J. Ziegler & A. Schmidt (Hrsg.): Mensch & Computer 2010
München: Oldenbourg Verlag, 2010, S. 281-290

Usability-Evaluation interaktiver Geräte: Online vs. Labor?

Knut Polkehn[1], Hartmut Wandke[1], Marcus Dahm[2]

[1] Institut für Psychologie, Humboldt-Universität zu Berlin

[2] FB Medien, Fachhochschule Düsseldorf

Zusammenfassung

Kommen aufwändige Usability-Tests von interaktiven Geräten im Labor zu denselben Ergebnissen wie Online-Tests mit browserbasierten Simulationen? In einem Vergleich der beiden methodischen Ansätze kann gezeigt werden, dass die Ergebnisse unterschiedlich ausfallen: Online wird die Effektivität über- und der Aufwand unterschätzt, während das Benutzererleben durchaus vergleichbar ausfällt. Für komparative Usability-Evaluationen sind beide Ansätze gleichermaßen geeignet. Defizite von Simulationen durch das Fehlen von taktilen und haptischen Affordances können durch minimale Instruktionen ausgeglichen werden.

1 Problemstellung und Stand der Forschung

Die vorliegende Studie geht der Frage nach, unter welchen Bedingungen es möglich ist, auch Hardwareaspekte der Mensch-Technik-Interaktion in Online-Experimenten auf Usability zu prüfen. Im Rahmen von Prototyping werden interaktive Geräte oft durch Software (mit dem Vorteil der Erfassung von Logfile-Potokollen) simuliert. Solche Simulationen können dann auch einem Online Usability-Tests unterzogen werden, wie sie z. B. bei Websites erfolgreich unter der Bezeichnung Remote Usability Testing (Lorenzen-Schmidt und Nufer, 2008) praktiziert werden.

In einer ersten Studie konnten Dahm et al. (2004) den Nachweis führen, dass die Simulation von Mobiltelefonen sehr gut in browserbasierten Online-Experimenten möglich ist und dass auf diese Weise sehr schnell aussagkräftige Daten zur Usability gewonnen werden konnten. Allerdings waren sämtliche Interaktionen mit den simulierten Telefonen diskret (Tastendrücke per Maus). Schwierig ist die Simulation von Geräten mit analogen Bedienelementen, wie sie von Dahm et al. 2007 durchgeführt wurde. Gegenstand der Untersuchung waren MP3-Player, bei denen z. T. gleitende und rotierende Bewegungen auf einer Touch-Oberfläche erforderlich waren. Die Autoren bildeten diese Interaktion mit einer rotierenden Mausbewegung nach, wobei die linke Maustaste gedrückt zu halten war und der Cursor als Hand mit

ausgestrecktem Zeigefinger zu sehen war. Im Gegensatz zum realen Finger gab der Cursor jedoch keinerlei Rückmeldung, wenn sich die tastbaren Oberflächeneigenschaften des Gerätes änderten. Damit fehlten in der Simulation die taktilen Affordances (Norman, 1999), die auf die erforderliche Interaktion hinweisen konnten.

Es überraschte daher nicht, dass der iPod-Player von Apple, dem der Ruf vorauseilte, auf besonders intuitive Weise bedienbar zu sein, eher schlecht im Vergleich von drei Playern abgeschnitten hat. Aber vielleicht war dies auch nur ein Artefakt, zurückzuführen auf die Online-Simulation, bei der man den Player nicht in der Hand halten und mit den Fingern die Oberfläche berühren konnte?

2 Untersuchung

2.1 Fragestellungen und Hypothesen

1. Wenn browserbasierte Software-Simulationen online einsetzt und mit den Original-Geräten in einem Labor verglichen werden, kommen dann die beiden Methoden zu vergleichbaren Ergebnissen? Üblicherweise wird Labortests wegen der besseren Bedingungskontrolle eine höhere Reliabilität zugeschrieben. Andererseits bildet das Online-Experiment besser die Kontextfaktoren der Nutzung ab.
2. Hardware bietet oft Affordances für die Ausführung von Interaktionen, insbesondere haptischer und taktiler Natur. Kann man ihr Fehlen in der Simulation dadurch kompensieren, in dem man minimale symbolische Instruktionen einbaut? Diese Frage wird am Beispiel des IClickWheel des iPod untersucht.

