Mobile Usability Evaluation – Field-Testing vs Labor-Testing

Jandl Christian
FH Sankt Pölten
Stift 19
A-3321 Ardagger
+43-664-1840997
dm141534@fhstp.ac.at

1. Abstract
Diese Arbeit befasst sich mit wissenschaftlichen Artikeln zur Usability Evaluation von Applikationen auf mobilen Endgeräten. In Form einer Literaturrecherche werden aussagekräftige Papers, welche sich mit dem Vergleich von Labor- und Field-Testing beschäftigen, analysiert und anschließend miteinander verglichen. Es wird dargelegt, dass beide Methoden zu einem gutem Ergebnis führen können, und das hauptausschlaggebend für eine zielführende Evaluation ein gutes Testsetting ist. Das Ziel dieser Arbeit ist es, einen objektiven Blick auf beide Evaluations-Techniken beim Testen von mobilen Anwendungen zu generieren um bei anstehenden Usability-Tests leichter zu einer Entscheidung der zu verwendenden Technik zu gelangen.

General Terms
H.5.2 User interfaces (evaluation/methodology)

Keywords
Usability evaluation, field test, laboratory test, experimental comparison.

2. Einleitung:
Smartphones und Tablets werden von einer immer breiter gestreuten Gruppe von Nutzern verwendet, dementsprechend veränderte sich auch die Nutzungsumgebung der mobilen Applikationen.
In einem bereits 1998 veröffentlichten Artikel der mobileHCI (Internationale Konferenz für Mobile Human Computer Interaction) ermutigte Johnson Usability-Experten sich in den Methoden zur Datenerfassung für die Bewertung mobiler Geräte und Anwendungen zu vertiefen. Er kritisierte, dass die herkömmlichen Usability-Labore nicht in der Lage sind, wichtige Aspekte wie Umwelteinflüsse, Nutzungsumgebung und gleichzeitig ausgeführte Aktivitäten des Nutzers, angemessen zu simulieren [1].
Trotz dieser frühen Erkenntnisse Johnsons zeigte Kjeldskov und Grahams in einer Studie, die auf eine Befragung namhafter Usability-Institute zwischen 2000 und 2002 basiert, dass 71% aller Evaluierungen von mobilen Geräten und Services im Labor durchgeführt wurden [2].

Die schnelle Entwicklung mobiler Web-Technologien hat zu einem exponentiellen Wachstum der Zahl von unerfahrenen Anwendern geführt. Auch Menschen, denen Kenntnisse in der Informatik fehlen, benutzen Tablets und Smartphones häufiger denn je. So gibt es nach Hassenzahl keine Garantie, dass die Benutzer auch tatsächlich das Produkt als solches wahrnehmen und in dieser Weise schätzen wie die Designer dachten, dass es wahrgenommen und geschätzt werden würde [3]. Darum ist es von bedeutender Wichtigkeit, dass die Gebrauchstauglichkeit von mobilen Anwendungen Usability Evaluierungen unterzogen wird, um reelle Einschätzungen der künftigen Nutzung zu erhalten. In dieser Arbeit, die auf einer Literaturrecherche beruht, werden verschiedene Publikationen der wissenschaftlichen Plattformen ACM und IEEE untersucht, die sich mit dem Vergleich von Usability Testing im Labor und dem bei Field-Testing beschäftigen.

2.1 DEFINITION USABILITY
Um die Gebrauchstauglichkeit von Software steigern zu können, sind folgende drei Aspekte zu berücksichtigen:

• Sie soll effizienter werden Es soll weniger Zeit in Anspruch genommen werden, um eine definierte Aufgabe erfüllen zu können.
• Sie soll leichter erlernbar werden Die Funktionen der Software sind in der Benutzerfläche ersichtlich und können vom Benutzer leicht erlernt werden.
• Sie soll benutzerfreundlicher werden Die Erwartungen des Benutzers an die Software sollen erfüllt werden und die User-Experience (UE) so gut als möglich sein. [4].

Usability ist die Herrausforderung der Software-Entwickler die Applikationen mit einem möglichst hohen Maß an User-Expierience auszustatten ohne auf den geforderten Funktionsumfang zu verzichten. Usability und User Experience werden von MacDonald und Atwood als die vierte und fünfte Evolutionsstufe der Mensch-Computer-Interaktion angeführt [5]. In den Anfängen der Computerwissenschaften waren Betriebssicherheit und Performance wesentliche Indikatoren für die Qualität einer Software. Diese Eigenschaften wurden durch den technologischen Fortschritt und die dadurch resultierende geringere Fehleranfälligkeit der Computer mittlerweile ein Standard der Software-Branche. Somit verschob sich der Fokus

der Forschung in Richtung der Benutzerfreundlichkeit und der User-Experience. Diese sind durch die breite Streuung der Nutzerschichten wesentliche Faktoren in der zu erwartenden Verbreitung und dem daraus resultierenden wirtschaftlichen Erfolg einer Applikation.

