Halbrügge, M. & Engelbrecht, K.-P. (2015). *Können Nutzer im Usability-Labor zwischen Interface-Varianten unterscheiden? Zwei Fallbeispiele aus dem Smart Home.* In Gramann, K.; Zander, T. O. & Wienrich, C. (eds.). 11. Berliner Werkstatt Mensch-Maschine-Systeme, pp27-32, Universitätsverlag der Technischen Universität Berlin

Können Nutzer im Usability-Labor zwischen Interface-Varianten unterscheiden? Zwei Fallbeispiele aus dem Smart Home

Marc Halbrügge[1] und Klaus-Peter Engelbrecht[1]

Keywords: Usability, Nutzerzufriedenheit, AttrakDiff, MMQQ, Smart Home

Zusammenfassung

Wenn Benutzeroberflächen auf verschiedenen Geräten dargestellt werden sollen, sind häufig Anpassungen an das jeweilige Gerät nötig. Aus Usability-Sicht stellt sich hier die Frage, wie die dabei entstehenden Varianten sinnvoll evaluiert werden können. Standardfragebögen sind an sich ein mächtiges Werkzeug, kommen hier aber unter Umständen an ihre Grenzen. Anhand zweier Untersuchungen an einer Smart-Home-Applikation beleuchten wir diese Grenzen und schlagen als Alternative ein Forced-Choice-Paradigma vor, das in unserem Fall sehr viel besser zwischen Interface-Varianten diskriminieren konnte.

Einleitung

Immer mehr Bereiche des täglichen Lebens werden immer stärker von Computertechnologie durchdrungen. Gleichzeitig fächern sich die daran beteiligten Geräte in immer neue und verschiedene Arten und Unterarten auf, auf der einen Seite zu immer kleineren wie „intelligenten" Uhren, auf der anderen Seite zu immer größeren, z.B. „intelligenten" Fernsehern. Viele Anwendungen möchte der Nutzer auf allen oder mehreren dieser Geräte nutzen, wobei aus Usability-Sicht eine konsistente, d.h., möglichst einheitliche Gestaltung der Benutzeroberfläche von Vorteil ist. Es ist jedoch offensichtlich, dass je nach Größe und Abstand des zu bedienenden Geräts unterschiedliche Variationen einer Benutzeroberfläche notwendig sind. Aus technischer Sicht kann dieses durch die Adaption einer Basis-Oberfläche gelöst werden (z.B. „Responsive Design", Marcotte, 2011).

Aus Sicht der Gebrauchstauglichkeit stellt sich hierbei die Frage, wie die unterschiedlichen Varianten sinnvoll Nutzertests unterzogen werden können. Insbesondere wenn eine Entscheidung zwischen zwei Design-Varianten getroffen werden soll, eignen sich prinzipiell Nutzerurteile als Entscheidungskriterium. Da aufgrund der sehr ähnlichen Varianten nur kleine Unterschiede in den Urteilen erwartet werden, würde man zunächst Versuchs-Designs mit Messwiederholung vorziehen, um große Fallzahlen zu vermeiden. In diesem Fall stellt sich die Frage, ob direkte quantitative Urteile oder Präferenzurteile geeigneter sind, um eine Entscheidung für ein Design zu treffen.

Wir präsentieren zwei empirische Untersuchungen der Gebrauchstauglichkeit eines Kochassistenten, bei denen verschiedene Varianten der Benutzeroberfläche einmal mit Fragebögen, einmal nach Präferenz verglichen wurden. Die Ergebnisse zeigen verschiedene Problematiken beider Methoden auf, wobei die Vorteile der Präferenzmethode insgesamt überwiegen.