Ausgehend von den beiden Fragestellungen wurden Forschungshypothesen entwickelt:

1. Die Usability von Original-Geräten und Online-Simulationen unterscheidet sich, was die absolute Ausprägung von gängigen Usability-Kriterien (Effektivität, Effizienz und Zufriedenstellung) betrifft. Wir nehmen an, dass die Richtung der Unterschiede geräte- bzw. simulationsspezifisch ist.
2. Auch wenn wir annehmen, dass die absolute Ausprägung von Usability-Maßen bei realen Geräten und Online-Simulationen verschieden sein wird, so erwarten wir doch, dass die *Relationen* zwischen den Usability-Maßen verschiedener Systeme identisch sein sollten. Es sollte also keine Interaktion zwischen den zu evaluierenden Systemen und der Evaluationsmethode geben: so wie ein Laptop-Bildschirm größer ist als das Display eines Handys und zwar unabhängig davon, ob die Größe beim Anblick in Zoll geschätzt oder in Millimeter gemessen wird.
3. Es wird angenommen, dass eine zusätzliche visuell-anschauliche Instruktion (Anreicherung der Simulation mit visuellen Affordances) beim iPod zu einer Verbesserung der Interaktion führt, weniger Fehler und Abbrüche auftreten, die Zeit für die Aufgabenlösung kürzer ist und das Gerät besser bewertet wird.

2.2 Untersuchungsdesign

Es wurde ein erweitertes 2 x 2-Untersuchungsdesign verwendet: Die Usability von zwei MP3-Playern wurde unter zwei Bedingungen – im Labor und online – untersucht, so dass der Bezug zur vorhergehenden Studie von Dahm et al. (2007) erhalten blieb. Beim iPod wurde online eine zusätzliche Version mit animierter Hilfe (sich drehender Pfeil) eingeführt, deren Start und Ende in der Abbildung 1 zu sehen ist.

Abbildung 1:links Samsung Touch & Click / rechts iPod ClickWheel (animierte Hilfe) a) klicken und halten b)drehen

Diese Hilfe wurde nur bei den ersten zwei mittels IClickWheel zu lösenden Aufgaben präsentiert. Tabelle1 gibt einen Überblick über das Untersuchungsdesign mit unabhängigen Parallelstichproben: Jede Vp testete nur einen Player, entweder im Labor oder online.

Labor	online simuliert
Originalgerät iPod	iPod ohne Hilfe
	iPod mit Hilfe
Originalgerät Samsung	Vereinfachter Samsung ohne Hilfe

Tabelle 1: 5 parallele Stichproben (2 im Labor / 3 online)

2.3 Aufgaben

Es wurden Aufgaben ausgewählt, bei denen sowohl die in der Online-Simulation schwer zu erkennenden analogen Bedienoperationen (Drehbewegung auf dem IClickWheel) erforderlich waren, als auch die am realen Gerät schwer zu erkennenden Doppelfunktionen eines Bedienelements (leichtes Antippen vs. kräftiges Drücken). Es handelte sich um genau dieselben Aufgaben, die auch von Dahm et al. (2007) verwendet wurden.

1. Player einschalten
2. Tastentöne ausschalten
3. Musiktitel auswählen
4. Lautstärke regulieren
5. Vorspulen
6. Zeit einstellen
7. Klangeinstellung auf Klassik umstellen
8. Player ausschalten

Die oben beschriebene animierte Hilfe für den iPod wurde in der Aufgabe 2 „Tastentöne ausstellen“ eingesetzt, welche als erste das Nutzen des IClickWheels erforderte. In der Aufgabe 3 „Einen Titel auswählen“ erschien die Hilfe zusätzlich im Untermenü Musik, wo die

Scroll-Funktion nötig war, um durch die Wahl der Kategorien (Titel, Alben etc.) zum gewünschten Titel zu gelangen.