2.2 Usability Testing

Usability-Tests sind ein wissenschaftlich valides Werkzeug um die Gebrauchstauglichkeit einer Applikation beurteilen zu können. Usability-Tests werden in der Regel mit Hilfe eines Loud Thinking - Protokoll durchgeführt. Den Testpersonen werden in einer definierten Testumgebung Aufgaben gestellt und ermutigt, während Erfüllung dieser laut zu denken. Das gibt Usability-Experten notwendige Informationen zu erkennen, ob die Mensch-Maschine-Schnittstelle der natürlichen menschlichen Denkweise entspricht [6].
Die gefundenen Probleme der Testpersonen werden anschließend gewichtet und nach deren Schwere klassifiziert.
Die Schwere eines Usability-Problems ist ein wichtiger Faktor bei der Festlegung der Dringlichkeit von Maßnahmen. Die dringlichsten Maßnahmen sind die, die verhindern, dass der Nutzer die Aufgabe abschließen kann. Dumas und Redish (1993) etablierten eine 4-Punkte-Skala, wobei der Schweregrad 1 die größten Problemen darstellen und 4 die Probleme mit den geringsten Auswirkungen.

- **Level 1:** Dieses Problem verhindert das Lösen der Aufgabenstellung
- **Level 2:** Probleme dieser Art führen zu Verzögerungen und Frustation der Testperson
- **Level 3:** Problem hat negativen Einfluss auf die Usability
- **Level 4:** Subtiles Problem, bei Weiterentwicklung der Anwendung berücksichtigen [7].

2.3 Mobile Usability

Mit der Bereitstellung von mobilen Technologien wurde es notwendig, Usability-Methoden an die technischen Neuerungen anzupassen, zu erweitern oder gänzlich neue Methoden zu entwickeln, um auf die Veränderungen in der Nutzungsumgebung und der Bedienung auf mobilen Geräten reagieren zu können.
Für Zhang [8] beinhaltet Mobile Usability neue mobil-bezogene Herausforderungen wie: mobiler Kontext, Verbindungsverfügbarkeit, kleine Bildschirmgrößen, verschiedene Bildschirmauflösungen, begrenzte Verarbeitungskapazität und Dateneingabemethoden.
Weiters haben Hersteller von mobilen Geräten und Betriebssystemen die Umsetzung ihrer eigenen Usability-Richtlinien eingefordert. Beispielsweise legen die Apple Human Interface Guidelines für iOS [9] Interaktionselemente fest, die es bei der Anwendungserstellung zu berücksichtigen gilt.
Beispiele dafür sind:
Die Interaktion mit dem Multi-Touch-Display, die Anzeige in verschiedenen Auflösungen und Abmessungen, das Reagieren auf Device-Orientation und vorgegebene Gesten wie Tap, Flick und Pinch.
Genauso wie Apple entwickelte auch der Hersteller des zweiten großen mobilen Betriebssystems, Google [10], Interface Guidelines für Android, mit dem Ziel, Entwickler zu einem einheitlichen Interaktions-Schema zu bewegen. Insbesondere Eigenschaften wie Touch-Bewegungen, Größe und Lage der Icons und Buttons, kontextabhängige Menüs und ihre Anpassung an Responsive Design, sowie Größe und Format von Text sollen so normiert werden.
Somit ergibt sich für die Usability Evaluation von mobilen Anwendungen einen viel größeren Einflussbereich auf die Qualität der Gebrauchstauglichkeit als dies bei Desktopanwendungen der Fall ist.

Weitere Einflussfaktoren auf die Usability und der User-Expierence im Bereich der mobilen Gebrauchstauglichkeit ist die Rücksichtnahme auf die Nutzungsumgebung und die Tätigkeiten, die während einer Interaktion mit dem mobilen Gerät zeitgleich ausgeführt werden und so zu einer verminderten Aufmerksamkeit des Nutzers führen. Dies ist im vermehrten Ausmaß zu berücksichtigen, wenn die Applikation zur Orientierung oder während sportlichen Tätigkeiten genutzt werden soll. Diesen Einflüssen gerecht zu werden kamen Zhang & Adipat 2005 zu dem Schluss, dass das Testen von Usability im Labor keine überzeugende Ergebnisse liefert [8]. So entstand die Methode des Field-Testings. Nun wurden Usability-Tests dort durchgeführt, wo man annahm, dass auch der Nutzer im realen Anwendungsfall dies tun würde.

