weitesten verbreiteten Methoden zur Usability-Inspektion sind die Heuristische Evaluation und der Kognitive und Pluralistische Walkthrough [BMO08]. Zur Repräsentation der Inspektionsmethoden werden deshalb die Heuristische Evaluation und der Kognitive Walkthrough gewählt. Der Kognitive Walkthrough wurde gewählt, weil bereits Forschungsarbeiten hinsichtlich seiner Anpassung an die Bedürfnisse des Usability-Testings für mobile Endgeräte bestehen. Der Pluralistische Walkthrough unterscheidet sich außerdem hauptsächlich dadurch vom Kognitiven Walkthrough, dass die Nutzungsszenarien nicht nur von Usability-Experten, son-

den zusätzlich von Anwendern und Entwicklern durchgegangen werden [Nie94c, Nie94b]. Es ist somit schwierig den Nutzungskontext zu berücksichtigen. Die Teilnehmer der Inspektion müssten dazu zum Beispiel in einem Feldversuch Pausen einlegen, um zu diskutieren. Die Praktikabilität dieser Methode wurde bisher nicht untersucht.

Des Weiteren sind Kosten und Zeitaufwand des Pluralistischen Walkthroughs möglicherweise höher als die des Kognitiven Walkthroughs, da die Testperson und die Zeit des Entwicklers entlohnt werden und die Diskussion der einzelnen Usability-Merkmale Zeit beansprucht. Die Bewertung der anderen Kriterien würde ähnlich dem des Kognitiven Walkthroughs ausfallen. Die Heuristiken für die Heuristische Evaluation wurden von Bertini et al. [BCD⁺08] angepasst, um für die Evaluation von mobilen Anwendungen geeignet zu sein. Die Beschreibung und Bewertung dieser Methode findet sich im Abschnitt 4.4.

4.3.1 Beschreibung

Bei einem Kognitiven Walkthrough untersuchen die Evaluatoren, häufig Usability-Experten, eine Anwendung hinsichtlich ihrer Usability. Der Kognitive Walkthrough wurde ursprünglich im Jahr 1990 von Lewis et al. [LPWR90] für Anwendungen, die Informationen bereitstellen, entwickelt. Später [WRLP94] wurde dieses Verfahren erweitert und somit für alle interaktiven Systeme anwendbar [BTT05, S. 230 ff.].

Der Kognitive Walkthrough kann sowohl mit Hilfe einer Beschreibung der Anwendung, eines Mock-Ups oder Prototypen, als auch mit einer vollständig funktionierenden Anwendung durchgeführt werden. Der Inspektor benötigt außerdem ein konkretes Szenario, in dem die Anwendung bedient wird. Dazu gehören Annahmen über die typische Nutzerschaft, den Nutzungskontext und die Ablauf der Aktionen, die dazu nötig sind, die zu analysierende Aufgabe zu erfüllen. Dabei können auch mehrere Szenarien untersucht werden. Zu diesem Zweck werden zum einen besonders gebräuchliche Aktionsabläufe untersucht und zum anderen weniger übliche, aber dafür kritische Aktionssequenzen. Unter den gegebenen Bedingungen untersucht der Inspektor die Anwendung bei jedem Schritt auf die folgenden Fragen [BTT05, WRLP94]:

- Werden die Personen, die die Anwendung benutzen, den richtigen Effekt erzielen wollen?
- Werden sie bemerken, dass die richtige Aktion verfügbar ist?
- Werden sie die korrekte Aktion mit dem Effekt, den sie zu erzielen versuchen, in Verbindung bringen?
- Wenn die Personen den richtigen Schritt getan haben, werden sie merken, dass sie einen Fortschritt auf dem Weg zu ihrem Ziel gemacht haben?

Sollte eine dieser Fragen verneint werden, ist ein Usability-Problem gefunden und wird notiert. Die gefundenen Probleme werden später mit den Entwicklern der Anwendung und / oder

Benutzeroberfläche diskutiert und Lösungen erarbeitet. Lewis et al. [WRLP94] gliedern die genannten notwendigen Schritte des Kognitiven Walkthroughs in fünf Phasen.

1. Definition des Inputs für den Walkthrough (Nutzerschaft, Beispielaufgaben, Aktionsabläufe zum Erfüllen der Aufgaben, Beschreibung oder Implementation der Benutzeroberfläche / der Anwendung)
2. Berufung der Analysten / Usability-Experten
3. *Durchschreiten* (walk through) der Aktionsabläufe für jede Aufgabe unter Berücksichtigung der oben genannten Fragestellungen
4. Notieren der kritischen Informationen (Anforderungen an die Vorkenntnisse der Nutzer, Annahmen über die Nutzerschaft, Notizen über Randerscheinungen und Designänderungen, Erklärungen des Nutzerverhaltens)
5. Überarbeitung der Anwendung zum Lösen der Probleme

Dadurch, dass diese Phasen ohne viel technischem Aufwand auskommen, lässt sich der Kognitive Walkthrough überall dort ausführen, wo sich das mobile Endgerät auch bedienen lässt. Der Walkthrough ist demnach nicht an ein Usability-Labor gebunden, sondern kann in verschiedenen Umgebungen durchgeführt werden. Der größte Unterschied zu dem Think Aloud Test im Feld ist dann, dass der Usability-Experte selbst die Rolle des Endnutzers annimmt.

4.3.2 Bewertung

In der Bewertung des Kognitiven Walkthroughs werden, soweit vorhanden, die Informationen verwendet, die dieses Verfahren für das Usability-Testing von mobilen Applikationen verwenden.

Zeitaufwand

Der Kognitive Walkthrough gilt als schnelle und kostengünstige Methode des Usability-Testing, denn bereits drei bis fünf Experten entdecken 75%-80% der Usability-Probleme [Nie94b]. Die Dauer des Tests hängt dennoch von der Anzahl der durchzuführenden Aufgaben und Länge der Aktionsabläufe ab und lässt sich demnach nicht konkret bestimmen. Gabrielli et al. gaben für die Dauer des Kognitiven Walkthroughs im Labor einen Mittelwert von 53 Minuten²⁷ und für die Dauer in natürlicher Umgebung 90 Minuten²⁸ an [GMKC05]. Da die Experten den Test parallel durchführen können, sollte die Dauer nur selten zwei Stunden überschreiten.

²⁷SD = 21,38 Minuten.

²⁸SD = 15 Minuten.

Der Zeitaufwand des Kognitiven Walkthroughs ist somit entsprechend der Bewertungsskala in Abschnitt 3.1 mit fünf Punkten zu bewerten.

Kosten

Die Kosten des Kognitiven Walkthroughs beschränken sich auf die Entlohnung der eingesetzten Usability-Experten. Bei dem Einsatz von fünf Usability Experten für einen neunzigminütigen Test ergeben sich die Kosten für die Usability-Experten:

$$\text{Kosten für die Usability-Experten} = 5 * 1,5h * 32,69\text{€}/h = 245,18\text{€}$$

Somit sind noch etwa 750€ verfügbar, die für die Bezahlung der Experten während der Analysephase oder für Diktiergeräte, Aufnahmemedien und ähnliches verwendet werden können, ohne dass sich die Bewertung von fünf Punkten für die Kosten des Kognitiven Walkthroughs ändern würde. Sind die Notizen des Usability-Experten während des Tests detailliert genug, können die Lösungen für die Usability-Probleme möglicherweise ohne ihn entwickelt werden, so dass das Budget für einen anschließenden weiteren Test aufgespart werden kann.

Zeitpunkt der Durchführung

Nach Wharton et al. genügt eine Beschreibung der Benutzeroberfläche oder Mock-Up in Papierform oder ein Prototyp zur Durchführung des Kognitiven Walkthroughs [WRLP94]. Mit diesen Mitteln kann somit bereits eine Überprüfung der Usability der Benutzeroberfläche stattfinden, für die Überprüfung auf Merkmale wie Fehlerrobustheit oder Effizienz ist, wie bei allen Usability-Methoden, allerdings eine funktionierende Anwendung nötig. Der Kognitive Walkthrough ist also in jeder Phase der Anwendungsentwicklung durchführbar und wird dementsprechend mit fünf Punkten bewertet.

Berücksichtigung der Usability-Attribute

Um zu überprüfen, ob der Kognitive Walkthrough die Usability-Attribute berücksichtigt, werden diese nachfolgend einzeln betrachtet und auf die Möglichkeit zur Anwendung der entsprechenden Metriken untersucht. Wie repräsentativ die so gewonnenen Ergebnisse sind, hängt jedoch von den Fähigkeiten des Usability-Experten und davon, wie gut er sich in einen möglichen Nutzern versetzen kann, ab.

Learnability / Erlernbarkeit:

Grundsätzlich ist es möglich, die Erlernbarkeit einer Anwendung mit Hilfe des Kognitiven Walkthroughs zu analysieren. Die dafür relevante Metrik ist der *Task Success*, der den Grad

des Erfolgs, bei dem Versuch eine Aufgabe abzuschließen, misst. Eine komplette Phase des Kognitiven Walkthroughs besteht daraus, bestimmte Aufgaben mit der Anwendung zu erledigen²⁹ und zu überprüfen, wie einfach dem Nutzer die Lösung dieser Aufgaben fällt [WRLP94].

Efficiency / Effizienz:

Die direkte Benutzung der Anwendung erlaubt dem Usability-Experten auch, die Zeit zu messen oder die Schritte zu zählen, die er für die Erledigung der gestellten Aufgabe benötigt. Anhand dieser Werte lässt sich die Effizienz der Anwendung bestimmen.

Memorability / Einprägsamkeit:

Um die Einprägsamkeit der Anwendung zu ermitteln, reicht gegebenenfalls die Erfahrung des Usability-Experten. Alternativ müsste der Usability-Experte nach einiger Zeit einen erneuten Walkthrough durchführen. Durch seine Arbeit besitzt der Usability-Experte jedoch vermutlich enorme Vorkenntnisse im Umgang mit Applikationen. Es ist nicht auszuschließen, dass diese Vorkenntnisse die Beurteilung der Einprägsamkeit einer mobilen Anwendung verfälschen. Die Berücksichtigung dieses Attributs ist also nur unter Vorbehalt gegeben.

Errors / Fehler:

Während der Benutzung der mobilen Anwendung macht der Usability-Experte gegebenenfalls Fehler, die das erfolgreiche Absolvieren der Aufgabe verhindern. Dies sind nach Möglichkeit alle Fehler, die der Endnutzer auch machen könnte. Die Berücksichtigung dieses Attributes ist somit gegeben.

Satisfaction / Zufriedenheit:

Die Zufriedenheit eines Benutzers kennzeichnet sich durch das Empfinden beim Bedienen der Anwendung und ist somit ein überwiegend subjektives Attribut. Sehr qualifizierte Usability-Experten können die Zufriedenheit der Anwender möglicherweise einschätzen, sie lässt sich jedoch, ähnlich wie die Einprägsamkeit, nicht optimal durch den Kognitiven Walkthrough bemessen.

Zusammenfassung:

Die Attribute Erlernbarkeit, Effizienz und Fehler lassen sich durch fähige Usability-Inspektoren durchaus vollständig messen und bewerten. Die Evaluation Einprägsamkeit der Anwendung und Zufriedenheit der Nutzer ist jedoch nur begrenzt möglich. Aus diesem Grund wird die Berücksichtigung der Usability-Attribute nur mit vier Punkten bewertet, auch wenn alle Attribute, zumindest teilweise, berücksichtigt werden.

Berücksichtigung des emotionalen Zustands

In einem Artikel über verschiedene Usability-Testmethoden vertritt Thomas James die Meinung, der Kognitive Walkthrough „*does not address user satisfaction*“ [Jam12]. Es ist korrekt,

²⁹Bei dem Testen mit einem Mock-Up oder ähnlichem kann die Aufgabe jedoch nur imaginär ohne erkennbaren Erfolg erledigt werden.

dass der Usability-Experte, der den Kognitiven Walkthrough durchführt, während des Tests über keinerlei Nutzerfeedback verfügt. Er erhält demnach keine direkten Informationen, was ein Nutzer vor und bei der Benutzung der Anwendung empfindet und ob ihm die Benutzung Freude bereitet. Das bedeutet, es ist dem Experten weder möglich *Self-Reported Metrics* noch *Behavioral and Physiological Metrics* einzusetzen. Bei dem Kognitiven Walkthrough ist der Usability-Experte jedoch ohnehin abhängig von seinen empathischen Fähigkeiten, um sich in einen potenziellen Nutzer hineinzuversetzen. Außerdem können ihm noch sein eigenes Empfindungsvermögen und seine Erfahrung aus bisheriger Arbeit mit ähnlichen Anwendungen und Tests, die Testpersonen einbeziehen, als Evaluationshilfe dienen. Die Berücksichtigung des emotionalen Zustands kann also noch als *sehr schwach* bewertet werden.

Berücksichtigung des physischen Zustands

Der Kognitive Walkthrough in seiner traditionellen Ausführung berücksichtigt Bewegungen der Nutzer während der Anwendung nicht. Dies ist beim Testen von Anwendungen, die auf stationären Computern ausgeführt werden, aus ersichtlichen Gründen auch nicht nötig. Jedoch haben sowohl Gabrielli et al. [GMKC05] als auch Po et al. [PHVS04] den Walkthrough in freier Umgebung durchführen lassen und kamen zu den zuvor beschriebenen Ergebnissen. Keine der Studien erwähnte explizit, dass die Inspektoren Anweisungen bekamen, sich während des Tests auf eine bestimmte Art und Weise zu bewegen. Der von Wharton definierte Input für einen Kognitiven Walkthrough [WRLP94] wäre jedoch leicht durch solche Anweisungen zu ergänzen. Nachteilig ist in diesem Fall allerdings, dass der Usability-Experte die Bewegung, sollte sie aktiv sein, unterbrechen muss, um die gewonnenen Erkenntnis zu notieren oder aufzunehmen. Die Natürlichkeit der Handhabung und Bewegung geht damit möglicherweise verloren. Ein Beispiel, welches dies verdeutlicht, ist eine Applikation, die von einer Person getestet werden soll, die sich in Eile befindet. Dabei könnte sich um eine Anwendung handeln, mit der sich die Öffnungszeiten von in der Nähe gelegenen Supermärkten vergleichen lässt und den Weg dorthin weist. Die Person geht bereits schneller, da der nächste Supermarkt bereits in fünf Minuten schließt, navigiert jedoch nebenbei mit Hilfe der Navigation der Anwendung und sucht möglicherweise bereits nach einer Alternative. Der Usability-Experte, der eine solche Anwendung für eine längere Strecke testen soll, muss möglicherweise zwischendurch stehen bleiben, um detaillierte Notizen über die Usability der Anwendung zu machen. Diese Pause benutzt er womöglich unbewusst, um über die nächsten nötigen Schritte mit der Anwendung nachzudenken. Außerdem dient diese Pause automatisch der Regeneration, so dass die physische Belastung insgesamt geringer ist als bei einer Person, die durchgängig schnell geht oder rennt. Dennoch ist die Berücksichtigung des physischen Zustands durchaus als *stark* einzuschätzen und wird dementsprechend mit vier Punkten bewertet.