2.4 Abhängige Variablen

Folgende Usability-Indikatoren wurden als abhängige Variablen verwendet:

Usability-Indikator	Beschreibung des Indikators
Effektivität	*Anzahl der gelösten Aufgaben* über alle Versuchspersonen dargestellt als Prozentsatz, bezogen auf die Gesamtanzahl der zu lösenden Aufgaben
Effizienz: Bearbeitungszeit	Zeit von Aufgabendarbietung bis Zielerreichung je Aufgabe
Effizienz: relative Zahl von Schritten in Menüs	Anzahl aller Menü-Bewegungen, inklusive der Abweichungen vom Optimalweg, auch Umwege, Abbrüche und Neuanfänge, sowie das „Überschießen“ von Menüzielen bei analogen Bewegungen und ihre Korrektur. Zur besseren Vergleichbarkeit der Aufgaben wurde dann das *Verhältnis* aus der bei der jeweiligen Aufgabe *registrierten Anzahl* der Menü-Bewegungen und der *minimal notwendigen Anzahl* gebildet.
Effizienz: relative Anzahl von Hardware-Aktionen	Es wurde je Aufgabe gezählt, wie viele Mausklicks, Tastenbetätigungen und IClickWheel-Aktionen durchgeführt wurden. Bei den Simulationen wurden auch Klicks auf andere Elemente und Flächen dazugezählt, die keine Bedienelemente waren. Zur besseren Vergleichbarkeit der Aufgaben wurde dann wie beim Menüaufwand das Verhältnis aus der bei der jeweiligen Aufgabe gemessenen und der beim Optimalweg notwendigen Anzahl an Hardware-Interaktionen gebildet, analog der Vorgehensweise von Dahm et al. (2007).
User Experience: Zufriedenstellung	Durch eine einfache Ratingskala, bei der die Extreme durch ein trauriges bzw. fröhliches Symbol visualisiert waren, wurde *unmittelbar nach jeder Aufgabe* erfasst, wie zufriedenstellend die Versuchspersonen die gerade abgelaufene Interaktion erlebt hatten.
User Experience: AttrakDiff2	Am Ende des Versuchs wurde ein *Gesamteindruck* (mit dem Attrakdiff2 von Hassenzahl et al. 2003) der Probanden zu den jeweils untersuchten realen oder simulierten Playern erhoben.

Tabelle 2: Usability-Indikatoren als abhängige Variable

Alle Effizienzmaße wurden nur für Aufgaben bestimmt, die auch gelöst wurden. Für die Beantwortung der IClickWheel-spezifischen Effizienz-Hypothese beim iPod wurden die Effizienz-Variablen zusätzlich zur Erfassung für die Gesamtaufgabe für die Zeit bis zum ersten Bedienen des IClickWheel in der Aufgabe separat erfasst.

2.5 Stichprobe

Als Zielpopulation wurden Personen beider Geschlechter unter 40 Jahre (typisch für mobile MP3-Hörer) und ohne Erfahrung in der Bedienung des jeweils in der Teilstichprobe untersuchten MP3-Players festgelegt, da ja gerade das Entdecken von Interaktionsmöglichkeiten interessierte. Für die Labor-Untersuchung wurden die Probanden aus einer Probandendaten-

bank rekrutiert, in der Studenten aller Jahrgangsstufen und uni-externe Personen verzeichnet sind. Von 67 Personen im Labor testeten 35 Personen den iPod und 32 Personen den Samsung-Player. Für die Rekrutierung von Teilnehmern für die Online- Untersuchung wurde der Untersuchungslink auf einschlägigen Websites veröffentlicht bzw. über unterschiedliche Mailinglisten von Universitäten verschickt. Nach Aussortieren von Teilnehmern mit technisch fehlerhaften Daten und von doppelten Teilnehmern verblieben online 908 Probanden. Weitere Aussortierungen erfolgten aufgrund des Alters (Personen > 40 Jahre) und der Kenntnis des untersuchten Players, so dass 544 Teilnehmer übrig blieben.

Während der Aufgabenbearbeitung brachen weitere Teilnehmer die Untersuchung ab. Der Typ des simulierten Players scheint einen Einfluss auf die *Abbruchrate* gehabt zu haben: Beim Samsung war sie mit 19 % signifikant kleiner als beim iPod (34% ohne Hilfe und 32 % mit Hilfe) (χ^2=13.918, p<.05, φ = -.16). Die beiden iPod-Online-Bedingungen unterscheiden sich aber in den Abbruchraten nicht signifikant (χ^2=0.137, p>0.05, φ = 02).