2.4 Usability Labor Testing

Usability-Labor-Tests finden in einem Usability-Labor, welches speziell für Untersuchungen der jeweiligen Applikation eingerichtet wird, statt. Meistens verfügen die Labore standardmäßig über eine klassische Wohnzimmer- oder Büroeinrichtung.
Darin werden Nutzungsszenarien, die einem bestimmten Anwendungsfall der Applikation entsprechen nachgestellt. Die Testperson soll sich so in einer natürlichen Umgebung, in derer sie auch im realen Nutzungskontext die Anwendung benützen würde, die gestellten Aufgaben lösen. Die Testperson befindet sich allein oder in Anwesenheit eines Usability-Ingenieurs im Testraum. In einem weiteren Raum befinden sich die technischen Aufzeichnungsgeräte und die Steuerung der verschiedenen Kameras die sich im Testraum befinden. Die Abtrennung besteht zumeist aus einer verspiegelten Oberfläche, durch der es den Testern möglich ist, die Testperson bei der Aufgabenerfüllung zu beobachten, ohne dass sich diese dadurch gestört fühlen könnte.
Durch Nachempfinden einer realen Umgebung und den vorgegebenen Aufgaben an die Testperson, ist es möglich alle Aspekte der Gebrauchstauglichkeit zu testen. Weiters können äußere Einflüsse vermieden werden, die keine Relevanz auf die Testergebnisse haben.
Diese Vorteile sind vor allem relevant, wenn der Test den Vergleich von verschiedenen Designs oder Interfaces zum Ziel hat, da die Testperson sich sehr gut auf die Aufgabe konzentrieren kann. Weiters ist es in Usability-Labors ohne großen Aufwand möglich, die Benutzeraktivität und das Verhalten der Testperson mittels Kameras und Mikrofone, die in der Einrichtung unsichtbar verbaut sind, aufzuzeichnen um später diese Daten zu analysieren.
Allerdings können durch das Isolieren der Testperson in Usability-Laboren äußere Faktoren, die die User-Experience beeinflussen würden, nicht erfasst werden. Da die Testperson sich stets der Ausnahmesituation der Testumgebung bewusst ist, kann es zu möglichen Messfehlern kommen [11].

2.5 Usability-Field-Testing

Eine Feldstudie ist eine wissenschaftlich anerkannte Methode zur Erfassung von Daten über Benutzer, Benutzerbedürfnisse und den Produktanforderungen. Infolge der weiten Verbreitung von mobilen Geräten und dem Nutzungskontext ortsabhängiger

Applikationen wie beispielsweise Software zur Orientierung spielen sie auch bei Mobile Usability Evaluierungen mittlerweile eine gewichtige Rolle zur Verbesserung der Gebrauchstauglichkeit. Durch Beobachtungen und Interviews der Testpersonen im Nutzungsumfeld können weitere relevante Daten zur Produktverbesserung gesammelt werden. Bei der Datensammlung liegt der Fokus auf dem Verlauf der Aufgabe, Ineffizienzen der Anwendung und der physikalischen Umgebung der Testperson.
Usability-Tester in Feldstudien beobachten die Testperson, stellen ihnen Fragen und notieren sich Beobachtungen, während die gestellten Aufgaben zu lösen versucht.
Diese Methode ist sehr nützlich in einem frühen Stadium der Produktentwicklung, wenn in Erfahrung gebracht werden soll, wo die genauen Anforderungen der Nutzer an die Applikation liegen.

Bei Field-Testing muss der User seine kognitiven Ressourcen auf das Benutzen der Applikation und das Ausführen einer primären Tätigkeit wie zum Beispiel der Fortbewegung aufteilen Blumenstein und Schmiedl [12] kamen 2011 zu der Erkenntnis, dass auch mit vergleichsweise günstigen Aufzeichnungsmethoden valide Ergebnisse zu erzielen sind. Sie testeten ein professionelles mobiles Eyetracking-System und ein selbst zusammengestelltes und modifiziertes System mit ähnlichen Komponenten. Der Unterschied in der Anschaffung betrug mehr als 20.000 € ohne dass sie jedoch signifikante Unterschiede in den Testergebnissen feststellen konnten. Vor allem die Kalibrierung des Eyetrackingsystems bei der High-Budget-Variante erwies sich als umständlich und zeitaufwendig, ohne bessere Ergebnisse zu liefern. Somit kann davon ausgegangen werden, dass durch die Auswahl der richtigen Komponenten und der Fähigkeit der Usability-Tester, Kameras und Mikrofone so anzubringen, dass sie die Testperson nicht behindern, Field-Testing eine vergleichsweise günstige Evaluierungsmethode mit höherem Zeitaufwand als im Labor darstellt [11].

3. Usability Tests im Labor und im Feldversuch

Diese Literaturrecherche befasst sich mit den Ergebnissen wissenschaftlicher Berichte, die jeweils die gleiche Applikation im Usability-Labor und im Field-Test evaluierten. Im Folgenden werden die wesentlichen Erkenntnisse dieser Studien zusammengefasst und am Ende dieser Arbeit miteinander verglichen, um Gemeinsamkeiten oder auch signifikante Unterschiede zu analysieren. Daraus erhofft sich der Autor ein schlüssiges Bild über die Einsatzmöglichkeiten der verschiedenen Techniken zu erhalten.