Berücksichtigung der natürlichen Nutzungsumgebung

In den Forschungsarbeiten, die den Kognitiven Walkthrough in natürlicher Umgebung durchführen ließen, wurden den Usability-Experten reale Orte genannt, an denen sie die Szenarien durchspielen sollten. Zu diesen Orten gehörten unter anderem der Flur einer Universität, ein Restaurant, ein Fahrstuhl und ein Balkon [GMKC05, PHVS04]. So lassen sich Orte mit verschiedensten Umgebungsvariablen wählen, die für den Test der Anwendung relevant sind. Durch einige dieser Orte entstehen lediglich Einschränkungen in der Wahl des Mediums zum Notieren der gewonnenen Erkenntnisse während des Tests. So lassen sich an sehr lauten Plätzen vermutlich Probleme feststellen, wenn der Usability-Experte ein Diktiergerät verwenden möchte, um seine Gedanken festzuhalten. An anderen Orten sind möglicherweise keine geeigneten Flächen vorhanden, auf denen Notizen auf einem Block oder Zettel gemacht werden können. Ein anderer Nachteil besteht darin, dass die Usability-Experten sich gegebenenfalls mit unbekannten und hinderlichen Bedingungen auseinandersetzen müssen, während sie die Usability der Anwendung testen (vgl. Abschnitt 4.3.1). Gabrielli et al. vermuten, dass diese Unannehmlichkeiten die Anzahl der gefundenen Usability-Probleme negativ beeinflussen [GMKC05]. Da der Kognitive Walkthrough jedoch ohne großem Aufwand an vielen verschiedenen Orten durchgeführt und direkt mit persönlichem Nutzungskontext und Umweltzuständen in Verbindung gebracht werden kann, lässt sich die Berücksichtigung der natürlichen Nutzungsumgebung zusammenfassend als *sehr stark* bewerten.

Berücksichtigung der mobilgerätespezifischen Ein- und Ausgabeprobleme

Sofern bereits ein auf mehreren Endgeräten funktionierender Prototyp einer Anwendung oder eine vollständig funktionierende Anwendung vorhanden ist, lassen sich die Ein- und Ausgabemedien ohne großen Aufwand auf mehreren Geräten testen. Dabei könnte sowohl ein Usability-Experte mehrere verschiedene Endgeräte zum Testen der Anwendung bei sich tragen oder jeder Usability-Experte, der für den Test eingesetzt wird, bekommt ein unterschiedliches Gerät zugewiesen. Es ist zu beachten, dass ersteres Verfahren die Gesamtdauer des Tests verlängert und somit auch die Kosten steigen. Es ist hingegen zu vermuten, dass dadurch auch mehr Usability-Probleme, die in Relation zu den Ein- und Ausgabemedien stehen, entdeckt werden. Außerdem werden so die Ergebnisse der verschiedenen Experten vergleichbarer³⁰. Es lässt sich also feststellen, dass die mobilgerätespezifischen Ein- / und Ausgabeprobleme bei Bedarf *sehr stark* berücksichtigt werden können.

³⁰Entdecken mehrere Experten die gleichen Probleme auf dem gleichen Endgerätetyp, ist es wahrscheinlich, dass auch mehr tatsächliche Benutzer der Anwendung dieses Problem erfahren.

Beeinflussung des natürlichen Nutzungsverhaltens

Das natürliche Nutzungserlebnis des Anwenders mit der mobilen Applikation wird allein durch die Tatsache, dass sich dieser Anwender zusätzlich in der Rolle des Evaluators befindet, gestört. Das Notieren der gewonnenen Erkenntnisse und das Nachdenken über die Usability-Faktoren hindern den Experten daran, die Anwendung vollständig unbeeinflusst zu benutzen. Je nach Nutzungssituation kann dies mehr oder weniger störend sein. Bei dem zuvor genannten Beispiel einer Applikation über Öffnungszeiten ist das Nutzungsverhalten vermutlich *stark* beeinflusst. Bei einer Anwendung, die häufig im Sitzen und ohne Zeitdruck benutzt wird, ist die Beeinflussung hingegen geringer einzuschätzen. Des Weiteren ist es positiv zu bewerten, dass das natürliche Nutzungsverhalten nicht durch Kameras oder andere an dem Test beteiligte Personen gestört wird. Insgesamt lässt sich infolgedessen von einer *mittleren* Beeinflussung des natürlichen Nutzungsverhaltens sprechen.

Gefundene Usability-Probleme

Po et al. [PHVS04] untersuchten den Kognitiven Walkthrough im Nutzungskontext, indem sie Experten die Evaluation in einem Feldversuch durchführen ließen. Die Ergebnisse verglichen sie mit der Heuristischen Evaluation und traditionellen Kognitiven Walkthrough in einer Laborumgebung. Dazu ließen sie elf Usability-Experten eine der Methoden auf einem mobilen Endgerät durchführen³¹ und analysierten die gefundenen Probleme. Bei diesem Versuch deckte die Heuristische Evaluation 7,5 Usability-Probleme auf, während die anderen beiden Methoden im Durchschnitt 19 Probleme fanden. Die gefundenen Probleme wurden in fünf Schweregrade eingeteilt. Bei den Anteilen der Probleme pro Schweregrad lassen sich auch Parallelen zwischen den beiden Walkthrough-Methoden ziehen. So sind bei beiden etwa 10% der gefundenen Probleme Usability-Katastrophen, während die Usability-Experten, die die heuristische Evaluation verwendeten, kein Problem als katastrophal einstufen. Bei näherer Betrachtung der gefundenen Probleme lassen sich jedoch auch Unterschiede zwischen dem Kognitiven Walkthrough im Labor und dem Walkthrough in der natürlichen Nutzungsumgebung feststellen. Unter Verwendung letzterer Methode wurden mehr Probleme, die sich auf Eingabemedium, Lichtverhältnisse und Verarbeitungsgeschwindigkeit des Endgerätes bezogen, gefunden. Ein weiterer Vorteil, den einer der Usability-Experten nannte, war, dass es für ihn in der natürlichen Umgebung einfacher war, sich in die Lage des Endnutzers zu versetzen. Die Autoren dieser Forschungsarbeit waren jedoch auch vor der Allgemeingültigkeit des Ergebnisses, da die Standardabweichung (SD) der Anzahl der gefundenen Probleme hoch ist und signifikant von dem eingesetzten Usability-Experten abhängt [PHVS04].

³¹Jeweils vier Experten verwendeten die Heuristische Evaluation und den Kognitiven Walkthrough, während drei den Walkthrough im Nutzungskontext durchführten.

Eine ähnliche Studie von Gabrielli et al. [GMKC05] vergleicht den traditionellen Kognitiven Walkthrough mit dem Walkthrough im Kontext und einem Kognitiven Walkthrough, der mit Videoaufzeichnungen unterstützt wird. Bei dem Walkthrough mit Videoaufzeichnung benutzen die Usability-Experten das Endgerät nicht selbst, sondern analysieren die Aufzeichnungen, die Testpersonen im Umgang mit der mobilen Anwendung zeigen. Die Testpersonen wurden außerdem dazu aufgefordert, während der Nutzung *laut zu denken*. Dieses Verfahren ist also der Think Aloud Methode sehr ähnlich und beinhaltet nicht mehr das aktive Benutzen der Applikation durch den Usability-Experten [GMKC05]. Die drei verschiedenen Methoden wurden von neun Usability-Experten angewandt, wobei jeweils drei zufällig den Verfahren zugeteilt wurden. Als Ergebnis wurde festgestellt, dass der Kognitiven Walkthrough im Nutzungskontext und die Think Aloud Methode im Durchschnitt zwölf Usability-Probleme aufdeckten, während mit dem traditionellen Kognitiven Walkthrough nur durchschnittlich 7,3 Probleme gefunden wurden. Der Anteil der Probleme, die sich auf den Nutzungskontext beziehen, lag jedoch bei allen drei Verfahren bei circa 75%. Wobei dieser Wert bei dem im Feld stattfindenden Kognitiven Walkthrough (CW) mit 72% am niedrigsten ist. Gabrielli et al. haben zwei mögliche Erklärungen dafür:

- „a) the capacity of CW method to prompt evaluators towards in depth analysis of the learnability of use of a specific application for its users;*
- b) the difficulties participants assigned to this condition found in accurately reporting on their protocols the problems experienced in situ, due to the somewhat unfamiliar and cumbersome conditions in which they were ask to conduct the evaluation“* [GMKC05]

Möglicherweise sind diese Werte jedoch auch auf die unterschiedlichen Usability-Experten, deren Erfahrung und Fähigkeiten zurückzuführen. Die Standardabweichung der gefundenen Usability-Probleme liegt bei den beiden Kognitiven Walkthroughs bei 2,88 bzw 3,46 Problemen. Der Unterschied in der Anzahl der gefundenen Probleme der Usability-Experten bei der Methode mit Videoaufzeichnung hingegen war wesentlich geringer³².

Gesamtbewertung

Ein großer Vorteil des Kognitiven Walkthroughs, der schon für das Testen von Anwendungen auf stationären Endgeräten bestand, ist die Möglichkeit, ihn schnell und günstig durchzuführen. Der Hauptgrund dafür ist, dass er ohne den Einsatz von Testpersonen auskommt. Die Variation dieses Verfahrens, den Kognitiven Walkthrough im Kontext durchzuführen, erweitert die positiven Eigenschaften der Methode noch um den Aspekt, dass dadurch auch die natürliche Nutzungsumgebung berücksichtigt wird. Diese Vorteile werden auch in Abbildung 4.5,

³²SD = 0,57 Probleme.

der Übersicht über die vorgenommene Bewertung, sichtbar. Hier wird jedoch auch die größte Schwäche des Verfahrens deutlich. Die Tatsache, dass keine Testpersonen an dem Verfahren teilnehmen, die keine Vorkenntnisse über Usability haben, sondern zur Zielgruppe der Anwendung gehören, erschwert die Analyse, wie zufrieden die Nutzer dieser Anwendung sein werden. Es können keine Informationen aus den direkten Reaktionen dieser Art Testnutzer gewonnen werden. Somit ist die Qualität des Testergebnisses sehr stark von dem Können und der Empathiefähigkeit des Usability-Experten abhängig.

Zeitaufwand	●●●●●
Kosten	●●●●●
Zeitpunkt der Durchführung	●●●●●
Berücks. der Usability-Attribute	●●●●○
Berücks. des emotionalen Zustands	●○○○○
Berücks. des physischen Zustands	●●●●○
Berücks. der natürlichen Nutzungsumgebung	●●●●●
Berücks. der mobilgerätespez. Ein-/Ausgabeprobleme	●●●●●
Beeinflussung des natürlichen Nutzungsverhaltens	●●○○○

Abbildung 4.5: Übersicht über die Bewertung des Kognitiven Walkthroughs

4.4 Heuristische Evaluation

„Heuristic Evaluation is the most informal method and involves having usability specialists judge whether each dialogue element follows established principles (the heuristics).“ [Nie94b]

Nach dieser Definition der Heuristischen Evaluation sind drei Komponenten für die Durchführung dieser Evaluationsmethode notwendig: der Usability-Spezialist, die zu analysierenden Elemente und die Heuristiken. Da die Heuristische Evaluation keine Testnutzer einsetzt, sondern ausschließlich durch die Usability-Experten durchgeführt wird, gilt sie, wie der Kognitive Walkthrough, als Usability-Inspektion (vgl. Abschnitt 4.3).

Bei der Heuristischen Evaluation ist es üblich, dass der Experte die gefundenen Usability-Problemen während der Evaluation auf einer beliebigen Skala hinsichtlich ihrer Dringlichkeit gewichtet. Es gibt verschiedene Sammlungen von Heuristiken, die bei der Evaluation der Usability von Anwendungen eingesetzt werden können. Zum einen solche, die für jede Anwendung verwendet werden können. Zum anderen solche, die speziell für eine Art von Anwendungen erarbeitet wurden [BTT05]. Eine Liste von zwölf allgemeinen Heuristiken stellt David Benyon in dem Buch „*Designing Interactive Systems*“ vor [BTT05, S. 90 & S. 228]. Außerdem beschreibt Jakob Nielsen in seinem Buch „*Usability-Engineering*“ zehn eigene Usability-Heuristiken [Nie94c, S. 115 ff.]. Diese Heuristiken wurden von Bertini et al. hinsichtlich ihrer Tauglichkeit für mobile Endgeräte und mobile Anwendungen untersucht, um anschließend eine

neue Sammlung von Heuristiken für das Usability-Testing von mobilen Endgeräten und Applikationen zu erarbeiten [BCD⁺08]. Diese Heuristiken (von Bertini et al. als *mobile Heuristiken* bezeichnet) sind Gegenstand der folgenden Beschreibung und Bewertung der Heuristischen Evaluation von mobilen Anwendungen.

4.4.1 Beschreibung

Bertini et al. entwickelten die *mobilen Heuristiken* in drei Phasen. In der ersten Phase untersuchten drei der Autoren, die an der Forschungsarbeit beteiligt waren, unabhängig voneinander die Heuristiken von Nielsen [Nie94c]. Dabei wurde erarbeitet, welche der Heuristiken irrelevant für die Charakteristiken der mobilen Geräte und Anwendungen waren und welche relevant waren, aber überarbeitet werden mussten. Außerdem wurden zusätzliche Heuristiken aufgestellt, die die relevanten Aspekte von mobilen Endgeräten und Anwendungen abdecken sollten. In der zweiten Phase glichen jeweils zwei Forscher ihre in der ersten Phase entstandene Tabelle der Heuristiken mit einem anderen teilnehmenden Autor ab. Dadurch sollte der Prozess zur Verbesserung der Heuristiken hinsichtlich ihrer Deutlichkeit und Relevanz bezüglich der Eigenschaften mobiler Geräte und Anwendungen beschleunigt werden. Die auf diese Weise verbesserten Tabellen wurden in einem gemeinsamen Treffen zu einer Sammlung von Heuristiken zusammengefügt, die anschließend an neunzehn Usability-Experten weitergeleitet wurden, um Resonanz einzuholen. Acht der kontaktierten Experten gaben Feedback zu den erarbeiteten Heuristiken. Diese Kommentare wurden von den Autoren genutzt, um den nachfolgenden finalen Satz von acht Heuristiken für mobile Endgeräte sowie mobile Anwendungen zu konzipieren [BCD⁺08].

- **Heuristik 1 - Visibility of the system status and losability/findability of the mobile device:** Der Benutzer muss immer darüber informiert sein, wie der Status seines mobilen Endgerätes ist. Dabei sollten wichtige Nachrichten über Informationen wie der Status der Batterie oder des Netzwerks vom System priorisiert werden. Des Weiteren sollte, da mobile Endgeräte häufig verloren gehen, entsprechend Anpassungen, wie Verschlüsselung der Daten, innerhalb des Systems oder der Anwendung vorgenommen werden. Wenn das Gerät verlegt wurden, sollte das Gerät selbst, das System oder die Anwendung das Wiederfinden ermöglichen.
- **Heuristik 2 - Match between system and the real world:** Dem Benutzer muss es möglich sein, die von der Anwendung gegebenen Informationen korrekt zu interpretieren. Dies geschieht, indem die Informationen in einer natürlichen und logischen Reihenfolge präsentiert werden. Wenn es möglich ist, sollte das System die Nutzungsumgebung wahrnehmen und die Präsentation entsprechend anpassen.