2.6 Durchführung

Die Untersuchung fand im Sommer 2008 statt. Für die Laboruntersuchung wurden die Interaktionen der Probanden mit dem jeweiligen MP3-Player per Video aufgezeichnet. Die Online-Interaktion mit den simulierten Playern wurde in Logfiles aufgezeichnet.

2.7 Ergebnisse

Über alle Teilnehmer einer Stichprobe hinweg wurden die Anzahl der gelösten Aufgaben aufsummiert und an der Gesamtzahl relativiert („iPod-Labor“: 97,4 %; „Samsung-Labor“ 98,4%; „iPod-online ohne Hilfe“: 90,9%; „iPod-online mit Hilfe“: 85,7%; „Samsung-online“: 95,5%).

Art des Testes	Labor				Online					
Player	iPod		Samsung		iPod o. H.		iPod m. H.		Samsung	
	MW	*Std.*	*MW*	*Std.*	*MW*	*Std.*	*MW*	*Std.*	*MW*	*Std.*
Bearbeitungszeit (sec)	502	*230*	298	*116*	300	115	276	102	179	56
rel. Menüschritte	6,0	*3,6*	2,0	*0,5*	5,6	4,5	5,6	4,3	1,5	0,3
rel. HW-Aktionen	7,5	*4,2*	4,9	*2,2*	6	4,4	4,9	3,3	3,6	0,9
Zufriedenstellung	14,7	*11,7*	27,1	*12,0*	7,2	13,0	20,0	16,7	24,8	13,3
AttrakDiff-ATT	33,6	*8,8*	33,4	*7,2*	32,2	7,4	34,3	6,1	31,0	5,6
AttrakDiff-HQI	32,7	*6,2*	30,8	*5,8*	31,5	5,4	31,4	5,8	29,0	4,1
AttrakDiff-HQS	34,5	*5,4*	23,8	*7,7*	32,8	5,0	31,0	6,5	23,4	6,7
AttrakDiff-PQ	27,5	*9,2*	32,0	*8,2*	28,4	8,2	32,3	6,7	33,7	6,5
% Clickwheel entdeckt	82		-		57,5		82		-	
MW: Mittelwert Std.:Standardabweichung leer: nicht berechenbar -:nicht erhoben										

Tabelle 3: Ergebnisse (Verteilungsparameter)

Tabelle3 zeigt die Ergebnisse für die aufgrund der Bedingungsvariation (UV: Testart und Player) resultierenden fünf Stichproben. Aufgrund der Kürze dieses Artikels werden in Tabelle3 nur Mittelwerte (MW) und Standardabweichungen (Std) berichtet. Auch die Ergebnisse der Hypothesen-prüfenden Tests werden im Folgenden nur zusammenfassend dargestellt.

Eine ausführlichere Darstellung der Ergebnisse kann unter http://www2.hu-berlin.de/psychologie/ingpsycscw/muc2010/MuC2010ergebnisse.pdf abgerufen werden.

Effektivität

Die Bedingungen „Labor“ und „Online“ wurden hinsichtlich der Anzahl der gelösten Aufgaben über alle Versuchspersonen hinweg, bezogen auf die Gesamtanzahl der zu lösenden Aufgaben, verglichen. Im Labor (iPod + Samsung) wurden 98,5% der Aufgaben gelöst, online (iPod + Samsung ohne Hilfe) 94,7%. Dieser Unterschied ist signifikant (χ^2 *(1)=13.397, p<.001,* φ*=-.07*).

Effizienz

Eine 2x2 MANOVA ergab signifikante Effekte sowohl hinsichtlich der UV „Testart“ (Online vs. Labor), als auch der UV „Player“ (iPod vs. Samsung). Es wurde jedoch keine signifikante Interaktion Testart * Player gefunden. Für die UV „Testart“ ergaben Vergleiche hinsichtlich der abhängigen Variablen *Bearbeitungszeit* (F(1,132)=46.25, p<0.001, ε^2=0.26) und *HW-Aktionen* (F(1,132)=7.66, p=0.006, ε^2=0.055) signifikante Unterschiede. Für die UV „Player“ galt das für die *Bearbeitungszeit* (F(1,132)=40.93, p<0.001, ε^2=0.24), die *Menüschritte* (F(1,132)=48.9, p<0.001, ε^2=0.27) und für die *HW-Aktionen* (F(1,132)=13.84, p<0.001, ε^2=0.095)