3.1 Vergleich einer Consumer-Applikation

Im Jahr 2005 veröffentlichten A. Kaikokonen und Titti Kallio eine Studie, in der sie unter gleichen Vorraussetzungen Usability Tests im Labor und im Feldversuch unternahmen [13].
Als Testapplikation verwendeten sie “Mobil Wire”, die sich zu diesem Zeitpunkt noch in der Entwicklung befand. Mit Hilfe dieser App ist es möglich Dateien vom Smartphone auf den Desktop-PC zu verschieben. Als Testpersonen dienten zwei Gruppen zu je 20 Personen mit annähernd gleichen Durchschnittsalter und ähnlicher technischer Affinität. Die ungewöhnliche Größe der Gruppe (bei normalen Tests 5-10 Personen) begründeten sie mit der Aussage Faulkners aus dem Jahr 2003 dass 95% der Usability Probleme bei einer Größe von 20 Testpersonen gefunden werden. Die Tests wurden sowohl im Labor sowie auch im Feldversuch durchgeführt. Den Testpersonen wurden 10 Aufgaben gestellt und sie wurden aufgefordert während der Lösung der Aufgaben laut zu denken. Der Labortest wurde in einem professionellen Usabilitylabor durchgeführt, wohingegen im Feldversuch die Testpersonen die Vorgabe bekamen, zur U-Bahnstation zu gehen, mit dieser zu einem Einkaufszentrum zu fahren und dort einen Bekannten zu treffen. Der Testleiter folgte unauffällig und gab der Testperson die zu lösenden Aufgaben per Funk durch.

Insgesamt hatten die Testpersonen zehn Aufgaben zu lösen, dabei wurden sowohl im Feldversuch wie auch im Labortest die gleichen 46 Usabilityprobleme gefunden. Daraus wurde eine Liste von 22 Problemen extrahiert, welche im Labor- und im Feldtest öfter als einmal angeführt wurden. Neun davon waren schwerwiegende Probleme und führten zu einem Abbruch der gestellten Aufgabe. Bei drei Usability-Problemen war ein signifikanter Unterschied zwischen Feldversuch und Labortest in der Anzahl der Personen die dieses Problem feststellten, zu erkennen. Alle drei wurden öfter im Feldversuch bemerkt. Es ließ sich jedoch nicht eindeutig belegen, wie dies zustande kam.

Die Ausführungsdauer der gestellten Aufgaben war sowohl im Feldversuch wie auch im Laborversuch gleich lang, nur die Vorbereitungszeit war bei Field-Testing doppelt so lang wie im Labor. Im Gesamten waren die Forscher jedoch von den ähnlichen Testergebnissen überrascht, da sie erwartet hatten, dass beim Feldversuch mehr Usability-Probleme gefunden werden würden. Die wesentlichen Erkenntnisse waren, dass es bei Consumer Applikationen kein wesentlicher Unterschied zwischen Labor- und Feldtests zu erkennen ist. Bei Feldversuchen ist die Vorbereitung wesentlich zeitaufwendiger, man muss ungefähr die doppelte Zeit pro Testperson kalkulieren. Dafür scheint es, als ob die Testpersonen im Feldversuch auskunftsfreudiger sind und offener über ihr Empfinden beim Ausführen der Aufgaben sprechen. Weiters wiesen sie auf die Gefahr hin, dass im Feldversuch es immer wieder zu unvorhergesehenen Zwischenfällen kommen kann, die den Test im schlimmsten Fall negativ beeinflussen können.

3.2 Vergleich einer Business-Applikation I

2004 publizierten Kjeldskov und Skov [14] das sogenannte Hassle-Paper. Ziel dieser vergleichenden Studie war es, eine Business-Applikation im Labor und parallel in der realen Nutzungsumgebung auf Gebrauchstauglichkeit zu testen.
Die Testpersonen waren jeweils sechs professionelle Pflegekräfte im Gesundheitsbereich mit gleichen IT-Grundkenntnissen. Beide Evaluierungen nutzen dieselbe Applikation, die zur Patientendatenverwaltung eingesetzt werden sollte. Der Feldversuch fand in einem Krankenhaus während der normalen Arbeitszeiten mit echten Patienten statt, während der Laborversuch in einem Usability-Labor stattfand. Dieses wurde so umgestaltet, damit es dem Setting eines Krankenhauses entsprach. Es wurden mehrere Krankenzimmer mit Betten und Tischen eingerichtet, in denen sich Akteure befanden, die die Patienten darstellten.
Die Datenanalyse brachte für jede Testumgebung eine Liste mit Usability-Problemen hervor, worin die Schwere (siehe Kapitel 2.2) und die Häufigkeit des Problems dargelegt wurden. Sie fanden heraus, dass in der Laborstudie mehr Usability-Probleme gefunden wurden. Hier fanden die Testpersonen 36, im Feldversuch 23 Usability-Probleme. 14 Probleme wurden nur im Labor gefunden, wohin gegen nur ein Problem im Feldversuch

gefunden wurde, dass im Labor nicht auftrat. Unterschiedlich war auch die benötigte Zeit für den Test. Im Labor wurden 34 und im Feldversuch 65 Mitarbeiterstunden benötigt. Basierend auf dieser empirischen Studie kamen sie zu folgenden Ergebnis:
"*expensive time in the field should perhaps not be spent on usability evaluation if its possible to create a realistic laboratory setup including elements of context and requiring mobility*" und "*field studies may instead be more suitable for obtaining insight needed to design the system right in the first place*" [Kjeldskov, 2004] .