- **Heuristik 3 - Consistency and mapping:** Das gedankliche Modell des Benutzers über die möglichen Funktionen und Interaktionen mit der Anwendung sollten konsistent mit dem Kontext sein. Dies ist besonders wichtig, wenn eine Zuordnung der Benutzeraktionen auf dem Endgerät mit der tatsächlichen Tätigkeit möglich ist und sich in Begrifflichkeiten widerspiegelt. Ein Beispiel dafür ist eine Anwendung zur Navigation.
- **Heuristik 4 - Good ergonomics and minimalist design:** Mobile Endgeräte sollten einfach und komfortabel zu halten und zu transportieren sein. Des Weiteren sollte das Gerät robust konstruiert sein, um möglichen Einflüssen von Außen standzuhalten. Da der Bildschirmplatz der mobilen Geräte nicht sehr groß sind, sollten mit ihm sparsam umgegangen werden. Dialoge sollten dementsprechend keine Informationen beinhalten, die selten oder nie benötigt werden.
- **Heuristik 5 - Ease of input, screen readability, and glanceability:** Mobile Systeme sollten einen einfachen Weg zu Dateneingabe ermöglichen und möglichst das Benötigen von zwei Händen zur Eingabe verhindern. Es sollte einfach sein, durch Inhalte zu navigieren und sie zu lesen, unabhängig der vorliegenden Lichtverhältnisse. Idealerweise sollte der Benutzer die wichtigsten Informationen erlangen können, wenn er nur einen kurzen Blick auf das System wirft.
- **Heuristik 6 - Flexibility, efficiency of use, and personalization:** Dem Benutzer muss es möglich sein, häufige Aktionen zu personalisieren oder anzupassen. Er sollte außerdem das System manuell an die äußeren Gegebenheiten anpassen können. Falls möglich, sollte das System vorteilhafte Anpassungen unterstützen und vorschlagen.
- **Heuristik 7 - Aesthetic, privacy, and social conventions:** Ästhetische und emotionale Aspekte des Endgeräte und mobilen Systems sollten berücksichtigt werden. Die Benutzerdaten müssen privat und sicher sein. Die Interaktion mit dem System sollte komfortabel sein und gesellschaftliche Konventionen respektieren.
- **Heuristik 8 - Realistic error management:** Wenn ein Fehler auftritt, soll dem Benutzer geholfen werden, den Fehler zu erkennen, zu diagnostizieren und, falls möglich, Unterstützung bei der Lösungsfindung bieten. Die Fehlermeldungen müssen klar und präzise sein. Wenn es keine Lösung zu dem Problem gibt oder der Fehler zu vernachlässigende Auswirkungen hat, sollte dem Benutzer dabei geholfen werden, mit dem Problem zurechtzukommen.

Da die Heuristiken sowohl für die Heuristische Evaluation von mobilen Endgeräten als auch für die von mobilen Anwendungen entwickelt wurden, sind nicht alle Inhalte für das Usability-Testing von mobilen Applikationen relevant. Besonders die Heuristiken bezüglich der Transportfähigkeit und Wiederauffindbarkeit der Endgeräte können im Kontext dieser Arbeit vernachlässigt werden.

4.4.2 Bewertung

Die Bewertung der Heuristiken wird anhand der Forschungsarbeit „*Appropriating Heuristic Evaluation Methods for Mobile Computing*“ von Bertini et al. [BCD⁺08] vorgenommen. Dabei wird angenommen, dass grundsätzliche Eigenschaften der Heuristiken, wie die benötigte Anzahl der Usability-Experten, Bestand haben, sofern es von Bertini et al. nicht anders erwähnt wurde. Ihre Arbeit umfasst, außer dem Entwickeln der mobilen Heuristiken, auch einen Usability-Test von zwei mobilen Anwendungen mit Hilfe der erarbeiteten Heuristiken und der ursprünglichen Heuristiken von Nielsen [Nie94c, S. 115 ff.]. Da bisher keine anderen Ergebnisse von Usability-Tests, die mit den mobilen Heuristiken von Bertini et al. durchgeführt wurden, veröffentlicht sind, werden die dort gewonnenen Daten als Richtwert für die Bewertung der Heuristischen Evaluation als Usability-Testmethode für mobile Anwendungen verwendet.

Zeitaufwand

Die Zeit, die zur Durchführung der Heuristischen Evaluation benötigt wird, ist abhängig von der zu testenden Anwendung und der Erfahrung der Usability-Experten. Eine generelle Zeitbeschränkung ist demnach nicht möglich. Bertini et al. ließen die mobilen Heuristiken von vier Usability-Experten gleichmäßig aufgeteilt auf zwei verschiedene Applikationen anwenden. Ein Experte benötigte für diese Aufgabe im Durchschnitt 72,75 Minuten. Die Standardabweichung von 44,776 Minuten ist jedoch ein Indiz für die hohe Korrelation zwischen der Erfahrung und Expertise der eingesetzten Usability-Spezialisten und der Dauer der Heuristischen Evaluation³³ [BCD⁺08]. Die Heuristische Evaluation kann, falls genügend Ressourcen verfügbar sind, von den Usability-Experten gleichzeitig durchgeführt werden, so dass sich die Länge des Usability-Tests, egal wie viele Experten eingesetzt werden, nicht verändert. Es ist also anzunehmen, dass die Heuristische Evaluation einer mobilen Anwendung nicht länger als zwei Stunden dauert. Dementsprechend kann der Zeitaufwand dieser Methode mit fünf Punkten bewertet werden.

Kosten

Um mit der Heuristischen Evaluation 75-80% der Usability-Probleme zu entdecken, werden drei bis fünf Usability-Experten benötigt [BCD⁺08, PHVS04]. Da keine Testnutzer in Anspruch genommen werden, muss also lediglich der Stundenlohn für die Experten berücksichtigt werden. Bei einer Vergütung von 32,69€ pro Stunde (vgl. Abschnitt 3.1.1) und einer Testdauer von 72,75 Minuten entstehen demnach Kosten von 118,91€ bis 198,18€. So sind über

³³Dabei ist zu beachten, dass die Standardabweichung zusätzlich durch das Testen von zwei verschiedenen Applikationen, die bereits unterschiedliche durchschnittliche Zeitspannen zur Evaluation benötigten, beeinflusst wird. Außerdem ist die relativ kleine Stichprobe der Größe vier (evaluierende Personen) zu berücksichtigen.

800€ verfügbar, um Aufnahmegeräte und Medien für die Usability-Experten zu finanzieren, auf denen die gefundenen Probleme aufgezeichnet werden, und trotzdem lägen die Kosten der Heuristischen Evaluation noch unter 1000€. Die Kosten der Heuristischen Evaluation können in diesem Fall also mit fünf Punkten bewertet werden. Diese Bewertung entspricht auch der Bezeichnung der Heuristischen Evaluation als eine der „*discount usability engineering methods*“ [Nie94a].

Zeitpunkt der Durchführung

Es gehört zu den Stärken der Heuristischen Evaluation, dass zu seiner Durchführung kein funktionierender Prototyp verfügbar sein muss. Die Heuristiken lassen sich auch auf Modellen, Mock-Ups oder Skizzen anwenden [BCD⁺08]. Dazu müssen jedoch entsprechende Informationen über das Verhalten der Anwendung vorliegen, da die Anwendung sonst nicht hinsichtlich Heuristiken wie dem *Realistic Error Management* geprüft werden können. Dennoch ist es grundsätzlich möglich, die mobilen Heuristiken in jeder Phase der Entwicklung anzuwenden [BCD⁺08].

Berücksichtigung der Usability-Attribute

Da die Heuristische Evaluation durch Usability-Experten durchgeführt wird und keine Personen ohne Hintergrundwissen zu Usability und Mensch-Computer-Interaktion involviert sind, ist die Berücksichtigung der Usability-Attribute ausschließlich abhängig von den Fähigkeiten der Experten. Es wird an dieser Stelle bewertet, ob die Heuristiken die entsprechenden Usability-Attribute in irgendeiner Form berücksichtigen. Da jedoch keine Metriken (vgl. Abschnitt 2.1.5) angewandt werden, ist die tatsächliche Qualität eines Usability-Attributs nicht messbar. Dadurch ist es auch nicht möglich, bei einer Veränderung der Anwendung und eines erneuten Usability-Tests die Verbesserung / Verschlechterung der Attribute, wie beispielsweise der Effizienz, zu messen und auszudrücken. Die durch die Heuristische Evaluation berücksichtigten Attribute werden aus diesem Grund in der Gesamtbewertung dieses Kriteriums nur als zur Hälfte erfüllt gewertet.

Learnability / Erlernbarkeit:

Die direkte Messung der Erlernbarkeit durch die Metriken *Task Success* und *Time-on-task* wird durch die mobilen Heuristiken nicht unterstützt. Die Heuristiken 2 (Match between system and the real world), 3 (Consistency and mapping) und 8 (Realistic error management) sollen das Erlernen der grundlegenden Funktionen einer Anwendung jedoch indirekt unterstützen, indem sie sicherstellen, dass die Anwendung intuitiv bedienbar wird und dem Benutzer bei Fehlereingaben geholfen wird, sein Ziel dennoch zu erreichen.

Efficiency / Effizienz:

Die Effizienz einer Anwendung kann hauptsächlich durch die Beachtung der Heuristiken 2 (Match between system and the real world), 3 (Consistency and mapping) und 6 (Flexibility, efficiency of use, and personalization) gewährleistet werden. Insbesondere Heuristik 6, die das Personalisieren von häufig verwendeten Funktionen empfiehlt, ist für die Effizienz einer mobilen Anwendung von Bedeutung.

Memorability / Einprägsamkeit:

Es lässt sich argumentieren, dass Konsistenz innerhalb der Anwendung und vor allem das Anpassen von Begriffen der Benutzeroberfläche an Wörter, die dem Benutzer in ähnlichem Kontext aus der realen Welt bekannt sind (Heuristik 3), dabei helfen, die Einprägsamkeit zu verbessern. Die Verwendung von Metaphern ist jedoch ein allgemeines Mittel, um das Erlernen einer Anwendung einfacher zu gestalten [Nie94c, S. 126 ff.], und somit liegt keine direkte Heuristik zur Verbesserung der Einprägsamkeit einer mobilen Anwendung vor.

Errors / Fehler:

Heuristik 8 dient dazu, dass der Benutzer einen von ihm begangenen Fehler erkennt und ihm bei der Problemlösung hilft. Die Heuristiken 2 (Match between system and the real world), 3 (Consistency and mapping) und 5 (Ease of input, screen readability, and glanceability) hingegen sollen sicherstellen, dass möglichst wenig Fehler begangen werden, in dem sie der Vereinfachung der allgemeinen Handhabung der Applikation dienen.

Satisfaction / Zufriedenheit:

Die Zufriedenheit der Nutzer entsteht schlussendlich aus einem positiven Empfinden während der Bedienung. Alle Heuristiken sind indirekt dafür verantwortlich, dass dem Anwender die Handhabung der Applikation so angenehm wie möglich gemacht wird. Im Speziellen wird in Heuristik 7 (Aesthetic, privacy, and social conventions) darüber aufgeklärt, dass die Interaktion mit der Anwendung so komfortabel wie möglich gestaltet und ästhetische Aspekte berücksichtigt werden sollen, um die Zufriedenheit des Anwenders zu gewährleisten.

Zusammenfassung:

Vier der fünf Usability-Attribute werden durch die Heuristiken indirekt berücksichtigt. Durch den zuvor beschriebenen Mangel an eingesetzten Metriken ergibt sich daraus eine Bewertung von zwei Punkten.

Berücksichtigung des emotionalen Zustands

Der emotionale Zustand der Benutzer wird durch die Heuristische Evaluation nicht direkt berücksichtigt. Es ist zwar wahrscheinlich, dass ein Benutzer die Anwendung benutzen kann, wenn die Heuristiken angewandt werden, ob ihm die Benutzung jedoch gefällt, ist dadurch nur begrenzt zu bestimmen. Heuristik 7 ist dazu ausgelegt, sicherzustellen, dass die Interaktion mit der Anwendung komfortabel und ästhetische Aspekte berücksichtigt werden. Diese Faktoren

sind entscheidend für das Empfinden des Nutzers bei der Bedienung der Anwendung. Durch ihren subjektiven Charakter ist die Expertise des Usability-Spezialisten jedoch entscheidend darüber, ob seine Wahrnehmung repräsentativ für den emotionalen Zustand eines Nutzers der Applikation ist. Aus diesen Gründen ist die Berücksichtigung des emotionalen Zustands durch die Heuristische Evaluation nur *sehr schwach* vorhanden.

Berücksichtigung des physischen Zustands

Nach Bertini et al. ist Heuristik 5 (Ease of input, screen readability, and glanceability) speziell dafür ausgelegt, die besonderen Umstände der mobilen Nutzung zu berücksichtigen. In der Heuristik heißt es, dass wichtige Informationen mit einem kurzen Blick auf das Gerät ersichtlich werden sollen. Außerdem soll ein einfacher Weg zur Dateneingabe gewährleistet werden, um die Anwendung beispielsweise während des Gehens oder während einer Unterrichtsstunde unter dem Tisch bedienen zu können [BCD⁺08]. Da der Usability-Experte diese Aktivitäten allerdings nicht tatsächlich ausprobiert, ist die Evaluation durch diese Heuristik erneut auf Wissen, Erfahrung und der Fähigkeit, sich in die entsprechende Lage hineinzusetzen, des Usability-Experten angewiesen. Daraus folgt, dass die Heuristische Evaluation nur eine *mittlere* Berücksichtigung des physischen Zustands zulässt.

Berücksichtigung der natürlichen Nutzungsumgebung

Mehrere der mobilen Heuristiken verfolgen den Zweck, den natürlichen Nutzungskontext zu berücksichtigen. Dabei ist das soziale Umfeld ein Bestandteil der Heuristik 7 (Aesthetic, privacy, and social conventions) [BCD⁺08]. Heuristik 6 (Flexibility, efficiency of use, and personalization) hingegen schreibt vor, dass der Nutzer in der Lage sein soll, die Anwendung an die durch die Nutzungsumgebung entstehenden Anforderungen anzupassen. Heuristik 2 (Match between system and the real world) besagt, dass das Anwendungssystem, falls möglich, die Umgebung wahrnehmen und sich daran anpassen soll [BCD⁺08]. Die mobilen Applikationen sollen also Gebrauch von den verschiedenen Sensoren der Endgeräte (vgl. Abschnitt 2.2.2) machen, um die Darstellung des Inhalte entsprechend anzupassen. Die genannten Heuristiken geben demnach vor, dass sich die mobilen Anwendungen die Nutzungsumgebung zur Präsentation der Daten und zum Einhalten von Sozialkonventionen berücksichtigen sollen. Wie gut sich eine Anwendung an Vorgaben dieser Heuristiken hält, liegt jedoch allein im Ermessen des Usability-Experten, ohne dass dieser die Anwendung im tatsächlichen Nutzungskontext benutzt. Ein Beispiel, wann dies zum Problem werden kann, ist, wenn der Usability-Spezialist feststellt, dass sich die Darstellung der Inhalte automatisch an die Lichtverhältnisse anpasst, diese Anpassung jedoch in einer natürlichen Umgebung bei sich schnell ändernden Lichtverhältnissen, wie etwa beim Fahren durch eine Allee, zum Flackern des Bildschirms führt. Dennoch ist die

Berücksichtigung der natürlichen Nutzungsumgebung durch die mobilen Heuristiken durchaus gegeben und wird mit drei Punkten bewertet.

Berücksichtigung der mobilgerätespezifischen Ein- und Ausgabeprobleme

Sofern die Heuristische Evaluation mit einem funktionierendem Prototypen auf einem mobilen Endgerät stattfindet, lässt sich das Endgerät leicht durch eines mit anderen Ein- / oder Ausgabemedien ersetzen. Die Heuristiken 4 (Good ergonomics and minimalist design) und 5 (Ease of input, screen readability, and glanceability) sind für die Ein- und Ausgabe der Daten relevant und empfehlen zum Beispiel den sorgsamsten Umgang mit dem vereinnahmten Platz auf dem Bildschirm und einen einfachen Weg Daten einzugeben, ohne dabei beide Hände zu benutzen. Werden also die Endgeräte mit den verschiedenen Medien hinsichtlich der genannten Heuristiken überprüft, ist eine *sehr starke* Berücksichtigung der mobilgerätespezifischen Ein- und Ausgabeprobleme möglich.