Folgende Einzelvergleiche erbrachten Mann-Whitney-U-Test ein signifikantes Ergebnis: hinsichtlich der abhängigen Variablen *Bearbeitungszeit:* „iPod Labor > iPod online ohne Hilfe“ (U=174, p=0.001), „iPod Labor > iPod online mit Hilfe“ (U=135, p=0.001) sowie „Samsung Labor > Samsung online“ (U=106, p=0.001). Hinsichtlich der *Menüschritte* „Samsung Labor > Samsung online“ (U=113, p=0.001) und hinsichtlich der *HW-Aktionen* „iPod Labor > iPod online mit Hilfe“ (U=220, p=0.012) bzw. „Samsung Labor > Samsung online“ (U=224, p=0.015).

User Experience

Die MANOVA ergab hinsichtlich der UV „Testart“ für keine abhängige UX-Variable signifikante Unterschiede. Für die UV „Player“ konnten für die abhängige Variable *Zufriedenstellung* (F(1,132)=22.06, p<0.001, ε^2=0.14), für AttrakDiff-HQI (F(1,132)=4.57, p=0.034, ε^2=0.033), für AttrakDiff-HQS (F(1,132)=72.08, p=0.001, ε^2=0.353), AttrakDiff-PQ (F(1,132)=7.84, p=0.006, ε^2=0.056) signifikante Unterschiede gezeigt werden.

Folgende Einzelvergleiche erbrachten im Mann-Whitney-U-Test ein signifikantes Ergebnis: hinsichtlich der abhängigen Variablen *Zufriedenstellung* „IPod Labor > IPod online ohne Hilfe“ (U=250, p=0.031 n.s nach α-Fehler-Adjustierung) sowie „IPod online ohne Hilfe < IPod online mit Hilfe“ (U=207.5, p=0.004). Hinsichtlich *AttrakDiff-HQS* gilt das für den Vergleich „IPod Labor > IPod online mit Hilfe“ (U=242, p=0.034 n.s nach α-Fehler-Adjustierung). Bezüglich der abhängigen Variable *AttrakDiff-PQ* fanden sich tendenziell signifikante Unterschiede für die Vergleiche „IPod Labor < IPod online mit Hilfe“ (U=262, p=0.076) sowie für „IPod online ohne Hilfe < IPod online mit Hilfe“ (U=280.5, p=0.1).

Hypothese 1: Vergleich Labor- und Onlinestudie

Die erste Hypothese postulierte Unterschiede zwischen den Usability-Maßen im Vergleich der Labor- und Online-Studie.
Wir betrachten zunächst die Effektivität (Anteil der gelösten Aufgaben). Tabelle zeigt die prozentualen Anteile der gelösten Aufgaben für alle fünf Stichproben. Obwohl die Lösungshäufigkeiten an sich sehr hoch sind, unterscheiden sich die beiden Bedingungen: im Labor wurden signifikant mehr Aufgaben gelöst als online, unabhängig davon, ob in die Auswertung die iPod-online-Version mit oder ohne Hilfe einbezogen wurde.
In der Online-Studie wird also im Vergleich zu Labor-Tests die Effektivität der Interaktion mit den zu beurteilenden Geräten leicht unterschätzt.
Bei der Analyse der Effizienzmaße finden sich im Vergleich zwischen Labor und Online-Studie lediglich Effekte hinsichtlich der Bearbeitungszeit (Labor>Online) und der Hardware-Aktionen (Labor>Online). Wie hier aus Platzgründen nicht berichtete Analysen zeigen, scheinen insbesondere für letzteres die auch im Labor erst zu entdeckenden, nicht bekannten Interaktionsmöglichkeiten (IClickwheel beim iPod sowie nicht simulierte Doppeltastenbelegung beim Samsung) verantwortlich zu sein.
Was die Zufriedenstellung mit der Aufgabenlösung betrifft, die unmittelbar nach jeder Aufgabe mit Hilfe einer einfachen bipolaren Slider-Skala erhoben wurde, so zeigt sich, dass die Mittelwerte im positiven Bereich (Skala von -50 bis +50) angesiedelt sind. Hier kann kein signifikanter Effekt der UV „Testart“ gefunden werden. Es zeigten sich playerspezifische Effekte: „iPod Labor > iPod online ohne Hilfe“ sowie „iPod online ohne Hilfe < iPod online mit Hilfe“.