3.3 Business-Applikation II

2006 brachte Nielsen [15] aus Dänemark seine Antwort auf das sogenannte Hassle-Paper in Form einer eigenen Testreihe. Er testete ebenfalls eine Business-Applikation sowohl im Labor als auch im Feldversuch, kam jedoch zu unterschiedlichen Ergebnissen.

Getestet wurde ein mobiles System, das von Facharbeitern zur Aufnahme des Materialbedarfs, der Kilometerleistung und der Zeit eines Auftrages verwendet wurde. Dafür wurde ein Barcode-Scanner mit dem Mobilgerät verbunden. GPRS wurde für die Datenübertragung verwendet. Die Anwendung war Teil einer Verwaltungsapplikation, die jedoch nicht in dieser Studie evaluiert wurde. Der Barcode-Scanner wurde verwendet, um von einer Werkzeug- und Materialliste die benötigten Elemente zu scannen. Die Testteilnehmer waren alle auszubildende einer technischen Schule und sollten Fachkräfte der Erdbautechnik simulieren. Insgesamt waren 14 Testpersonen beteiligt, welche in zwei gleich große Gruppen geteilt wurde. Alle waren im täglichen Umgang mit Mobilgeräten vertraut, jedoch hatte noch niemand zuvor mit Barcode-Scannern hantiert. Die Evaluierung im Labor fand im klassischen Setting eines Usability-Testraumes statt. Die Testpersonen saßen an einem Tisch und arbeiteten dort an den gestellten Aufgaben. Der Feldversuch fand in einem Warenlager statt, wo die Testpersonen selbst ihren Arbeitsbereich wählen mussten, da keine Tische vorhanden waren.

Insgesamt wurden 76 verschiedene Usability-Probleme erkannt, 27 davon wurden als kritisch, 30 als erheblich und 19 als leicht eingestuft. In der Labor-Evaluation wurden 104 Vorkommen von Usability-Probleme und in der Feldauswertung 123 Fälle ermittelt.

Nach der Entfernung von Instanzen gleicher Usability-Probleme blieben 48 verschiedene Probleme im Labor und 60 verschiedene Probleme im Feldversuch übrig. 44 der insgesamt 76 Probleme traten nur in einem Testsetting auf. Dieses Ergebnis legt nahe, dass es von Bedeutung sein könnte, jeweils beide Evaluierungen durchzuführen.

Die Dauer der Aufgabenerfüllung war bei den meisten Aufgaben im Feldversuch länger als im Labor, jedoch fiel dies nicht so signifikant wie in den beiden anderen Testberichten aus.
Zusätzlich zu der Auswertung beider Testmethoden erstellten sie noch eine Tabelle, die die gefunden Usability-Probleme in Bereich wie Ergonomie, kognitive Last, Information, Interaktionsstil, Sichtbarkeit unterteilte. Darin wurde sichtbar, dass die nur im Feldversuch gefundenen Probleme hauptsächlich unter die Bereiche Kognitive Last, Arbeitsablauf, Interaktionsstil fielen, was die These der Forscher unterstützt, dass Field-Testing besser in der Lage ist Usability-Probleme zu finden, als Labortests, auch wenn diese zeitaufwendiger ist.

„Thus the overall conclusion of is that it is worthwhile conducting user-based usability evaluations in the field, even though it is more complex and time-consuming. The added value is a more complete list of usability problems that include issues not detected in the laboratory setting" [Nielsen,2006].

4. Zusammenfassung der Testreihen

In diesem Kapitel der Arbeit werden die wesentlichen Aspekte der verschiedenen Testreihen miteinander verglichen. Dies soll helfen, die Testreihen besser analysieren zu können und die Validität der Ergebnisse einstufbar zu gestalten.

4.1 Vergleich der Labor-Settings

Die Usability-Evaluation der Consumer-Applikation "Mobil Wire" und die der Business-Applikation in Verbindung mit dem Barcode-Scanner wurde beide in professionellen Usability-Labors vorgenommen, ohne dass jedoch die vorhandene Einrichtung maßgeblich verändert wurde. Dies konnte zur Folge haben, dass sich die Testperson ständig der Testsituation bewusst ist und im schlimmsten Fall sich selber getestet fühlt. Dies beeinflusst in unbestimmten Grad das Ergebnis der Evaluierung womit ein valides Ergebnis dadurch schwer zustande kommen kann. Diese Labortests sind zwar reproduzierbar und die Vorbereitungszeit zur Testreihe und die Konfiguration für die einzelnen Testpersonen ist gering, doch wird der Test dadurch in einer unnatürlichen Umgebung, in welcher die Software wahrscheinlich nie zum Einsatz kommen würde, getestet.