Anwendung: Der Kochassistent im Smart Home

Anwendungen im Smart Home müssen auf vielen unterschiedlichen Geräten dargestellt werden und eignen sie sich daher besonders für unsere Fragestellung. Die Wahl fiel auf einen

[1] Quality and Usability Lab, Telekom Innovation Laboratories, Technische Universität Berlin
Ernst-Reuter-Platz 7, 10587 Berlin, {marc.halbruegge, klaus-peter.engelbrecht}@telekom.de

Kochassistenten, der zum einen auf großen Touchscreens, zum anderen auf tragbaren Geräten, hier Tablets, benutzt wird (siehe Abbildung 1). Der Kochassistent beinhaltet eine schlagwortbasierte Rezeptsuche, einen Zutatenrechner und führt durch die Zubereitung eines Rezepts. Von diesen Funktionen betrachten wir im Rahmen dieser Arbeit insbesondere die Suche und Auswahl von Rezepten sowie Zutaten.

Experiment 1

Die Untersuchung fand im Zuge einer größeren Studie statt (siehe Quade et al., 2014). Von 20 Teilnehmern musste eine Person wegen fehlender Daten ausgeschlossen werden, von den verbleibenden waren 9 Frauen, das mittlere Alter betrug 28 Jahre (SD=9.7). Das benutzte Tablet war hier ein Apple iPad, der benutzte Touchscreen ein 19 Zoll Bildschirm, der fest in einen der Küchenschränke integriert war. Da die Benutzeroberfläche nicht auf das benutzte Tablet und den dort verwendeten Browser Safari optimiert war, ergaben sich mehrere relativ schwerwiegende Usability-Probleme:

1. Suchergebnisse wurden nicht in Listen, sondern als Drop-Down-Menü ohne Vorauswahl angezeigt. Dadurch war es den Nutzern nicht möglich, das Ergebnis einer Suche direkt wahrzunehmen. Jegliches visuelle Feedback, ob eine Suche erfolgreich war oder nicht, fehlte.
2. Die Schaltflächen zur An- oder Abwahl von Suchattributen waren deutlich kleiner als die üblicherweise geforderten 10mm (Park & Han, 2010), dadurch kleiner als eine menschliche Fingerkuppe und dementsprechend schwer zu treffen. Zusammen mit einer allgemeinen Zeitverzögerung von über 300ms bis zur visuellen Rückmeldung eines Klicks war es den Nutzern nur schwer möglich, den aktuellen Zustand des Systems zu erfassen.
3. Da der Küchenassistent nicht für Mehrfinger-Interaktion programmiert war, passierte es sehr schnell, dass Nutzerklicks vom System fälschlicherweise als Gesten zum Einzoomen oder Scrollen interpretiert wurden. Insbesondere nach Zoom-Fehlern waren manche Teilnehmer auf Hilfestellung vom Versuchsleiter angewiesen.

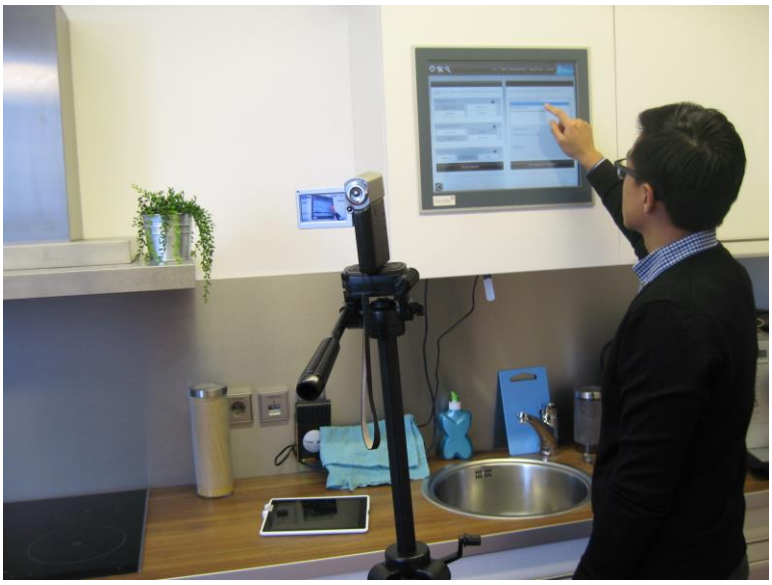

Abb.1: Küche als Teil eines Smart-Home-Systems inkl. Aufbau des Experiments

Eine Hälfte der Teilnehmer begann mit dem Tablet, die andere mit dem Touchscreen. Nach einem Aufgabenblock wurde zum jeweils anderen Gerät gewechselt. Nach beiden Blöcken füllten die Versuchspersonen sowohl das MultiModal Quality Questionnaire (MMQQ,

Wechsung, 2014) als auch den AttrakDiff mini (AttrD, Hassenzahl & Monk, 2010) aus. Die Auswertung erfolgte mit R (R Core Team, 2014), bei den angegebenen Effektstärken handelt es sich um das generalisierte Eta-Quadrat (Olejnik & Algina, 2003).