Beeinflussung des natürlichen Nutzungsverhaltens

Die Usability-Experten werden bei der Verwendung beziehungsweise Evaluation der mobilen Anwendungen nicht durch am Gerät montierte Kameras oder beobachtende Personen beeinträchtigt. Jedoch befindet sich auch hier der Usability-Experte in der Rolle des Nutzers und kann das natürliche Nutzungsverhalten nur simulieren. Aufgrund des gleichzeitigen Nutzens und Bewertens findet insgesamt eine mittlere Beeinflussung des natürlichen Nutzungsverhaltens statt.

Gefundene Usability-Probleme

Durch die Arbeit von Bertini et al. [BCD⁺08] ist lediglich ein Vergleich der Ergebnisse der mobilen Heuristiken mit denen der Heuristiken von Nielsen [Nie94c, S.115 ff.] verfügbar. Bei den zwei auf diese Weise getesteten Applikationen wurde mit Hilfe der mobilen Heuristiken 64 Usability-Probleme identifiziert, während die Usability-Experten mit Nielsens Heuristiken 50 Probleme aufdeckten. Dabei fällt auf, dass über 43 der Probleme, die mit den mobilen Heuristiken gefunden wurden, von den Experten als bedeutend oder unbedeutend einstufen. Nur jeweils etwa zehn Probleme wurden hingegen als kosmetisch oder katastrophal gewertet. Dieses Ergebnis stimmt mit früheren Beobachtungen überein, die besagen, dass mobile Heuristiken nur wenige kosmetische Probleme aufdecken und aus diesem Grund hauptsächlich in frühen Entwicklungsphasen eingesetzt werden sollten. Des Weiteren empfehlen Bertini et al. die Heuristische Evaluation nicht als Alternative zu Usability-Tests mit Benutzerstudien zu betrachten, sondern als Ergänzung dazu [BCD⁺08].

Gesamtbewertung

Abbildung 4.6 offenbart die deutlichen Nachteile der Heuristischen Evaluation, die dadurch entstehen, dass ein Usability-Experte die Rolle des Nutzers begeben muss und sich dabei an festgeschriebene Strukturen hält. Die Strukturen sind jedoch nicht sehr detailliert beschrieben. So besagt eine Heuristik zwar, dass die Interaktion mit der Anwendung komfortabel und ästhetische Aspekte berücksichtigen soll. Es liegt jedoch in der Empfindung des Experten, einzuschätzen, wie komfortabel oder ästhetisch die Nutzerschaft eine Anwendung letztendlich findet. Somit können vermeintliche Emotionen und natürliches Verhalten der potenziellen Nutzer mit der Heuristischen Evaluation nur begrenzt berücksichtigt werden, was sich auch in der Bewertung ausdrückt. Es ist jedoch positiv zu bemerken, dass die Heuristische Evaluation sehr schnell und kosteneffizient durchführbar ist.

Zeitaufwand	●●●●●
Kosten	●●●●●
Zeitpunkt der Durchführung	●●●●●
Berücks. der Usability-Attribute	●●○○○
Berücks. des emotionalen Zustands	●○○○○
Berücks. des physischen Zustands	●●●○○
Berücks. der natürlichen Nutzungsumgebung	●●●○○
Berücks. der mobilgerätespez. Ein-/Ausgabeprobleme	●●●●●
Beeinflussung des natürlichen Nutzungsverhaltens	●●○○○

Abbildung 4.6: Übersicht über die Bewertung der Heuristischen Evaluation

4.5 Interview

Das Interview im Sinne der Usability-Testmethoden gehört zu der Gruppe der Usability-Befragungen. Bei den Usability-Befragungen werden die Nutzer dazu aufgefordert, ihre Meinung über die Anwendung und ihre Usability abzugeben. Dies kann sowohl verbal als auch schriftlich erfolgen. In diesem Abschnitt wird zunächst das Interview beschrieben und bewertet. Eine ähnliche Variante sind die sogenannten Fokusgruppen. Im Gegensatz zu dem Interview werden dabei mehrere Benutzer gleichzeitig befragt und diskutieren gemeinsam über die mobile Anwendung, ihre Funktionen und die vom Moderator gestellten Fragen. Nielsen nennt geringe Validität und die Schwierigkeit, die Ergebnisse auszuwerten, als größte Nachteile dieser Methode [Nie94c, S. 209 ff.]. Unabhängig davon, welche Usability-Befragungsmethode angewandt wird, empfiehlt Holzinger, dass sie in Verbindung mit direkten Usability-Tests, zum Beispiel der Think Aloud Methode, durchgeführt werden [Hol05].

Nach Bernhaupt et al. eignen sich die Usability-Befragungen, die in frühen Entwicklungsstadien eingesetzt werden, wie Beobachtung, Interviews und Workshops, nicht für das Testen

der Usability von mobilen Anwendungen. Diese Verfahren lassen sich zwar verwenden, wenn die Benutzung der Anwendung strukturiert ist und sich auf einige wenige (Arbeits-)Kontexte beschränkt, für komplexe Anwendung sei es jedoch unbrauchbar [BMO08]. Bernhaupt et al. schreiben dazu:

„[...] users have difficulties recalling their activities, needs, and preferences since the private or mobile contexts are not as well structured as their work counterparts.“ [BMO08]

Um die Usability-Befragung und die Forschung hinsichtlich der Gebrauchstauglichkeit dieser Methodengruppe zu berücksichtigen, werden Interviews und Fragebögen in dieser Arbeit dennoch untersucht. Sie gelten als effizienteste Varianten zur Usability-Befragung [TA08, Nie94c].

4.5.1 Beschreibung

Das Interview wird von einem Moderator durchgeführt, der sich die zu stellenden Fragen im Vorfeld überlegt und aufgeschrieben hat. Es unterscheidet sich in dieser Hinsicht nicht von einem gewöhnlichen Interview. Der Interviewpartner ist in diesem Fall ein Nutzer, der die Anwendung oder einen Prototypen der Anwendung bereits ausprobiert hat. Dabei ist zu beachten, dass die Antworten der Testperson nützlicher werden, wenn sie die Anwendung kurz vor dem Interview testen konnte. Damit das Interview Aussagekraft besitzt, sollten mindestens fünf Personen befragt werden [Nie94c, S. 209 ff.].

Bei dem Interview über die Usability einer Anwendung gilt es, die Meinung der Testperson über vorher festgelegte Schwerpunkte zu erfassen. So kann ermittelt werden, was dem Benutzer an der Anwendung und der Benutzeroberfläche gefällt und was ihn stört. Dabei ist es wichtig, dass viele offene Fragen gestellt werden, die der Nutzer nicht mit *ja* oder *nein* beantworten kann. Außerdem muss sich der Moderator nicht strikt seine erarbeiteten Fragen durcharbeiten, sondern kann auf für ihn interessant wirkende Antworten des Interviewpartners eingehen und Folgefragen stellen. Daraus resultiert jedoch, dass die Ergebnisse der Interviews schwieriger zu vergleichen sind. Eine Frage, die möglichst immer gestellt werden sollte, ist die Frage nach kritischen Problemen (critical incidents), auf der Benutzer eventuell gestoßen ist. Des Weiteren ist es wichtig, dass der Moderator keine Zustimmung oder Ablehnung gegenüber den Antworten der Testperson zeigt. Er sollte auch nicht versuchen, dem Benutzer zu erklären, warum sich die Anwendung in einer bestimmten Weise verhält, auch wenn sich der Nutzer stark über etwas beklagt. Nur so ist gegeben, dass der Moderator weiterhin unvoreingenommene Antworten erhält [Nie94c, Hol05]. Unter Umständen ist es dennoch problematisch, die Antworten der Interviewpartner richtig zu interpretieren. Der Grund dafür besteht darin, dass Personen dazu neigen Antworten zu geben, von denen sie denken, dass sie erwartet werden. Die gilt speziell

für Antworten, die sie für peinlich oder inakzeptabel halten [Nie94c]. In diesem Sinne könnte es jemandem zum Beispiel unangenehm sein, die Wahrheit zu sagen, wenn er gefragt wird, wie lange er für die Bedienung einer bestimmten Funktion gebraucht hat.

4.5.2 Bewertung

Es folgt die Bewertung des Interviews als Usability-Testmethode. Hierbei wurden weitestgehend Annahmen getroffen, da noch keine Arbeit veröffentlicht wurde, die das Usability-Testing von mobilen Anwendungen mit Hilfe des Interviews durchführt oder analysiert. Es wird davon ausgegangen, dass ein Testmoderator zur Verfügung steht, der fünf Interviews mit fünf verschiedenen Testpersonen durchführt. Die Testpersonen hatten zuvor die Möglichkeit, die Anwendung oder den Prototypen zu testen. Diese Testphase kann sowohl in geschlossenen Räumen, wo anschließend das Interview durchgeführt wird, stattfinden oder aber dem Nutzer wird mehr Zeit mit der Anwendung gewährt, so dass er sie in der natürlichen Nutzungsumgebung ausprobieren kann. Es ist auch denkbar, dass das Interview erst durchgeführt wird, wenn die Anwendung bereits auf dem Markt ist. In diesem Fall könnten tatsächliche Anwender eingeladen werden, die bereits einige Erfahrung mit der Applikation gesammelt haben.

Zeitaufwand

Unter der Annahme, dass die Befragung einer Person nicht mehr Zeit als zwei Stunden in Anspruch nimmt und somit innerhalb der von Dumas und Loring vorgeschlagenen maximalen Länge³⁴ für einen Usability-Test mit einem Moderator und einem Befragten liegt [DL08, S. 2], würde die Befragung von fünf Teilnehmern etwa zehn Stunden in Anspruch nehmen. Dabei ist nur die Zeit berücksichtigt, die direkt für das Interview aufgebracht werden muss. Die Dauer der Testphase für die Befragten kann je nach Anwendung / Aufgabenstellung stark variieren. Die Teilnehmer können jedoch, sofern genügend Testobjekte (Endgeräte, Prototypen, Mock-Ups) zur Verfügung stehen, gleichzeitig mit der Anwendung interagieren. Die Dauer des Tests und Interviews würde demnach vermutlich zwei Tage nicht überschreiben. Daraus folgt eine Bewertung des Zeitaufwands des Interviews von einem Punkt³⁵.

³⁴Die von Dumas und Loring vorgeschlagene maximale Länge für ein Interview beträgt exakt zwei Stunden [DL08, S. 2].

³⁵Auch wenn viele Vermutungen angestellt werden mussten, um den Zeitaufwand zu bewerten, scheint das Ergebnis die Tatsachen widerzuspiegeln. Jakob Nielsen bezeichnet die benötigte *usability staff time* und den allgemeinen Zeitaufwand als große Mankos des Interviews [Nie94c, S. 210 & S. 224].

Kosten

Durch die in Abschnitt 3.1.1 ermittelten Kosten in Höhe von 123,4€ pro Testperson und 32,69€ pro Stunde für einen Usability-Experten als Moderator lassen sich folgende Gesamtkosten berechnen:

$$\text{Gesamtkosten} = 5 * 123,4\text{€} + 10h * 32,69\text{€/h} = 943,9\text{€}.$$

Die Kosten des Interviews werden somit noch mit fünf Punkten bewertet. Dabei ist jedoch zu bedenken, dass weitere Kosten für die Analyse der Interviewaufzeichnungen entstehen.

Zeitpunkt der Durchführung

Die Befragung des Benutzers kann, mit entsprechenden Einschränkungen, in jeder Phase der Anwendungsentwicklung stattfinden. Die Einschränkungen bestehen jedoch darin, wie viel Informationen über die vom Nutzer wahrgenommene Usability aus dem Interview gewonnen werden können. Ein Benutzer, der nur eine Zeichnung der Benutzeroberfläche vorgelegt bekommt, wird in diesem Sinne nichts über die Performance der Anwendung aussagen können, und ein Interviewpartner, der die Anwendung an einem Emulator getestet hat, wird beispielsweise keine Auskünfte über die mobilegerätespezifischen Ein- und Ausgabeprobleme geben können. Da es aber theoretisch möglich ist, das Interview zu jedem Zeitpunkt stattfinden zu lassen, wird dieses Kriterium mit fünf Punkten bewertet.

Berücksichtigung der Usability-Attribute

Während beim Think Aloud Test die Meinung der Nutzer über die Usability noch weitestgehend unberücksichtigt bleibt, wird bei einem Interview der Fokus eben auf diese Wahrnehmung der Testpersonen gelegt. Daher ist der Grad der Berücksichtigung der Usability-Attribute bei einem Interview eng an die gestellten Fragen gebunden. Ferner ist die Tatsache, dass die gegebenen Antworten lediglich die Meinung der Befragten widerspiegelt [Nie94c, S. 209 ff.], ein Grund dafür, dass die Qualität der einzelnen Attribute nur stark subjektiv bewertet wird. Hinzu kommt die beschränkte Validität der Aussagen der Interviewpartner. Ein Beispiel dafür ist eine Studie, die Benutzer vor und nach Einführung neuer Funktion in einem System befragte. Dabei wurde festgestellt, dass lediglich 28% der Personen, die zuvor angaben, die neuen Funktionen zu mögen, die Funktionen nach der Implementierung auch tatsächlich positiv bewerteten. Ein weiteres Beispiel hierfür ist eine Befragung von Nutzern, die nach der Verständlichkeit einer Anleitung für ein mobiles Telekommunikationssystem befragt wurden. Unter 25 Befragten gaben 24 Personen an, dass die Anleitung überdurchschnittlich gut oder einfach zu verstehen war.

Bei einem anschließenden Test ergab sich jedoch, dass lediglich 50% der Befragten das System tatsächlich bedienen konnten [Nie94c, S. 209 ff.]. Um diesen Faktor sowie die zuvor beschriebene Eigenschaft der Menschen, Antworten zu geben, bei denen sie denken, dass sie von ihnen erwartet werden, zu kompensieren, werden die entsprechenden Attribute nur als zur Hälfte berücksichtigt bewertet.

Die Fragen, um die Usability-Attribute zu evaluieren, können direkt aus den Metriken (vgl. Abschnitt 2.1.5) gewonnen werden. So kann zum Beispiel zur Messung der Effizienz nach der Dauer zum Erledigen einer bestimmten Tätigkeit gefragt werden, was der Metrik *Time-on-task* entspricht.

Learnability / Erlernbarkeit:

Da die Erlernbarkeit einer Anwendung daraus besteht, wie einfach oder schwierig es für jemanden ist, eine Aufgabe mit der Anwendung zu bedienen, ist es denkbar, dass die Antworten der befragten Personen nicht die Realität widerspiegeln. Sie fühlt sich womöglich unwohl dabei, zuzugeben, dass sie Probleme bei der Bedienung hatte.

Efficiency / Effizienz:

Auch bei diesem Attribut besteht die Möglichkeit, dass die Testperson nicht zugeben möchte, eine lange Zeit für eine vermeintlich einfache Aufgabe gebraucht zu haben. Den Anwendern ist es häufig bewusst, dass das Erledigen einer bestimmten Aufgabe nicht besonders lange dauern dürfte. Wenn sie dann nach einiger Zeit den Weg finden, der sie zum Ziel führt, könnten sie diesen Weg im Nachhinein als einfach empfinden und sind beschämt, nicht vorher auf die Idee gekommen zu sein.