Für das Laborsetting von Business-Applikation I wurde ein Usability-Labor so umkonstruiert, damit es dem Feldversuch so nahe wie möglich kommt. Die Testpersonen wussten zwar, dass sie sich im einem Usability-Labor befinden, konnten aber anhand der naturgetreuen Nachbildung von Gängen und für den Test extra engagierten Patienten in Krankenbetten die Testsituation weniger intensiv wahrnehmen und ihrer normalen Tätigkeit nachgehen und sogleich getestet werden. Wie viel Aufwand es war, das Labor derart naturgetreu zu gestalten wird in der Studie nicht angeführt, es ist jedoch davon auszugehen, dass dies einen großen Teil der gesamten Vorbereitungszeit in Anspruch genommen hat.

4.2 Natürliche Umgebung im Feldversuch

In Business-Applikation I wurde die Usability-Evaluierung während dem normalen Betrieb in einem Krankenhaus durchgeführt. Die Testpersonen waren Angestellte und somit in vertrauter Umgebung. Auch die Patienten waren echt.

Business-Applikation II hingegen fand in einer Warenhalle statt, wo die Testpersonen selber wählen konnten wo sie die gestellten Aufgaben lösen möchten. In der Arbeit selbst finden sich nur wenig Angaben wie die Testpersonen darauf reagierten. Einziger Hinweis ist ein Foto einer Testperson, die am Boden kniend versucht einen Barcode auf der Tabelle zu scannen. Ob dies wiederum den tatsächlichen Nutzungsbedingungen der Anwendung entspricht wurde nicht angeführt.

Im Fall der Consumer-Anwendung ist das Field-Setting als realitätsnah gewählt. Den Anwendern wurde eine Route vorgegeben auf der sie die kurze Strecken im Außen- und Innenbereich zu Fuß gingen aber auch die U-Bahn zur Fortbewegung nutzen. Sie sollten eine bekannte Person im Einkaufszentrum treffen. Die Testpersonen reagierten auf komplizierte Aufgaben, indem sie sich ruhige Plätze suchten und stehenblieben. Einfache Aufgaben lösten sie im Gehen ohne dass es zu besonderen Zwischenfällen kam. In diesem Feldversuch ist anzumerken, dass das Testsetting einer Consumer-Applikation gut entspricht. Ob es für diese Applikation jedoch als realitätsnahes

Testszenario einzustufen ist, bleibt fraglich. Das Verschieben von Dateien zwischen Smartphone und Desktop-PC findet meist deshalb statt, um sie im nächsten Arbeitsschritt am PC zu bearbeiten. Darum ist eher davon auszugehen, dass man die Applikation großteils in einem Büro nutzen wird.

4.3 Die Wahl der Testpersonen

In jeder der drei Studien wurde darauf geachtet, dass die Testpersonen ähnliche Fähigkeiten im Umgang mit technischen Geräten haben. Bei der Consumer-Applikation wurden die meisten Testpersonen eingesetzt (20 Pers/Gruppe). Die hohe Anzahl wurde mit der Aussage von Faulkner begründet, dass 20 Personen 95% [16] der Usability-Probleme finden würden. Ob diese ungewöhnliche Gruppengröße auch den gewünschten Effekt erzielt hat, wird nicht näher ausgeführt.

Bei den beiden anderen Arbeiten entspricht die Größe der Gruppen jeweils einer realistischen Zahl bei professionellen Usability-Tests (5 – 10 Personen finden 80% der Usability-Probleme) [17]. Business-Applikation I verwendet als Testpersonen die zukünftigen Anwender, dies ist auf jeden Fall ein Vorteil gegenüber Business-Applikation II welche Schüler einer technischen Ausbildungstätte einsetzt, obwohl die App für Fachkräfte im Hoch- und Tiefbau konzipiert wurde. Inweit dies einen Einfluss auf das Testergebnis hat ist fraglich.