Damit können wir auf die erste und zentrale Hypothese weitere Antworten geben. Offensichtlich ist es so, dass die Aufwandsmaße (Interaktion mit der Hardware sowie Zeit für die Aufgabenbearbeitung) bei der Verwendung originaler Geräte höhere Werte annehmen als bei den simulierten Geräten, während in den Menüschritten sowie bei der Zufriedenstellung mit der Aufgabenbearbeitung die Maße zwischen Laboruntersuchungen und Online-Simulationen keine Unterschiede aufweisen.
Die bisher dargestellten Effektivitäts- und Aufwandsmaße kennzeichnen, ebenso wie das unmittelbare Erleben in der Interaktion, die standardmäßig erhobenen Aspekte von Usability. Aber gerade MP3-Player sind ja dazu gedacht, über den reinen Abspielvorgang hinaus Unterhaltung zu schaffen, zu einem spielerischen Umgang anzuregen und durch ihr Design eine besonders ausgeprägte Form von User Experience zu ermöglichen. Hier zeigen die AttrakDiff-2 Daten, dass das besondere Image des iPod sich auch in den Attraktivitätsurteilen zeigt, obwohl die Interaktionsparameter eher auf Usability-Probleme bei diesem Player hinweisen. Auf der Skala „pragmatische Qualität“ schneidet der Samsung-Player (in Übereinstimmung mit den bisherig dargestellten Verhaltensdaten) besser ab als der iPod, aber bei der Skala „Stimulation“ ist es genau umgekehrt. Für unsere Fragestellung ist jedoch wichtiger und entscheidend, dass sich die Bewertungen für die realen Geräte und die Online-Simulationen hinsichtlich aller subjektiven Maße nicht unterscheiden.

Hypothese 2: Relationen zwischen den Usability-Maßen verschiedener Systeme

Diese Hypothese ist wichtig für die komparative Evaluation verschiedener Systeme. Erwartet wurde eine konsistente ordinale Relation: System A ist besser als B (real und simuliert).

Tabelle 4 zeigt, dass für traditionelle verhaltensbezogene Usability-Maße (Zeit und relative Menü-Schritte) im Labor und online analoge Unterschiede zwischen den Playern zu finden sind. Für die von der Simulierbarkeit abhängige AV „HW-Aktionen" trifft das nicht zu. Bei den subjektiven Daten lassen sich im Labor und online vergleichbare Ergebnisse für die AVs „Zufriedenstellung", „ATT", „HQS" und „PQ"finden.

Effizienz	Online	Vergleich	Labor
Bearbeitungszeit	iPod o.H. > Samsung	≈	iPod o.H. > Samsung
rel. Menüschritte	iPod o.H. > Samsung	≈	iPod o.H. > Samsung
rel. HW-Aktionen	iPod o.H. = Samsung	≠	iPod o.H. > Samsung
User Experience			
Zufriedenstellung	iPod o.H. < Samsung	≈	iPod o.H. < Samsung
AttraDiff-ATT	iPod o.H. = Samsung	≈	iPod o.H. = Samsung
AttraDiff-HQI	iPod o.H. > Samsung	≠	iPod o.H. = Samsung
AttraDiff-HQS	iPod o.H. > Samsung	≈	iPod o.H. > Samsung
AttraDiff-PQ	iPod o.H. < Samsung	≈	iPod o.H. < Samsung
<	*signifikanter Unterschied zwischen den Systemen*		
=	*kein signifikanter Unterschied zwischen den Systemen*		
≈	*konsistente ordinale Relation zwischen Online und Labor*		
≠	*inkonsistente ordinale Relation zwischen Labor und Online*		
Die zugrundeliegenden Signifikanztests sind abrufbar unter http://www2.hu-berlin.de/psychologie/ingpsycscw/muc2010/MuC2010ergebnisse.pdf			