Memorability / Einprägsamkeit:

Zur Ermittlung der Einprägsamkeit muss ein Testnutzer die Anwendung nach einiger Zeit erneut verwenden und anschließend befragt werden. Ein wichtiger Faktor der Einprägsamkeit ist die Effizienz. Da die Gültigkeit der Effizienz jedoch bereits nur eingeschränkt gegeben ist, gilt das gleiche für das Usability-Attribut Einprägsamkeit.

Errors / Fehler:

Einige Fehler, die der Befragte gemacht hat, könnten ihm unangenehm sein, so dass er nicht erwähnt, ihn begangen zu haben. Dies ist ähnlich dem Verhalten von Menschen, wenn sie etwas nicht verstanden haben und sie wissen, dass ihr Gegenüber die Antwort kennt. Es ist zu beobachten, dass viele Menschen in diesen Fällen trotzdem nicht nachfragen, besonders wenn weitere Personen anwesend sind. Bei einem Interview könnte das zum Beispiel ein Protokollant sein.

Satisfaction / Zufriedenheit:

Die Zufriedenheit des Anwenders an der Bedienung der Applikation ist ein ausschließlich subjektives Kriterium, was aus diesem Grund im Interview durch entsprechende Fragestellungen des Moderators in Erfahrung gebracht werden kann. Es gibt zwar einige Einschränkungen, die in dem nachfolgenden Abschnitt über die Berücksichtigung des emotionalen Zustandes ausführlich beschrieben werden, grundsätzlich kann die Zufriedenheit des Nutzers jedoch festgestellt

werden.

Zusammenfassung:

Die Berücksichtigung der Usability-Attribute durch das Interview ist aus den genannten Gründen nur unter Vorbehalt gegeben und eine eindeutige Bewertung ist nur eingeschränkt möglich. Mit Hilfe des Kompromisses, dass Attribute, die eigentlich nicht durch subjektive Aussagen zu evaluieren sind, nur als zur Hälfte berücksichtigt gelten, ergibt sich eine Bewertung von drei Punkten.

Berücksichtigung des emotionalen Zustands

Die Interview wird von Tullis & Albert schon als „*Self-Reported Metric*“ bezeichnet [TA08]. Diese Metrik zeichnet sich dadurch aus, dass mit ihrer Hilfe die subjektiven Daten gewonnen werden können. Demnach eignet sich das Interview zur Berücksichtigung des emotionalen Zustands. Es gibt jedoch einige Einschränkungen: Zum einen ist es bei dieser Methode nicht vorgesehen, den Benutzer während des Tests der Anwendung zu beobachten. Der Moderator ist also auf die Aussagen und Körpersprache des Interviewpartners während des Interviews angewiesen und kann keine Informationen aus Mimik und Gestik der Testperson während der Benutzung gewinnen. Es werden außerdem keine Video- oder sonstige Aufzeichnungen des Nutzers, die nachträglich analysiert werden könnten, während des eigentlichen Bedienens der Anwendung gemacht. Zum anderen wird der Benutzer erst nach der Benutzung befragt. Das heißt, er muss sich erinnern, was genau er in speziellen Situationen empfunden hat. Dadurch werden die Ergebnisse möglicherweise weniger verlässlich. Des Weiteren ist die Anwendung von „*Behavioral and Physiological Metrics*“ kein Bestandteil des Interviews als Usability-Testmethode. Aus diesen Gründen wird die Berücksichtigung des emotionalen Zustands mit drei Punkten bewertet.

Berücksichtigung des physischen Zustands

Dadurch, dass keine Informationen darüber verfügbar sind, wie sich das Testen der Anwendung im Feld und ein anschließendes Interview auf die Ergebnisse des Usability-Tests auswirken, ist es schwierig zu ermitteln, wie gut der physische Zustand tatsächlich berücksichtigt werden kann. Grundsätzlich ist es jedoch vorstellbar, dass eine Person, die an dem Interview teilnimmt, zunächst dazu aufgefordert wird, die Anwendung in seiner natürlichen Nutzungsumgebung zu verwenden. Je nach Anwendung wird die Person sich dann auch auf verschiedene Art und Weise bewegen / nicht bewegen. Anschließend kann diese Person dann nach den Problemen, die in diesen Situationen auftraten, befragt werden. Ein großer Nachteil ist dabei jedoch, dass die Zeit zwischen Test und Interview länger wird und sich die Testperson eventuell nicht mehr an jede Einzelheit erinnern kann. Es wäre auch möglich, die für die Applikation üblichen

Bewegungsarten im Labor zu simulieren und die Testperson zum Beispiel einen Parkour absolvieren zu lassen, der die Anwendung dieser Bewegungsarten beinhaltet. Dies würde jedoch zu Lasten der Berücksichtigung der natürlichen Nutzungsumgebung gehen. Bei Bedarf lässt sich demnach mit dem Interview vermutlich eine *starke Berücksichtigung* des physischen Zustands erreichen.

Berücksichtigung der natürlichen Nutzungsumgebung

Für die Berücksichtigung der natürlichen Nutzungsumgebung gilt ähnliches wie für die des physischen Zustands: Es ist möglich, den Testteilnehmer die Anwendung in der natürlichen Nutzungsumgebung testen zu lassen, jedoch ist die zeitliche Differenz zwischen Test und Interview dabei ein großer Nachteil. Da das Interview erst nach dem Test stattfindet, ist es schwieriger, einen Zusammenhang zwischen den Kriterien herzustellen. Der Nutzer könnte vergessen, wesentliche Aspekte zu erwähnen, was Auswirkungen auf die gewonnenen Informationen hat. Zum Beispiel könnte ihm eine Funktion, deren übliche Nutzung in Bewegung stattfindet, intuitiv vorgekommen sein, aber nur, weil er genügend Zeit hatte, die Bedienelemente in Ruhe anzuschauen, während er an einer roten Ampel wartete. Ein anderes Beispiel beinhaltet die Ein- und Ausgabe des Gerätes. Die Sprachbedienung einer Applikation könnte der Testperson nicht negativ aufgefallen sein, aber nur, weil die Umgebungslautstärke während der Bedienung sehr gering war. Sollte der Moderator nicht explizit nach solchen Variablen fragen, erwähnt der Interviewpartner sie möglicherweise auch nicht. Unter Umständen fragt der Moderator nach solchen Merkmalen, aber der Interviewpartner hat während des Tests nicht auf die Gegebenheiten geachtet und kann die Situation nicht rekapitulieren. Dies führt zu einer Bewertung der Berücksichtigung der Nutzungsumgebung mit drei Punkten.

Berücksichtigung der mobilgerätespezifischen Ein- und Ausgabeprobleme

Unter der Voraussetzung, dass der Testteilnehmer die Anwendung zuvor an einem mobilen Endgerät und nicht mit Hilfe eines Emulators testen konnte, können Probleme, die im Zusammenhang mit der Ein- / und Ausgabe stehen, getestet und erkannt werden. Hinzu kommt, dass Ein- und Ausgabeprobleme dadurch, dass sie häufig offensichtlich sind, vermutlich von jedem Testnutzer erkannt werden. Schlechte Lesbarkeit und wiederholtes Drücken der falschen Buttons haben einen direkten Effekt auf die Benutzung der Anwendung und fallen somit besonders negativ auf. Das Testen der Anwendung auf mehreren verschiedenen Endgeräten, um verschiedenartige Ein- und Ausgabemedien zu testen, ist möglich. Zu beachten ist dabei allerdings, dass mehr Zeit des Testnutzers in Anspruch genommen werden muss und somit gegebenenfalls eine höhere Vergütung nötig ist. Da die einzige Einschränkung darin besteht, dass der Interviewpartner die Probleme in Interview rekapitulieren muss, der Charakter der Probleme jedoch

ein leichteres Erinnern ermöglicht, kann die Berücksichtigung der mobilgerätespezifischen Ein- und Ausgabeprobleme als *stark* bezeichnet und demnach mit vier Punkten bewertet werden.

Beeinflussung des natürlichen Nutzungsverhaltens

In der hier vorgestellten und bewerteten Variante des Interviews findet während der Testnutzung keinerlei Videoaufzeichnung oder Beobachtung der Testperson statt. Der Teilnehmer kann das Endgerät so halten und die Anwendung so bedienen, wie sie es unter normalen Umständen täte und wird nicht von Gegenständen oder Betreuern des Tests dabei gestört. Das Nutzungsverhalten wird folglich nicht beeinflusst und das Kriterium mit fünf Punkten bewertet.

Gefundene Usability-Probleme

Über die Anzahl der gefundenen Usability-Probleme des Interviews als Usability-Testmethode für mobile Anwendungen waren zum Zeitpunkt des Schreibens dieser Arbeit keine Informationen verfügbar.

Gesamtbewertung

Abbildung 4.7 zeigt eine Übersicht über die Bewertung der einzelnen Kriterien. Dabei fällt die lange Dauer der Interview besonders negativ auf. Diese wäre unter Berücksichtigung der Vor- und Nachbereitung solcher Interviews sogar noch erhöht. Im Gegensatz dazu ist jedoch positiv zu beurteilen, dass die Nutzer während des Umgangs mit der mobilen Anwendung nicht gestört werden und somit eine natürliche Handhabung gewährleistet ist. Hinzu kommt, dass die Interviewpartner nicht zwangsläufig mit einer fertigen Applikation konfrontiert werden müssen. So lässt dieses Verfahren auch zu, sich Meinungen der Testpersonen über Zeichnungen oder Prototypen von Anwendungen einzuholen.

Zeitaufwand	●○○○○
Kosten	●●●●●
Zeitpunkt der Durchführung	●●●●●
Berücks. der Usability-Attribute	●●●○○
Berücks. des emotionalen Zustands	●●●○○
Berücks. des physischen Zustands	●●●●○
Berücks. der natürlichen Nutzungsumgebung	●●●○○
Berücks. der mobilgerätespez. Ein-/Ausgabeprobleme	●●●●○
Beeinflussung des natürlichen Nutzungsverhaltens	●●●●●

Abbildung 4.7: Übersicht über die Bewertung des Interviews

4.6 Fragebogen

Als zweite Methode der Usability-Befragungen, neben dem Interview (vgl. Abschnitt 4.5), wird im folgenden Abschnitt die Möglichkeit, Fragebögen für das indirekte Testen der Usability einzusetzen, beschrieben und bewertet.

Fragebögen, die beispielsweise in Papierform, per E-Mail oder als Online-Dokument ausgefüllt werden, unterscheiden sich von Interviews hauptsächlich dadurch, dass das Interview variabler gestaltet werden kann. Dem Moderator ist es möglich, seine Fragen anzupassen, auf Fragen besonders einzugehen oder einige auszulassen. Außerdem kann er dem Interviewpartner unverständliche Fragen erklären. Mit Fragebögen hingegen ist es möglich, große Gruppen oder sogar die gesamte Nutzerschaft einer Anwendung zu befragen. Sie können sogar mehrfach ohne Veränderungen eingesetzt werden, um Fortschritte festzustellen [Nie94c, S. 209 ff.].

Väätäjä & Roto entwickelten ein Verfahren [VR10] zur Umsetzung von Fragebögen auf mobilen Endgeräten. Es beinhaltet einen weiteren großen Unterschied: Der Fragebogen wird auf dem mobilen Endgerät ausgefüllt und kann somit direkt während der Benutzung oder unmittelbar danach zum Einsatz kommen. Da der Fragebogen sich dadurch noch weiter vom Interview als Usability-Testmethode abgrenzt, dient das Verfahren von Väätäjä & Roto als Grundlage für die Bewertung.

4.6.1 Beschreibung

Die Fragen, die in Fragebögen eingesetzt werden, sind in der Regel so gestellt, dass derjenige, der die Fragen beantwortet, keine langen Antworten geben muss. Tatsächlich ist es sogar wahrscheinlicher, dass ein Fragebogen ausgefüllt wird, wenn Antworten kurz gehalten werden können und der gesamte Fragebogen nur ein bis zwei Seiten umfasst [Nie94c, S. 213]. Die Antwortmöglichkeiten sind häufig vorgegeben oder bestehen aus der Angabe von Zahlen, zum Beispiel wie lange eine bestimmte Aktion dauerte. Zudem ist es üblich, dass Meinungen auf einer Beurteilungsskala anzugeben sind. Diese Verfahrensweise macht die Auswertung der Fragebögen leichter und erlaubt quantitative Datenanalysen [Nie94c, VR10].

In Abbildung 4.8 ist ein Beispiel der mobilen Fragebögen von Väätäjä und Roto zu sehen. Sie zeigt eine Möglichkeit, nach der Zufriedenheit des Benutzers mit Hilfe von Symbolen oder nach Qualitätsmerkmalen mit Hilfe von Skalen zu fragen. Väätäjä & Roto testeten ihre Entwürfe für Fragebögen auf mobilen Endgeräten mit Hilfe von sechs Usability-Experten als Anwender des mobilen Systems. Die Experten mussten während des Tests zwei Aufgaben mit einer multi-medialen Anwendung erledigen und nach jeder Aufgabe wurde mit Hilfe eines Fragebogens nach ihren Emotionen und ihrer Zufriedenheit im Umgang mit der Anwendung gefragt. Im Anschluss an die Aufgaben und die darauf folgenden kurzen Fragebögen folgte ein weiterer Fragebogen. Väätäjä & Roto fassen ihn als *„Mobile questionnaire on pragmatic and hedonic*



Abbildung 4.8: Fragebögen auf mobilen Endgeräten [VR10]

quality perceptions and judgement of appeal“ zusammen. Der gesamte Test wurde in einem Labor durchgeführt und nicht auf seine Gebrauchstauglichkeit im freien Feld hin überprüft.

Das Ergebnis der Arbeit sind Richtlinien, die dabei helfen sollen, einen Fragebogen möglichst komfortabel auf einem mobilen Endgerät bedienbar zu gestalten. Diese Richtlinien bestehen aus Ratschlägen dazu, wie der Fragebogen leicht zu lesen und auszufüllen ist sowie wenig Ablenkung bedeutet. Dabei spielen die kleine Bildschirmgröße, die Dateneingabe, der mobile Nutzungskontext und die gewählte Implementierung (über den mobilen Browser abrufbare Internetseite, SMS/MMS, mobile Anwendung) eine tragende Rolle [VR10].

4.6.2 Bewertung

Es folgt die Bewertung des Fragebogens. Damit sich die Methode weiter von dem Interview abgrenzt und somit in dieser Arbeit ein breiteres Spektrum von Methoden zum Usability-Testing analysiert wird, findet die Bewertung speziell für das Verfahren statt, das die Nutzer den Fragebogen auf dem mobilen Endgerät ausfüllen lässt.

Zeitaufwand

Es lässt sich keine genaue Zeit bestimmen, die für die Anfertigung und das Ausfüllen der Fragebögen benötigt wird. Das Verfahren „*Mobile Questionnaires for User Experience Evaluation*“ von Väättäjä & Roto sieht vor, dass nach jeder Aufgabe, die mit dem Endgerät erledigt wurde, ein Fragenkatalog beantwortet werden soll. Die Dauer des Tests hängt demnach von der Anzahl zu erledigenden Aufgaben ab. Diese Fragen nahmen in der Studie jedoch maximal sieben Minuten in Anspruch. Diese Zeit verringerte sich maßgeblich, als die Testpersonen den Fragebogen ein zweites Mal (nach der zweiten Aufgabe) ausfüllen sollten. Zu diesem Zeitpunkt wurden bei vier von sechs Personen nur noch ein bis zwei Minuten gebraucht³⁶ [VR10]. Diese

³⁶Eine Person benötigte weniger als eine Minute und eine Person drei bis vier Minuten.

kurzen Bearbeitungsdauern und die Tatsache, dass diese Fragebögen immer wieder verwendet werden können und dass bei Fragen mit Antwortauswahlmöglichkeiten eine maschinelle Auswertung möglich ist, rechtfertigt die Bewertung des Zeitaufwands mit fünf Punkten.