4.4 Ergebnisse im Vergleich

Die Ergebnisse in Qualität und Quantität der gefundenen Usability-Problemen in den drei Testreihen lässt keine valide Aussage zu, ob nun Labor- oder Field-Testing die bessere Variante ist. In der Evaluation der Consumer-App traten in beiden Testumgebungen dieselben Probleme auf, einzig in der Häufigkeit unterschieden sie sich. Hier erschien der Feldversuch als geeigneter.
Die Ergebnisse der Business-App I ergaben ein völlig anderes Bild. Hier wurden von insgesamt 36 Usability-Problemen 14 Beanstandungen der Gebrauchstauglichkeit nur im Labor erkannt, wohingegen in der realen Umgebung nur ein Problem gefunden wurde, dass im Labor unerkannt blieb.
Business-Applikation II brachte wieder ein anderes Ergebnis. Hier wurden nach Bereinigung gleicher Probleme in den Labortests 48 und im Feldversuch 60 verschiedene Usability-Probleme ermittelt. Die hohe Anzahl der gefundenen Probleme lässt hier den Schluss zu, dass sich die Applikation in einer sehr frühen Entwicklungsstufe befunden hat.In Tabelle 1 werden die Ergebnisse der drei Studien vergleichend dargestellt.

Applikation	Testpersonen		Usability-Probleme		Zeitaufwand relativ	
	Labor	*Field*	*Labor*	*Field*	*Labor*	*Field*
Mobile Wire *Consumer*	20	20	22	22	1	2
Krankenhaus *Business*	6	6	36	23	1	2
Hoch-Tiefbau *Business*	7	7	48	60	1	1.3

Tabelle 1 Auflistung der Testergebnisse

5. Diskussion

Die Frage, ob Field-Testing oder Labor-Testing angemessener ist, lässt sich auch im Einzelfall nie zu gänzlichen Gewissheit entscheiden.

Im September 2014 erschien eine Studie von Kjeldskov und Skov [18], die sie *"Was it worth the hassle?"* nannten. Darin erforschten sie die Folgestudien ihrer eigenen 2004 erschienen Studie, dem sogenannten Hassle-Paper, auf welches in einem vorigen Kapitel bereits näher eingegangen wurde. Diese Arbeit wurde insgesamt 191 mal in anderen wissenschaftlichen Arbeiten zitiert und eignete sich daher gut als Ausgangsbasis, um zu sehen, in welcher Art und Weise andere Forscher ihre Ergebnissen bewerteten. Von den 191 Arbeiten, die Kjeldskov und Skov zitierten blieben nach dem ersten Querlesen 142 Arbeiten über, die tatsächlich Bezug auf das Hassle-Paper nahmen und in einer Sprache verfasst waren, die die Forscher lesen konnten. Diese teilten sie in drei Gruppen:

- Arbeit behandelt entweder Labor- oder Field-Testing (62 Arbeiten)
- Arbeit vergleicht Labor- und Field-Testing (16 Arbeiten)
- Diskussion über Labor- oder Field-Testing (64 Arbeiten)

Die große Anzahl dieser Arbeiten zeigt, welchen Einfluss diese Studie auf nachfolgende Forschungsprojekte hatte. Anschließend wurden Arbeiten analysiert und die interessantesten Forschungsergebnisse zusammengefasst. Im Zusammenhang mit dieser Arbeit sind die Ergebnisse aus der Gruppe die ebenfalls einen direkten Vergleich zwischen Labor- und Feldversuch unternahmen, am aussagekräftigsten. Hier ist zu bemerken, dass trotz der intensiven Diskussion die nach dem Erscheinen des Hassle-Papers in Gang gesetzt wurde, keine der 16 vergleichenden Arbeiten die Ergebnisse von Kjeldskov und Skov anzweifelten oder versuchten in einer empirischen Studie die Testreihe zu reproduzieren. Anstatt die Ergebnisse zu überprüfen, wurden immer neue Projekte evaluiert. Dadurch variierten die Simulationen im Labor sehr stark, was wiederum Kjeldskovs Aussage widerspricht, dass im Labor durchgeführte Studien so realitätsnah wie möglich sein sollen um dieselben oder mehr Usability-Probleme zu finden als im Feldtest. Weiters stellten sie fest, dass nicht klar definiert ist, was einem Feldtest entspricht und wie dieser auszuführen ist. So finden sich Feldversuche wo die Testpersonen in öffentlichen Verkehrsmittel fahren, zu Fuß gehen oder sich in Geschäften befinden. Andere Feldversuche wiederum finden in einem Fußballstadium oder nur im Außengelände des Forschungsinstituts statt. Dies wirft die Frage auf, welche Grundsätze ein Feldtest zu erfüllen hat, um als solcher zu gelten.

Der Umstand dass man Feldversuche nicht miteinander vergleichen kann, da das Feld nie gleich ist, macht es auch schwierig Labor und Feldversuch miteinander zu vergleichen und es scheint, dass der Ausgang einer solchen Studie nie vorher mit Gewissheit zu sagen ist.

"Since no answer to the lab versus field question seems to be found, we have argued that the important question is not if or why one should do lab or field studies, but rather when we should do what, and how we should then do it." [Kjeldskov, 2014]

6. Fazit

Auch wenn es nicht einfach ist, die Gemeinsamkeiten der unterschiedlichen Studien zu extrahieren und zu einem eindeutigem Ergebnis zu gelangen, sind im Zuge dieser Arbeit einige Fragen beantwortet worden.