Tabelle 4: komperative Player-Evaluation: Vergleich der Bedingungen online und Labor

Hypothese 3: Vergleich iPod-Simulation ohne und mit IClickWheel-Hilfe

Mit dieser Hypothese wurde angenommen, dass eine zusätzliche visuell-anschauliche Instruktion (Anreichung der Simulation mit visuellen Affordances) zu einer Verbesserung der Interaktion führt, weniger Fehler und Abbrüche auftreten, die Zeit für die Aufgabenlösung kürzer ist und das Gerät besser bewertet wird.
Wurde die Funktionsweise des IClickWheels durch Hilfestellung besser entdeckt?

Um diese Frage zu beantworten, betrachteten wir die Aufgabe 2, bei der das erste Mal das IClickWheel einzusetzen war (Tabelle4 unten).

Hypothesenkonform wurde im Labor die Funktionsweise des IClickWheel am realen Gerät häufiger entdeckt (82%) als in der Online-Simulation (57,5 %). Allerdings hilft die Animation (drehender Pfeil) bei der Online-Simulation erheblich. Die Rate der Vpn, die die Funktionsweise erkennen, steigt auf 82 % und hat damit denselben Wert erreicht, wie er in der Laboruntersuchung auftrat. Der $\chi 2$-Test zeigt einen signifikanten Einfluss der Bedingung auf das Entdecken des IClickWheel ($\chi 2(2)=17.700$; $p<.05$, Cramer-V=.27). Drei Einzelvergleiche ergaben, dass im Labor das IClickWheel öfter entdeckt wurde als bei der Online-Version ohne Hilfe ($\chi 2(1)=7.189$, $p<.0167$, einseitig, $\varphi = -.23$). Auch kann gezeigt werden, dass mit animierter Mini-Instruktion das IClickwheel häufiger entdeckt wurde als in der Version ohne Hilfe ($\chi 2(1)=14.729$, $p<.0167$, einseitig; $\varphi = .27$). Zwischen der Labor- und der Online-Version mit Hilfe ist der Unterschied nicht signifikant.

Damit ist auch die dritte Hypothese klar bestätigt worden: In der Tat, fehlende taktile Rückmeldungen führen in der Simulation zu Erkennungsproblemen, aber man kann diese Defizite der Online-Simulation durch einfache Mittel beheben.
Die Daten zeigen aber auch, dass die Integration zusätzlicher Onlinehilfen zu Veränderungen in Verhaltensdaten (HW-Aktionen), als auch in subjektiven Bewertungen (HQS, PQ) führen kann. Der direkte Vergleich der iPod-Online-Varianten mit und ohne Hilfe, weist bei der Betrachtung über alle Aufgaben hinweg signifikante Unterschiede in der erwarteten Richtung lediglich für die Zufriedenstellung und tendenziell PQ aus. Hier ist zu berücksichtigen, dass die Hilfe nur in zwei Aufgaben dargeboten wurde.

3 Fazit

Insgesamt stützen die Ergebnisse das Vorgehen von Dahm et al. (2007). Es ist durchaus möglich, auch mobile Geräte mit analogen Eingabeelementen per Mausaktionen in einer Online-Simulation über das WWW zu untersuchen und daraus valide und bedeutsame Schlussfolgerungen zu ziehen. Allerdings müssen folgende Besonderheiten im Vorgehen und bei der Interpretation der Daten beachtet werden.