Kosten

Für eine aussagekräftige Usability-Befragung mit Hilfe eines Fragebogens sind nach Nielsen mindestens 30 Personen zu befragen. In der Praxis werden häufig zwischen 50 und 1000 Nutzer zufällig ausgewählt. Die genaue Anzahl der Nutzer hängt davon ab, wie detailliert die gewonnenen Daten sein sollen [Nie94c, Hol05]. Es ist allerdings nicht nötig, den Teilnehmern einer Befragung dieselbe Zeitaufwandsentschädigung zu bezahlen, wie es bei zuvor evaluierten Methoden der Fall ist. Bei Befragungen reichen häufig geringere Anreize oder die Chance auf einen Gewinn [Bro09]. In dem Artikel „7 aspects of successful usability questionnaires“ nennt Steve Balmer einen 25€ Gutschein als Beispiel für eine Incentivierung. Andere Möglichkeiten sind Preisnachlässe oder Gratisprodukte [Bro09]. Während sich Ausgaben von 25€ pro Person bei 50 Teilnehmern auf 1250€ belaufen, sind die Ausgaben mit 1000 Teilnehmern mit 50000€ bereits deutlich höher. Es empfiehlt sich demnach, die Incentivierung abhängig von der Anzahl der Benutzer zu machen und bei einem Bedarf von vielen Teilnehmer ein Gewinnspiel der garantierten Entlohnung vorzuziehen. Ausgaben zwischen 1001€ und 2000€ scheinen aus diesem Grund eine angemessene Schätzung zu sein. Daraus ergibt sich nach der Tabelle in Abschnitt 3.1.1 eine Bewertung von vier Punkten. Es fallen jedoch zusätzliche Kosten zur Analyse der gewonnenen Daten an.

Zeitpunkt der Durchführung

Die Befragungsmethode von Väättäjä & Roto benötigt eine funktionierende Anwendung, damit die Benutzer die Aufgabe mit dem mobilen Endgerät durchführen und anschließend den Fragebogen ausfüllen können. Denkbar ist, dass lediglich einen Funktionen der Anwendung implementiert wurden und anschließend hinsichtlich ihrer Usability getestet werden. Es ist also zumindest ein Prototyp erforderlich, um den Usability-Test durchführen zu können. Das Kriterium *Zeitpunkt der Durchführung* wird dementsprechend mit zwei Punkten bewertet.

Berücksichtigung der Usability-Attribute

Für Fragebögen gilt das gleiche wie für Interviews: Die gegebenen Antworten müssen nicht zwangsläufig immer der Wahrheit entsprechen. Das gilt besonders für Fragen, bei denen sie das Gefühl haben, dass sie die Wahrheit blamieren könnte [Nie94c, S. 214]. Eine ausführlichere

Beschreibung dieses Phänomens ist in der Bewertung des Interviews in Abschnitt 4.5.2 gegeben. Dennoch ist es den Menschen in der Regel unangenehmer gegenüber anderen Personen die Tatsachen nicht zu beschönigen als gegenüber Maschinen [Nie94c, S. 214]. Es wird davon ausgegangen, dass dieses Verhalten, die größere Zahl der befragten Personen und die dadurch gewonnene Masse an Daten sowie das direkte Beantworten der Fragen nach dem Erledigen einer Aufgabe ausreichend ist, um die möglichen Beschönigungen auszugleichen. Anders als bei der Bewertung des Interviews werden hier berücksichtigte Usability-Attribute, sofern nicht anderes erwähnt, vollständig gewertet.

Learnability / Erlernbarkeit:

Die Frage danach, wie einfach oder schwierig ein Benutzer das Erledigen einer Aufgabe findet, und damit indirekt nach dem *Task Success* zu fragen, gehört zu den grundlegenden Bestandteilen eines Fragebogens. Ebenso kann der Nutzer danach gefragt werden, wie lange er zum Erledigen der Aufgabe benötigt hat und ob ihm die Länge angemessen erschien. Das Usability-Attribut Erlernbarkeit wird somit durch den Fragebogen berücksichtigt.

Efficiency / Effizienz:

Die Effizienz einer mobilen Anwendung kann durch den Fragebogen ermittelt werden, wenn die Benutzer dazu aufgefordert werden, eine Aufgabe mehrfach durchzuführen und somit mit der Anwendung vertraut werden. Der Nutzer kann zwar anschließend gefragt werden, wie lange er für die Aufgabe gebraucht hat, eine Frage nach der Anzahl der Schritte, die er benötigt hat, oder wie viele Aktionen er durchführen musste, kann jedoch sehr komplex und ungeeignet für den mobilen Fragebogen werden. Es ist essentiell, dass diese Fragebögen sehr kurz und leicht verständlich sind, damit sie auch im Nutzungskontext ausgefüllt werden können. Jede Erklärungen, wie zum Beispiel, was genau als eine Aktion definiert wird, verlängern den Ausfüllprozess, was gegen die Richtlinien des Verfahrens verstößt [VR10]. Die Effizienz kann also nur teilweise berücksichtigt werden.

Memorability / Einprägsamkeit:

Da sich die Einprägsamkeit durch die Veränderung der Effizienz nach einer bestimmten Zeit definiert, kann dieses Attribut zum selben Grad wie die Effizienz berücksichtigt werden. Erneutes Ausfüllen des Fragebogens durch die gleichen Nutzer ist möglich, die Bestimmung der Effizienz mit Hilfe der Angaben aus dem Fragebogen jedoch nur teilweise. Somit ist auch die Einprägsamkeit einer Anwendung nur begrenzt überprüfbar.

Errors / Fehler:

Der mobile Fragebogen von Väättäjä & Roto sieht keine Felder vor, in denen freier Text geschrieben wird. Da Fehler jedoch sehr individuell und verschiedenster Ausprägung sein können, ist es nicht möglich, nach Fehlern zu fragen, sondern es besteht lediglich die Möglichkeit, eine Auswahl an Fehlern zu nennen. Nach fehlerhafter Bedienung kann aus diesem Grund nur ungenügend gefragt werden.

Satisfaction / Zufriedenheit:

Der Fragebogen lässt direkte Fragen bezüglich der Zufriedenheit der Testpersonen mit der An-

wendungen oder Bestandteilen der Anwendung zu. In Abbildung 4.8 sind fröhliche und traurige Gesichter als Beispiel für die Antwortmöglichkeiten bei Fragen nach dem Empfinden zu sehen. Die Zufriedenheit kann somit unter Anwendung der „*Self-Reported Metrics*“ (vgl. Abschnitt 2.1.5) berücksichtigt werden.

Zusammenfassung:

Die zuvor beschriebenen Eigenschaften hinsichtlich der Berücksichtigung der einzelnen Usability-Attribute führt zu einer Gesamtbewertung dieses Kriteriums von drei Punkten.

Berücksichtigung des emotionalen Zustands

Die Berücksichtigung des emotionalen Zustands der Benutzer durch den Fragebogen ist mit der durch das Interview zu vergleichen. Es werden subjektive Daten durch direkte Fragen gewonnen, es ist jedoch nicht möglich, „*Behaviorial and Physiological Metrics*“ anzuwenden, da der Nutzer während des Umgangs mit der Anwendung nicht beobachtet oder aufgezeichnet wird. Vorteilhaft ist jedoch, dass in dem Verfahren von Vääätäjä & Roto die Testnutzer direkt nach Erledigen einer Aufgabe nach ihrer Meinung gefragt werden. Dies erleichtert ihnen das Beantworten der Fragen, weil sie zwischendurch nicht durch andere Umstände abgelenkt werden.

Ein Nachteil, der bei Fragebögen im Gegensatz zu Interviews besteht, liegt darin, dass Mimik und Gestik des Nutzers auch während der Beantwortung der Fragen nicht sichtbar sind und keine Informationen daraus gewonnen werden können. Des Weiteren sind ihre Antwortmöglichkeiten auf eine Skala beschränkt. Während in einem Interview noch nach dem Grund für Gefühle wie Ärger oder Enttäuschung gefragt werden kann, ist dies im Verfahren von Vääätäjä & Roto nicht vorgesehen [VR10]. Zusammenfassend lässt sich also die Berücksichtigung des emotionalen Zustands durch den mobilen Fragebogen als *schwach* bezeichnen.

Berücksichtigung des physischen Zustands

Der physische Zustand kann im begrenzten Rahmen mit Hilfe des mobilen Fragebogens berücksichtigt werden. Zu beachten ist dabei, dass die Antwortmöglichkeiten das gesamte Spektrum der Bewegungsmöglichkeiten abdecken. Möglicherweise lässt sich auch durch Abfrage der Nutzungsumgebung eine Vermutung über den physischen Zustand anstellen. Dabei ist jedoch zu beachten, dass, nur weil ein Testnutzer beispielsweise angibt, er sei in einem Fahrzeug, dies noch nicht bedeutet, dass er sich fortbewegt. Aus dem Grund, dass die Fragen des mobilen Fragebogens Antwortmöglichkeiten vorgegeben haben und der Benutzer nicht beobachtet wird, ist es außerdem schwierig, etwas über die genaue Geschwindigkeit der Fortbewegung zu erfahren. Der Fragebogen könnte sonst schnell unübersichtlich und nicht mehr zügig zu bearbeiten werden, was gegen die von Vääätäjä & Roto vorgeschlagenen Richtlinien verstößt [VR10]. Grund-

sätzlich ist mit diesem Verfahren also eine *mittlere* Berücksichtigung des physischen Zustands möglich.

Berücksichtigung der natürlichen Nutzungsumgebung

Der Usability-Test mit Hilfe des mobilen Fragebogens ist dadurch, dass der Fragebogen direkt auf dem mobilen Endgerät ausgefüllt wird, an jedem Ort und zu jeder Zeit durchführbar. Die Motivation, eine Orientierungshilfe für mobile Fragebögen zu entwickeln, besteht laut Vääätäjä & Roto sogar daraus, Fragebögen von Nutzern in natürlichen Nutzungskontexten ausfüllen zu lassen [VR10]. Es ist jedoch nicht vorgesehen, dass der Nutzer explizit danach gefragt wird, wie der gegenwärtige Nutzungskontext ist. Eine Antwort über Unzufriedenheit mit einer bestimmten Funktion oder Eigenschaft des Endgerätes kann somit nicht mit der Nutzungsumgebung in Zusammenhang gebracht werden. Aus diesem Grund lässt sich nur von einer *starken* Berücksichtigung der natürlichen Nutzungsumgebung sprechen.

Berücksichtigung der mobilgerätespezifischen Ein- und Ausgabeprobleme

Es ist möglich, den Fragebogen auf verschiedensten Endgeräten zu beantworten. In dem Pilotversuch von Vääätäjä & Roto wird den mobilen Endgeräten eine Kurznachricht mit einem Hyperlink geschickt. Die teilnehmenden Endgeräte benötigen in diesem Fall also einen Browser, es ist aber auch möglich, den Fragebogen mit entsprechendem Mehraufwand als eigene Anwendung auf den Endgeräten zu installieren. Sollte die mobile Anwendung auf den privaten Endgeräten der Testnutzer getestet werden, könnten die Teilnehmer so ausgewählt werden, dass die Anwendung auf allen relevanten Endgeräten getestet wird. Wenn die mobilen Endgeräte gestellt werden, müssen entsprechend verschiedenartige Geräte verteilt werden. In jedem Fall lässt sich die Heterogenität der Endgeräte und ihre speziellen Ein- und Ausgabemedien *sehr stark* berücksichtigen, was eine Bewertung von fünf Punkten rechtfertigt.

Beeinflussung des natürlichen Nutzungsverhaltens

Der natürliche Umgang der Testteilnehmer mit dem mobilen Endgerät und der zu testenden Anwendung wird nicht durch angebrachte Kameras oder Beobachter des Tests und ähnliches beeinträchtigt. Die Beantwortung der Fragen nach jeder erledigten Aufgabe verlängert jedoch den Umgang mit dem Gerät und beeinflusst somit das natürliche Nutzungsverhalten. Der Nutzer soll die Anwendung nicht direkt erneut verwenden, sondern zunächst für einige Minuten den Fragebogen ausfüllen und wird dadurch möglicherweise in seinem gewöhnlichen Workflow unterbrochen. Da dies jedoch nicht zwingend der Fall ist und keine Beeinflussung durch den Test bedingte äußere Umstände stattfindet, wird die Beeinflussung als *schwach* bewertet.

Gefundene Usability-Probleme

Es sind bisher keine Studien darüber Verfügbar, wie der mobile Fragebogen im Vergleich zu anderen Usability-Testmethode hinsichtlich der gefundenen Usability-Probleme steht. Auch Vääätäjä & Roto geben keine Informationen darüber, welche Schwachstellen der von ihnen gewählten Beispielanwendung aufgedeckt werden konnten. Der Schwerpunkt ihrer Arbeit besteht weniger aus dem Messen der Leistungsfähigkeit des mobilen Fragebogen als viel mehr aus dem Erarbeiten eines optimal nutzbaren Fragebogens [VR10].

Gesamtbewertung

Bei der Betrachtung der Übersicht über die Bewertung in Abbildung 4.9 wird ersichtlich, dass die Schwachpunkte des mobilen Fragebogens darin liegen, dass der Benutzer nicht beobachtet werden kann und seine Antwortmöglichkeiten vorgegeben sind. So ist es nur begrenzt möglich, detailliert auf Antworten der Testperson einzugehen beziehungsweise über Mimik und Gestik des Nutzers etwas über dessen emotionalen Zustand zu erfahren. Die Vorteile der Methode hingegen liegen im Bereich der Zeit und der Kosten. Der mobile Fragebogen macht es möglich, viele Nutzer gleichzeitig in den Test zu involvieren. Ein weiterer positiver Aspekt des mobilen Fragebogen ist, dass der Anwender das mobile Endgerät in seiner natürlichen Nutzungsumgebung bedienen kann und nicht auf andere Personen angewiesen ist.

Zu Berücksichtigen ist jedoch, dass noch zusätzliche Zeit für die Vorbereitung des Tests und für die Analyse der Daten eingeplant werden muss. So müssen die Fragen des Fragebogens müssen hinsichtlich ihrer Verständlichkeit überprüft werden, bevor der Usability-Test mit diesem Verfahren beginnen kann [VR10].

Zeitaufwand	●●●●●●
Kosten	●●●●●○
Zeitpunkt der Durchführung	●●○○○○
Berücks. der Usability-Attribute	●●●○○○
Berücks. des emotionalen Zustands	●●○○○○
Berücks. des physischen Zustands	●●●○○○
Berücks. der natürlichen Nutzungsumgebung	●●●●●○
Berücks. der mobilgerätespez. Ein-/Ausgabeprobleme	●●●●●●
Beeinflussung des natürlichen Nutzungsverhaltens	●●●○○○

Abbildung 4.9: Übersicht über die Bewertung des Fragebogens

4.7 Zusammenfassung

Abbildung 4.10 stellt alle Usability-Testmethoden und die Bewertung der einzelnen Kriterien im Vergleich dar. Mit Hilfe dieser Abbildung lässt sich zum einen schnell ein Überblick darüber gewinnen, wie eine Methode in einem bestimmten Kriterium bewertet wurde. Zum anderen ist es jedoch auch leicht möglich, die Methoden, die bei einem bestimmten Kriterium besonders gut oder schlecht abgeschnitten haben, zu ermitteln. Besonders auffällig ist in diesem Sinne die gute Bewertung der *Beeinflussung des natürlichen Nutzungsverhaltens* des Interviews. Diese Methode ist die einzige der betrachteten, die den Nutzer während der Bedienung in keiner Weise ablenkt und mit *echten* Nutzern (keine mit überdurchschnittlicher Vorkenntnis über Usability) durchgeführt wird.