Feldversuche benötigen in jedem der untersuchten Fälle mehr Vorbereitungsaufwand als Tests im Labor. Durschnittlich muss mit der doppelten Zeit pro Testperson gerechnet werden.

Weiters wurde erkannt, dass der Feldversuch auf jeden Fall ein wichtiges Werkzeug beim Finden von Usability-Problemen darstellt. Besonders hilfreich können sie in einem frühen Entwicklungstadium der Applikation sein, wo es gilt technische Vorraussetzungen und Nutzeranforderungen zum reibungslosen Funktionieren der App zu klären.

Entscheidet man sich zu einer Evaluation im Labor ist darauf zu achten, dass die Nutzer auch der Zielgruppe der Anwendung entsprechen und das Labor-Setting so realitätsnah wie möglich zu gestalten ist.

Grundsätzlich können beide Evaluierungs-Techniken zum gewünschten Erfolg führen. Maßgeblich ist nicht die verwendete Technik, sondern ein ausgereiftes und durchdachtes Testsetting.

Literaturverzeichnis:

1. Johnson, P. (1998) Usability and Mobility; Interactions on the move. Proc. Mobile HCI´98, GIST Technical Report G-98-1
2. Kjeldskov, J., Skov, M. (2003) A Review of Mobile HCI Research methods. In Proceedings of the 5th International Mobile HCI Conference, Udine, Italy, Sringer-Verlag
3. Hassenzahl, M., (2002)The effect of perceived hedonic quality on product appealingness. International Journal of Human-Computer Interaction
4. ISO/IEC 9126 Software product Evaluation Quality Characteristics and Guidelines for the User, International Organization for Standardization, Geneva, Switzerland
5. MacDonald, C., M., Atwood, M., E., (2013) Changing Perspectives on Evaluation in HCI: Past Present and Future, In Proceedings of the CHI ´13 Extended Abstracts on Human Factors in Computing Systems
6. Ericsson, K.A., Simon, H.A. (1980) Verbal Reports as Data. Psychological Review, 1980, 87, 215-251
7. Dumas, J.R., (1993) A Practical Guide to Usability Testing. Ablex Publishing Corporation Norwood
8. Zhang, D., Adipat, B., (2005) Challenges, methodologies, and issues in the usability testing of mobile applications, International Journal of Human-Computer Interaction, vol. 18, no. 3, pp. 293
9. Apple iOS Human Interface Guidelines: URL: https://developer.apple.com/library/iad/documentation/UserExperience/Conceptual/MobileHIG/MobileHIG.pdf (2014)
10. Google User Interface Guidelines: URL: https://developers.google.com/wallet/objects/ui_guidelines/index.html (2014)
11. Nayebi, F.; Desharnais, J.-M.; Abran, A., (2012) The state of the art of mobile application usability evaluation," *Electrical & Computer Engineering (CCECE), 2012 25th IEEE Canadian Conference* doi: 10.1109/CCECE.2012.6334930 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6334930&isnumber=6334811
12. Blumenstein, K., Schmiedl, G., (2011) Usablilty-Testing mobiler Szenarien als Sekundärtask – Vergleich verschiedener Ansätze, Forum Medientechnik FH St Pölten URL:http://mfg.fhstp.ac.at/cms/wp-content/uploads/2011/12/Blumenstein_Schmiedl_Usability_s1.pdf
13. Kaikkonen, A., Kallio, T., Kekäläinen, A., Kankainen, A., Cankar, M., (2005), Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing. Journal of Usability Studies 1, 1 4-16.
14. Kjeldskov J., Skov, M.B., Als, B.S. and Hoegh, R.T.(2004) Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. Proc. Mobile HCI0`4, Springer
15. Nielsen, C.M., Overgaard, M., Bach Pedersen, M., Stage, J., Stenild, S., (2006). It's worth the hassle!: the added value of evaluating the usability of mobile systems in the field. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles* (NordiCHI '06), Anders Mørch, Konrad Morgan, Tone Bratteteig, Gautam Ghosh, and Dag Svanaes (Eds.). ACM, New York, NY, USA, 272-280. DOI=10.1145/1182475.1182504 URL: http://doi.acm.org/10.1145/1182475.1182504
16. Faulkner, L. (2003) Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. Behavoir Research Methods, Instruments and Computers 2003, 35 (3), 379-383
17. Nielsen, J., and Landauer, T. K.: "A mathematical model of the finding of usability problems," *Proceedings of ACM INTERCHI'93 Conference* (Amsterdam, The Netherlands, 24-29 April 1993), pp. 206-213. URL: www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/
18. Kjeldskov, J., Skov, M.B., (2014) Was it worth the hassle?: ten years of mobile HCI research discussions on lab and field evaluations. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services* (MobileHCI '14). ACM, New York, NY, USA, 43-52. DOI=10.1145/2628363.2628398 URL: http://doi.acm.org/10.1145/2628363.2628398