1. *Effizienz „Online vs. Labor“:* Es ist davon auszugehen, dass es einen starken Selbstselektionseffekt unter den Online-Teilnehmern gibt. Durch die relativ hohen Abbrecherquoten bleiben am Ende nur besonders motivierte und befähigte Teilnehmer in der Stichprobe. Dies hat zur Folge, dass die Online-Teilnehmer eine höhere Effizienz aufweisen als die Teilnehmer im Labor, die sich natürlich nicht trauen, einfach aufzustehen und das Labor zu verlassen. Bei Online-Simulationen ist daher immer damit zu rechnen, dass die dort ermittelte Effizienz überschätzt wird.
2. *Effektivität „Online vs. Labor“:* Hier sieht es umgekehrt aus. Online-Teilnehmer lösen weniger Aufgaben als Teilnehmer im Labor. Die Abweichungen in der Lösungsrate sind zwar nicht gravierend, aber bei Anwendungssystemen, bei denen eine hohe Sicherheit und Lösungsrate gefordert ist, sind die Effektivitätswerte in Online-Studien tendenziell als zu gering zu betrachten. Auch hierfür ist als Grund anzunehmen, dass die Teilnehmer eher abbrechen (zwar nicht den kompletten Versuch, aber doch die aktuell bearbeitete Aufgabe).
3. *ökologische Validität:* Im Fall von MP3-Playern kann man allerdings auch eine andere Sicht einnehmen und davon ausgehen, dass die Ergebnisse einer Online-Studie ein höheres Maß an Validität besitzen. In einem privaten Kontext außerhalb des Labors werden Benutzer von realen Geräten ebenfalls weniger Aufwand treiben und bei Schwierigkeiten auch eher die Aufgabenbearbeitung abbrechen.
4. *Komperative Evaluation:* Wenn es um vergleichende Evaluationen (siehe Hypothese 2) geht, können Online-Simulationen sehr gut eingesetzt werden. Unterschiede zwischen den Playern zeigen sich überwiegend in gleicher Richtung und auch in etwa gleichem Abstand in Labor- und Online-Test. Dies gilt für die „klassischen“ Usability-Kriterien, wie auch (unter bestimmten Voraussetzungen – siehe Hypothese 2) für die Skalen des AttrakDiff.

5. *Affordances in Simulierter Hardware:* Es ist durchaus möglich, Unzulänglichkeiten der browserbasierten Simulation (insbesondere das Fehlen von taktilen und haptischen Sinneseindrücken und den damit verbundenen Affordances) durch symbolische Instruktionen zu kompensieren. Dabei ist zu berücksichtigen, dass die Hilfe selbst Einfluss auf die subjektiven Bewertungen des Gerätes haben kann.

Diese insgesamt ermutigenden Schlussfolgerungen basieren jedoch nur auf dem Vergleich von zwei realen Playern und drei Online-Simulationen. Es ist in weiteren Studien zu prüfen, ob sie auch für andere interaktive Systeme und für weitere Usability-Kriterien Gültigkeit besitzen. Zunächst sind weiterführende Untersuchungen geplant, in denen die Konfundierung zwischen „Simulation“ und „Online-Studie“ aufgehoben wird. Zu diesem Zweck sollen die browserbasierten Simulationen unter Laborbedingen getestet werden.

4 Schlussteil

Literaturverzeichnis

Dahm, M., Günter, M., Hassing, J., & Bach, M. (2007). Interaktionsformen und Usability von MP3-Playern. In Gross, T. (Hrsg.): *Konferenz für interaktive und kooperative Medien*. München: Oldenbourg Verlag, S. 89-98.

Dahm, M., Felken Ch., Klein-Bösing, M., Rompel, G., & Stroick, R. (2004). Handyergo: *Breite Untersuchung über die Gebrauchstauglichkeit von Handys*, In Keil-Slawik, Selke, Szwillus (Hrsg): *Mensch und Computer 2004 – Allgegenwärtige Interaktion*. München: Oldenbourg Verlag, S.75-84.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In Ziegler, J. & Szwillus, G. (Hrsg.): *Mensch & Computer 2003. Interaktion in Bewegung*. Stuttgart: B.G. Teubner*, S.* 187-196.

Lorenzen-Schmidt, O., Nufer, S. (2008). From a Distance: Usability Testing aus der Ferne. *i-com,* 7(1), München: Oldenbourg Verlag, S. 44-46.

Norman, D. A. (1999). Affordances, Conventions and Design. *Interaction,* 6(3), ACM Press, S.38-43.

Danksagung

Die Autoren danken Frau Ch. Karsten für die Durchführung der Studie und die Auswertung der Daten und Herrn M. Günter für die technische Umsetzung der animierten iPod-Hilfe, sowie für die Bereitstellung der Infrastruktur für die Online-Studie.

Kontaktinformationen

Knut Polkehn (Humboldt-Universität zu Berlin)

Telefon: +49 (30) 20939357

E-Mail: knut.polkehn@psychologie.hu-berlin.de