Ein weitere Auffälligkeit ist das verhältnismäßig gute Abschneiden der Methoden bei dem Kriterium Berücksichtigung der *mobilergerätespezifischen Ein- und Ausgabeprobleme*. Dies ergibt sich daraus, dass die Methoden alle auf tatsächlichen Endgeräten durchgeführt werden können. Es ist jedoch denkbar, dass zukünftig auch Methoden zum Einsatz kommen, die nur auf einem Emulator durchführbar sind. Eine solche Usability-Testmethode könnte zum Vorteil haben, dass die Benutzeroberfläche und Positionierung bestimmter Bildelemente während des Tests leicht variierbar sind.

Die Abbildung 4.10 zeigt außerdem die ungewichtete Gesamtbewertung der Methoden. Dabei handelt es sich also um die Kumulation der vergebenen Punkte. Die Aussagekraft dieser Gesamtpunkte ist allerdings beschränkt. Zwar ist der Kognitive Walkthrough die Methoden mit den meisten Punkten, jedoch ist mit ihr nur eine *sehr schwache* Berücksichtigung des emotionalen Zustands und somit der Zufriedenheit der Nutzer mit der Anwendung möglich. In der Regel sollte die Zufriedenheit der Nutzer jedoch einen hohen Stellenwert einnehmen, damit sichergestellt werden kann, dass die mobile Anwendung von der potenziellen Nutzerschaft akzeptiert wird. Im Gegensatz dazu sind die beiden Methoden, die die niedrigste Gesamtpunktzahl erreichten, der Think Aloud Test im Labor und der im Feld, die Usability-Testmethoden, mit denen sich emotionalen Zustand der Nutzer am besten berücksichtigen lässt.

	Zeitaufwand	Kosten	Zeitpunkt der Durchführung	Berücks. der Usability-Attribute	Berücks. des emotionalen Zustands	Berücks. des physischen Zustands	Berücks. der natürlichen Nutzungsumgebung	Berücks. der mobilgerätespezif. Ein-/Ausgabeprobleme	Beeinflussung des natürlichen Nutzungsverhaltens	Gesamt (ungewichtet)
Fragebogen	●●●●●	●●●●○	●●●●●	●●●●○	●●●●○	●●●●○	●●●●○	●●●●●	●●●●○	31
Interview	●○○○○	●●●●●	●●●●●	●●●●○	●●●●○	●●●●○	●●●●○	●●●●○	●●●●●	33
Heuristische Evaluation	●●●●●	●●●●●	●●●●●	●●●●○	●●●●○	●●●●○	●●●●○	●●●●●	●●●●○	31
Kognitiver Walkthrough	●●●●●	●●●●●	●●●●●	●●●●○	●●●●○	●●●●○	●●●●●	●●●●●	●●●○○	36
Think Aloud Test (im Feld)	●○○○○	●●●○○	●○○○○	●●●○○	●●●●●	●●●●●	●●●○○	●●●○○	●○○○○	28
Think Aloud Test (im Labor)	●○○○○	●●●○○	●●●●●	●●●○○	●●●●●	●○○○○	●○○○○	●●●●●	○○○○○	27

Abbildung 4.10: Übersicht über die Bewertung aller Usability-Testmethoden

Kapitel 5

Fazit und Ausblick

In diesem letzten Kapitel werden die Erkenntnisse, die durch diese Arbeit gewonnen werden, in kurzer Form rekapituliert. Dabei werden auch die aufgetretenen Probleme genannt und die Konsequenzen, die daraus gezogen werden mussten, aufgeführt. Anschließend soll der Ausblick Möglichkeiten aufzeigen, diese Forschungsarbeit fortzuführen oder darauf aufzubauen.

5.1 Fazit

Ein Ziel der Arbeit war es, grundlegende Kriterien zu erarbeiten, die es erlauben, eine Bewertung von existierenden Usability-Testmethoden für mobile Anwendungen vorzunehmen, und dabei die Möglichkeit zu haben, einige Kriterien stärker in die Bewertung einfließen zu lassen als andere. Diese Kriterien sollten außerdem angewandt werden, um eine repräsentative Auswahl dieser Testmethoden zu evaluieren.

Zur Erarbeitung der notwendigen Kriterien wurden in Abschnitt 2.1 die wesentlichen Aspekte der Usability und des Usability-Engineerings erläutert. In diesem Zusammenhang erfolgte die Beschreibung von fünf Attributen, durch die sich die Usability definiert. Diese Attribute sind *Erlernbarkeit*, *Effizienz*, *Einprägsamkeit*, *Fehler* und *Zufriedenheit*. Wie gut eine Applikation diese Merkmale der Usability erfüllt, kann mit Hilfe von Metriken ermittelt werden. In dieser Arbeit wurden dazu die von Tullis & Albert [TA08] zusammengetragenen Metriken verwendet und in Abschnitt 2.1.5 vorgestellt.

Des Weiteren wurde beschrieben, wie sich das Usability-Engineering in den Softwareentwicklungsprozess eingliedert, und dabei festgestellt, dass eine kontinuierliche Verbesserung der Usability während des ganzen Prozesses und darüber hinaus stattfinden sollte. Die Analyse bestehender Standards und Normen ergab zudem, dass Nutzer und Nutzungskontext bereits möglichst in der Planungsphase in die Anwendungsentwicklung einbezogen werden sollten, um die sich gegenseitig beeinflussenden Faktoren Benutzer, Kontext, Arbeitsziele und Anwendungsfunktionen schon früh berücksichtigen zu können.

Damit die speziellen Faktoren der mobilen Nutzung der Anwendungen in den Bewertungskriterien berücksichtigt werden können, gibt eine Erläuterung der Nutzungskontexte von Applikationen in Abschnitt 2.2 Aufschluss darüber, was das Besondere an dieser Art von Anwendungen ist. Dabei wurde zwischen den Variablen, die direkt von dem Nutzer ausgehen, und den Variablen, die von außen auf den Nutzer einwirken, unterschieden. Zu den erstgenannten gehören die Emotionen des Nutzers, die sowohl durch äußere Umstände als auch durch die Anwendung selbst beeinflusst werden können, und die Bewegungsarten beziehungsweise Körperhaltungen des Nutzers, wie beispielsweise rennen, gehen, sitzen oder liegen, die in dieser Arbeit als physischer Zustand bezeichnet wurden. Die von außen auf den Nutzer einwirkenden Variablen,

stellen seine Umwelt dar. Diese gliedert sich in den physischen Kontext, der aus auditiven und visuellen Elementen besteht, und den sozialen Kontext, der sich aus der Interaktion mit anderen Menschen während der Bedienung der Anwendung ergibt.

Außerdem wurden die Charakteristika der mobilen Nutzungsgeräte geschildert. Sie kennzeichnen sich hauptsächlich durch die Heterogenität der erhältlichen mobilen Endgeräte und deren, im Vergleich zu Desktop Computern, niedrige Leistungsfähigkeit sowie die besonderen Medien zur Ein- und Ausgabe der Daten.

Aus den grundlegenden Aspekten des Usability-Engineerings und Besonderheiten der mobilen Nutzung wurden Kriterien zur Bewertung der Usability-Testmethoden erarbeitet und in Kapitel 3.1 vorgestellt. Um zu verdeutlichen, welche Kriterien speziell für Usability-Tests von mobilen Anwendungen relevant sind, fand dabei eine Unterteilung in geräteabhängige und geräteunabhängige Kriterien statt. Die geräteabhängigen Kriterien sind die Beeinflussung des natürlichen Nutzungsverhaltens, Berücksichtigung des physischen Zustands, der natürlichen Nutzungsumgebung und der mobilgerätespezifischen Ein- und Ausgabeprobleme. Wobei letzteres bewerten soll, wie gut die Heterogenität der mobilen Endgeräte durch die jeweilige Usability-Testmethode berücksichtigt werden kann. Zu den geräteunabhängigen Kriterien gehören der Zeitpunkt der Durchführung, die Berücksichtigung der Usability-Attribute und die des emotionalen Zustands. Zudem wurde versucht, wirtschaftliche Aspekte in die Bewertung einer Methode mit einfließen lassen zu können, indem auch die Ressourcen Zeit und Geld, die für eine Methode zur Verfügung stehen müssen, als geräteunabhängige Kriterien aufgeführt werden. Daraus ergaben sich die Kriterien Zeitaufwand und Kosten. Da der Anspruch dieser Arbeit jedoch war, dass auf ihrer Grundlage weitere Usability-Testmethoden bewertet werden können, mussten alle Kriterien so aufgebaut sein, dass ihre Bewertung nicht relativ, sondern absolut stattfindet. Dies bedeutet, dass die in Kapitel 4 berücksichtigten Methoden möglichst nicht direkt miteinander verglichen wurden, um eine Rangfolge pro Kriterium aufzustellen. Ansonsten wäre es ohne größere Umstände nicht möglich gewesen, weitere Methoden auf dieselbe Weise zu evaluieren, da davon alle vorherigen Ergebnisse beeinflusst werden würden.

Für die Dauer und Kosten der Methoden führte dies aufgrund der engen Beziehung der beiden Kriterien zueinander und der Abhängigkeit von den vorhandenen Ressourcen zu Problemen. Da es sich bei diesen Attributen jedoch häufig um wichtige Faktoren in der Entscheidung für oder gegen ein Verfahren handelt, wurden diese Kriterien dennoch aufgeführt und es wurde versucht, diese durch Schätzungen bestmöglich zu berücksichtigen. Die auf diese Weise entstandenen Werte sind in keinem Fall repräsentativ, dienen aber dazu, die in dieser Arbeit berücksichtigten Methoden vergleichbar zu machen.

Damit in einem solchen Vergleich eine Gesamtbewertung unter Einbeziehung von Prioritäten auf bestimmte Kriterien berechnet werden kann, sind mathematische Verfahren für die Gewichtung der Kriterien hilfreich. In Abschnitt 3.2 wurde ein solches Verfahren vorgestellt.

Die in Kapitel 4 durchgeführte Beschreibung und Bewertung von sechs Usability-Testmethoden (Think Aloud Test im Labor, Think Aloud Test im Feld, Kognitiver Walkthrough, Heuristische

Evaluation, Interview, Fragebogen) ermittelte den Kognitiven Walkthrough als Methode mit der höchsten Gesamtpunktzahl. Dies bedeutet jedoch nicht, dass sich dieses Verfahren für jeden Zweck am besten eignet, da in dieser Gesamtpunktzahl keine Prioritäten gesetzt wurden. So ist der Kognitive Walkthrough beispielsweise ungeeignet, wenn der Fokus bei einem Usability-Test auf den Reaktionen beziehungsweise der Zufriedenheit der Nutzer liegen soll. Er gehört wie die Heuristische Evaluation zu den Usability-Inspektionen und wird von Usability-Experten ohne Einbeziehung von Testnutzern durchgeführt. Die Stärken dieser Verfahren sind insbesondere die geringen Kosten und flexible, schnelle Durchführung.

Im Gegensatz dazu steht der Think Aloud Test, der durch ausführliche Kommunikation mit Testnutzern geprägt ist. Dieses Verfahren kann im Labor und in natürlicher Umgebung eingesetzt werden. Beide Varianten werden in dieser Arbeit berücksichtigt. Durch die enge Zusammenarbeit mit Testpersonen kann die Zufriedenheit der Nutzer mit Hilfe dieser Usability-Testmethoden sehr stark berücksichtigt werden. Die Schwächen des Think Aloud Tests im Labor, dass die natürliche Nutzungsumgebung und der physische Zustand der Nutzer nur begrenzt berücksichtigt werden können, werden bei dem Think Aloud Test im Feld zu Lasten der Kosten, Dauer und dem frühestmöglichen Zeitpunkt der Durchführung behoben.

Eine weitere Gruppe von Verfahren des Usability-Testings sind die Befragungsmethoden. Als prominente Vertreter dieser Gruppe wurden das Interview und der mobile Fragebogen vorgestellt und bewertet. Ein negativer Aspekt beider Methoden ist die Tatsache, dass die Testpersonen während der Nutzung der Anwendung nicht beobachtet werden können und dementsprechend keine Informationen aus der Mimik und Gestik gewonnen werden können. Das Interview ist jedoch besonders hinsichtlich der Beeinflussung des natürlichen Nutzungsverhaltens hervorzuheben. Es ist das einzige der betrachteten Verfahren, das in diesem Kriterium die volle Punktzahl erreichte. Außerdem ist es sehr flexibel einsetzbar, da der Interviewpartner nicht zwangsläufig über eine funktionierende Anwendung befragt werden muss. Die Anwendung des mobilen Fragebogens benötigt hingegen zumindest einen Prototypen. Dafür ist es mit dem Fragebogen leicht möglich, sehr viele Personen auf einmal zu befragen, was sich positiv auf die Gesamtdauer auswirkt. Außerdem entstehen in der Regel keine hohen Kosten, da die Methode nicht sehr aufwendig ist und den Testpersonen ein kleinerer Anreiz als in den anderen betrachteten Verfahren gegeben werden muss.

Es lässt sich also abschließend feststellen, dass keine der ausgewählten Methoden für jeden Zweck perfekt geeignet ist. Vielmehr ist es abhängig von dem Typ der zu testenden Anwendung sowie den Zielen und Mitteln ihres Vertreibers. Mit Hilfe von Prioritäten können diese Faktoren entsprechend berücksichtigt werden, um die bestmögliche Methode für die jeweilige Situation zu ermitteln.

5.2 Ausblick

Eine Möglichkeit zur Fortführung dieser Forschungsarbeit besteht darin, die auf Literaturrecherche basierenden Ergebnisse durch praktische Studien zu validieren. Auf diese Weise ist es zudem möglich, mit allen Usability-Testmethoden dieselbe Anwendung und dieselben Funktionen zu testen. So kann eine bessere Bewertung des Kriteriums Zeitaufwand und möglicherweise auch der entstehenden Kosten durchgeführt werden.

Die in dieser Arbeit geschaffenen Basis zur Bewertung von Methoden könnte außerdem als Grundlage für die Entwicklung eines neuen Verfahrens zum Testen der Usability von mobilen Anwendungen genutzt werden. Die Kriterien dienen dabei als Orientierungshilfe, was für die Qualität einer solchen Methode von Relevanz ist. Außerdem können die Beschreibungen und Bewertungen der berücksichtigten Methoden als Quelle für Ideen zur Entwicklung des neuen Verfahrens genutzt werden.

Schlussendlich kann die Arbeit jedoch auch genutzt werden, um die bisher nicht berücksichtigten Methoden oder neu entstandene Methoden zu bewerten. Ein interessanter Ansatz solcher nicht berücksichtigter Methoden ist beispielsweise das automatisierte Testen der Usability. Au et al. [ABWD08] entwickelten ein solches „Automated Usability Testing Framework“. Die bisherigen Informationen über dieses Verfahren genügten bisher jedoch nicht, um es schon in dieser Arbeit zu berücksichtigen. Die steigende Popularität mobiler Applikationen, die kürzer werdenden Entwicklungszyklen und die Möglichkeit, den Endnutzern auf einfache Art und Weise Updates zur Verfügung zu stellen, führen vermutlich dazu, dass sich die Verfahren zum Testen der Usability ändern. Denkbar ist eine engere Einbeziehung der tatsächlichen Benutzer in die Verbesserung der mobilen Anwendungen. Dies kann aus Sicht des Benutzers passiv durch Protokollierung der Benutzeraktionen geschehen oder aktiv, zum Beispiel durch Einholung von Resonanz in jeglicher Form.

Die Funktion, Kommentare zu einer Anwendung zu hinterlassen, ist bereits in Programmen zum Herunterladen von mobilen Anwendungen³⁷ integriert. Dadurch besteht bereits eine erste Variante, das Feedback der eigenen Nutzer zu erfassen, um die Anwendung an die Wünsche dieser anzupassen. Somit ist auch diese Funktion eine Möglichkeit, nach der bereits in der Einleitung zitierten These von Larry Marine zu handeln:

„It is far better to adapt the technology to the user than to force the user to adapt to the technology.“

³⁷Die Kommentarfunktion ist zum Beispiel im Appstore von Apple, von Amazon und im Android Market implementiert.

Literaturverzeichnis

- [ABWD08] AU, F.T.W. ; BAKER, S. ; WARREN, I. ; DOBBIE, G.: Automated usability testing framework. In: *Proceedings of the ninth conference on Australasian user interface-Volume 76* (2008), S. 55–64
- [AM09] AGARWAL, A. ; MEYER, A.: Beyond usability: evaluating emotional response as an integral part of the user experience. In: *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems* (2009), S. 2919–2930
- [App11] APPLE: View and Window Architecture. (2011). http://developer.apple.com/library/ios/#documentation/WindowsViews/Conceptual/ViewPG_iPhoneOS/WindowsandViews/WindowsandViews.html, Abruf: 08.09.2011
- [Bam] BAME, M.: Definition: Human Factors Engineering. <http://defense.about.com/od/HIJKLM/g/Human-Factors-Engineering.htm>, Abruf: 28.07.2011
- [Bar10] BARNUM, C.: *Usability Testing Essentials: Ready, Set ...Test!* Elsevier Science & Technology, 2010
- [BCD⁺08] BERTINI, E. ; CATARCI, T. ; DIX, A. ; GABRIELLI, S. ; KIMANI, S. ; SANTUCCI, G.: Appropriating Heuristic Evaluation Methods for Mobile Computing. In: *Handbook of Research on User Interface Design and Evaluation for Mobile Technology, IGI Global, S* (2008), S. 780–801
- [BCD⁺09] BERTINI, E. ; CATARCI, T. ; DIX, A. ; GABRIELLI, S. ; KIMANI, S. ; SANTUCCI, G.: Appropriating heuristic evaluation for mobile computing. In: *Int. J. Mobile Hum. Comput. Interact* 1 (2009), Nr. 1, S. 20–41
- [BCK⁺03] BECK, E.T. ; CHRISTIANSEN, M.K. ; KJELDSKOV, J. ; KOLBE, N. ; STAGE, J.: Experimental evaluation of techniques for usability testing of mobile systems in a laboratory setting. In: *Proceedings of Ozchi* (2003), S. 106–115
- [Bec78] BECHMANN, A.: *Nutzwertanalyse, Bewertungstheorie und Planung*. Haupt, 1978
- [Bla11] BLACKBERRY: Developing applications for different screen sizes. (2011). http://docs.blackberry.com/en/developers/deliverables/29251/Screens_1602941_11.jsp#Developing_apps_for_different_screen_sizes_1579421_11, Abruf: 08.09.2011
- [BMO08] BERNHAUPT, R. ; MIHALIC, K. ; OBRIST, M.: Usability evaluation methods for mobile applications. In: *Handbook of Research on User Interface Design and Evaluation for Mobile Technology, IGI Global, S* (2008), S. 745–758

Literaturverzeichnis

- [Bro09] BROMLEY, S.: 7 aspects of successful usability questionnaires. (2009). <http://www.stevebromley.com/blog/2009/11/11/7-aspects-of-successful-usability-questionnaires/>, Abruf: 23.11.2011
- [BTT05] BENYON, D. ; TURNER, P. ; TURNER, S.: *Designing interactive systems*. Addison-Wesley, 2005
- [Cha11] CHARNY, Ben: Write once, run anywhere' not working for phones. (2011). http://news.cnet.com/Write-once%2C-run-anywhere-not-working-for-phones---page-2/2100-1037_3-5788766-2.html?tag=mncol, Abruf: 08.09.2011
- [Chi07] CHISNELL, D.: When to ask participants to think out loud. (2007). <http://usabilitytestinghowto.blogspot.com/2007/04/when-to-ask-participants-to-think-out.html>, Abruf: 27.10.2011
- [Chr08] CHRISTOPOULOU, E.: Context as a necessity in mobile applications. In: *Handbook of Research on User Interface Design and Evaluation for Mobile Technology* 53 (2008)
- [Com11] COMPANY, The N.: In US, Smartphones Now Majority of New Cellphone Purchases. (2011). <http://blog.nielsen.com/nielsenwire/?p=28237>, Abruf: 07.09.2011
- [Dev11] DEVELOPERS, Android: Supporting Multiple Screens. (2011). http://developer.android.com/guide/practices/screens_support.html, Abruf: 08.09.2011
- [DL08] DUMAS, J.S. ; LORING, B.A.: *Moderating usability tests: principles and practice for interacting*. Morgan Kaufmann, 2008
- [DR99] DUMAS, J.S. ; REDISH, J.C.: *A practical guide to usability testing*. Intellect, 1999
- [DTC06] DUH, H.B.L. ; TAN, G.C.B. ; CHEN, V.H.: Usability evaluation for mobile device: a comparison of laboratory and field tests. In: *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services* (2006), S. 181–186
- [Ega88] EGAN, D. E.: Individual differences in human-computer interaction. In: *Human-Computer Interaction* (1988), S. 543–568
- [Far11] FARAGO, Peter: iOS & Android Apps Challenged by Traffic Acquisition Not Discovery. (2011). <http://blog.flurry.com/bid/76874>, Abruf: 13.12.2011

- [FMK⁺10] FALAKI, H. ; MAHAJAN, R. ; KANDULA, S. ; LYMBEROPOULOS, D. ; GOVINDAN, R. ; ESTRIN, D.: Diversity in smartphone usage. In: *Proceedings of the 8th international conference on Mobile systems, applications, and services* (2010), S. 179–194
- [Gar11] GARTNER.COM: Gartner Says Worldwide Mobile Device Sales to End Users Reached 1.6 Billion Units in 2010; Smartphone Sales Grew 72 Percent in 2010. (2011). <http://www.gartner.com/it/page.jsp?id=1543014>, Abruf: 13.12.2011
- [GBU88] GOULD, J.D. ; BOIES, S.J. ; UKELSON, J.: How to design usable systems. In: *Handbook of human-computer interaction* (1988), S. 757–789
- [GM03] GORLENKO, L. ; MERRICK, R.: No wires attached: Usability challenges in the connected mobile world. In: *IBM Systems Journal* 42 (2003), Nr. 4, S. 639–651
- [GMKC05] GABRIELLI, S. ; MIRABELLA, V. ; KIMANI, S. ; CATARCI, T.: Supporting cognitive walkthrough with video data: a mobile learning evaluation study. In: *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services* (2005), S. 77–82
- [Hol05] HOLZINGER, A.: Usability engineering methods for software developers. In: *Communications of the ACM* 48 (2005), Nr. 1, S. 71–74
- [Hur09] HURLEY, B.: GUI Mock Up Software Tools. (2009). <http://purplejunction.com/2009/08/01/gui-mock-up-software-tools/>, Abruf: 23.11.2011
- [ISO91] ISO/IEC: *ISO/IEC 9126: Software engineering – Product quality*. ISO/IEC, 1991
- [ISO99] *ISO 9241: Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten Teil 11: Anforderungen an die Gebrauchstauglichkeit - Leitsätze*. ISO, 1999 (DIN EN ISO 9241-11)
- [ISO06] *ISO 9241: Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten Teil 110: Grundsätze der Dialoggestaltung*. ISO, 2006 (DIN EN ISO 9241-110)
- [Jam12] JAMES, T.: Comparison of Usability Evaluation Methods. (2012). http://www.ehow.com/info_10072678_comparison-usability-evaluation-methods.html, Abruf: 10.11.2011
- [Jor02] JORDAN, P.W.: *Designing pleasurable products: an introduction to the new human factors*. Taylor & Francis, 2002

- [Kar02] KARAT, J.: Beyond task completion: evaluation of affective components of use. In: *The human-computer interaction handbook* (2002), S. 1152–1164
- [Kha01] KHALID, HM: Can customer needs express affective design. In: *Proceedings of the international conference on affective human factors design* (2001), S. 190–198
- [KK05] KALLIO, T. ; KAIKKONEN, A.: Usability testing of mobile applications: A comparison between laboratory and field testing. In: *Journal of Usability Studies* 1 (2005), S. 4–16
- [KKL⁺02] KIM, H. ; KIM, J. ; LEE, Y. ; CHAE, M. ; CHOI, Y.: An empirical study of the use contexts and usability problems in mobile Internet. In: *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on* (2002), S. 1767–1776
- [kom10] KOMPAKT, Usability E.: Glossar. (2010). <http://www.benutzbar.ch/index.php?title=Glossar>, Abruf: 24.10.2011
- [Kra] KRANNICH, Dennis: *Mobile Usability-Testing. Ein toolbasiertes Vorgehensmodell zum Rapid-Prototyping und Usability-Testing von Mobilien Systemen im originalen Benutzungskontext*. Bremen, Deutschland, Universität Bremen., Dissertation (2010). – URL: <http://elib.suub.uni-bremen.de/diss/docs/00011859.pdf>,
- [Kru06] KRUG, S.: *Don't Make Me Think*. New Riders, 2006
- [LG04] LEE, K.B. ; GRICE, R.A.: Developing a new usability testing method for mobile devices. In: *Professional Communication Conference, 2004. IPCC 2004. Proceedings. International* (2004), S. 115–127
- [Lik32] LIKERT, R.: A technique for the measurement of attitudes. In: *Archives of psychology* (1932)
- [LPWR90] LEWIS, C. ; POLSON, P.G. ; WHARTON, C. ; RIEMAN, J.: Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In: *Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people* (1990), S. 235–242
- [MS09] MACK, Z. ; SHARPLES, S.: The importance of usability in product choice: A mobile phone case study. In: *Ergonomics* 52 (2009), Nr. 12, S. 1514–1528
- [NHH⁺08] NIEGEMANN, Helmut M. ; HESSEL, Silvia ; HUPFER, Matthias ; DOMAGK, Steffi ; HEIN, Alexandra ; ZOBEL, Annett: *Kompendium multimediales Lernen*. Berlin, Heidelberg : Springer-Verlag, 2008 (X.media.press)

- [Nie94a] NIELSEN, J.: Guerrilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier. (1994). http://www.useit.com/papers/guerrilla_hci.html, Abruf: 30.11.2011
- [Nie94b] NIELSEN, J.: Usability inspection methods. In: *Conference companion on Human factors in computing systems* (1994), S. 413–414
- [Nie94c] NIELSEN, Jakob: *Usability Engineering*. San Francisco, California : Morgan Kaufmann Publishers, 1994
- [Nie00] NIELSEN, J.: Why you only need to test with 5 users. 19 (2000). <http://www.useit.com/alertbox/20000319.html>, Abruf: 28.07.2011
- [Nie03a] NIELSEN, J.: Recruiting Test Participants for Usability Studies. (2003). <http://www.useit.com/alertbox/20030120.html>, Abruf: 23.10.2011
- [Nie03b] NIELSEN, J.: Usability 101: Introduction to Usability. (2003). <http://www.useit.com/alertbox/20030825.html>, Abruf: 23.10.2011
- [Nie11] NIELSEN, J.: Salary Trends for Usability Professionals. (2011). <http://www.useit.com/alertbox/salaries.html>, Abruf: 23.10.2011
- [NL93] NIELSEN, J. ; LANDAUER, T.K.: A mathematical model of the finding of usability problems. In: *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (1993), S. 206–213
- [OST71] OSGOOD, C.E. ; SUCI, G.J. ; TANNENBAUM, P.H.: *The measurement of meaning*. University of Illinois Press, 1971
- [PHVS04] PO, S. ; HOWARD, S. ; VETERE, F. ; SKOV, M.B.: Heuristic evaluation and mobile usability: Bridging the realism gap. In: *Mobile Human-Computer Interaction–MobileHCI 2004* (2004), S. 591–592
- [RC08] RUBIN, J. ; CHISNELL, D.: *Handbook of Usability Testing: How to plan, design and conduct effective tests*. Wiley-India, 2008
- [Res11] RESEARCH, ABI: 44 Billion Mobile App Downloads by 2016. (2011). <http://www.abiresearch.com/press/3668-44+Billion+Mobile+App+Downloads+by+2016>, Abruf: 13.12.2011
- [Ric10] RICHTER, Markus Flückiger M.: *Usability Engineering kompakt. Benutzbare Software gezielt entwickeln*. 2. Heidelberg : Spektrum, 2010

- [ROH⁺04] ROTO, V. ; OULASVIRTA, A. ; HAIKARAINEN, T. ; KUORELAHTI, J. ; LEHMUS-KALLIO, H. ; NYSSONEN, T.: Examining mobile phone use in the wild with quasi-experimentation. In: *Helsinki Institute for Information Technology (HIIT), Technical Report 1* (2004)
- [Rot11] ROTHSTOCK, Karin: MOBILE EFFECTS - wie geht die mobile Reise in 2011 weiter? In: *Mobile Effects 2011* (2011). http://www.tomorrow-focus-media.de/fileadmin/customer_files/public_files/downloads/studien/Mobile_Effects_2011.pdf, Abruf: 13.09.2011
- [Rub95] RUBIN, J.: *Handbook of usability testing*. Washington: Society for Technical Communication, 1995
- [Sch10] SCHRODER, Anders.: *The Usable Mobile Phone*. <http://theusablemobilephone.com/>, 2010
- [SM08] SEFFAH, Ahmed ; METZKER, Eduard: *Adoption-centric Usability Engineering: Systematic Deployment, Assessment and Improvement of Usability Methods in Software Engineering*. 1. Springer, 2008
- [Sta04] STAAL, M.A.: Stress, cognition, and human performance: A literature review and conceptual framework. In: *NASA Technical Memorandum 212824* (2004)
- [SWL07] SCHUSTERITSCH, R. ; WEI, C.Y. ; LAROSA, M.: Towards the perfect infrastructure for usability testing on mobile devices. In: *CHI'07 extended abstracts on Human factors in computing systems* (2007), S. 1839–1844
- [TA08] TULLIS, T. ; ALBERT, W.: *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Morgan Kaufmann, 2008
- [tob11] TOBII.COM: Tobii Introduces Most Precise Eye Tracking Solution Available for Mobile Device Testing. (2011), May. <http://www.tobii.com/en/group/news-and-events/press-releases/>, Abruf: 14.10.2011
- [VR10] VÄÄTÄJÄ, H. ; ROTO, V.: Mobile questionnaires for user experience evaluation. In: *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems* (2010), S. 3361–3366
- [WBH88] WHITESIDE, J. ; BENNETT, J. ; HOLTZBLATT, K.: Usability engineering: Our experience and evolution. In: *Handbook of human-computer interaction* (1988), S. 791–817

- [WRLP94] WHARTON, C. ; RIEMAN, J. ; LEWIS, C. ; POLSON, P.: The cognitive walkthrough method: A practitioner's guide. In: *Usability inspection methods* (1994), S. 105–140

Erklärung der Urheberschaft

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form in keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ort, Datum

Unterschrift