Wie in Tabelle 1 gezeigt korrelieren die Skalen MMQQ Ease of Use und AttrakDiff pragmatische Qualität (PQ) im Bereich der Skalenreliabilitäten, dasselbe gilt für MMQQ Joy of Use und AttrakDiff hedonische Qualität (HQ). Aus Sicht der klassischen Testtheorie messen beide Skalenpaare also jeweils dasselbe Konstrukt, daher wird für die weitere Analyse der Mittelwert der Skalenpaare verwendet. Die resultierenden Skalen werden im Folgenden mit „Ease" bzw. „Joy" bezeichnet, Cronbachs Alpha beträgt für beide Skalen .93. Weder ergaben sich Zusammenhänge mit dem Alter (Ease: r=-.17, p=.48; Joy: r=-.01, p=.97), noch zeigten sich Unterschiede zwischen den Geschlechtern (Ease: $t_{15.4}$=.3, p=.76; Joy: $t_{16.0}$=1.5, p=.15).

Tab.1: Korrelationen der Skalen zum 1. Messzeitpunkt

	AttrD PQ	AttrD HQ	MMQQ Ease	MMQQ Joy
AttrD PQ	*(.92)*	.72	1.03	.73
AttrD HQ	.63	*(.84)*	.56	.96
MMQQ Ease	.94	.49	*(.91)*	.73
MMQQ Joy	.67	.84	.66	*(.91)*

Anmerkungen: Unkorrigierte Korrelationen unterhalb, doppelt minderungskorrigierte Korrelationen oberhalb der Diagonale. Auf der Diagonale ist Cronbachs Alpha abgetragen.

Nach der Durchführung des zweiten Aufgabenblocks mit dem jeweils anderen Gerät wurden beide Fragebögen wieder vorgelegt, die Korrelationen der erhaltenen Werte finden sich in Tabelle 2. Auffällig ist der starke Zusammenhang zwischen Ease und Joy zum zweiten Messzeitpunkt. Mit der minderungskorrigierten Korrelation von .99 besteht kein Unterschied zwischen den Dimensionen mehr, die Änderung zum ersten Durchgang ist signifikant (Steiger Test nach Bortz, 2005, S. 223; z=-3.5, p<.01).

Tab.2: Korrelationen der kombinierten Skalen zwischen 1. und 2. Messzeitpunkt

	Ease T1	Joy T1	Ease T2	Joy T2
Ease T1	*(.93)*	.69	.81	.61
Joy T1	.64	*(.93)*	1.00	1.02
Ease T2	.75	.92	*(.92)*	.99
Joy T2	.58	.97	.94	*(.97)*

Anmerkungen: Unkorrigierte Korrelationen unterhalb, doppelt minderungskorrigierte Korrelationen oberhalb der Diagonale. Auf der Diagonale ist Cronbachs Alpha abgetragen.

Eine anschließend durchgeführte Varianzanalyse mit den Faktoren Gerät und Reihenfolge (Tablet zuerst oder Touchscreen zuerst) ergab keine signifikanten Ergebnisse. Der aufgrund der Usability-Probleme des Tablets erwartete Effekt des Geräts ist mit η^2=.005 deutlich geringer als der gleichwohl insignifikante Effekt der Gerätereihenfolge mit η^2=.06. Die grafische Inspektion der Mittelwerte zeigt, dass Teilnehmer, die mit dem großen Touchscreen angefangen haben, das Tablet nur wenig schlechter als den Bildschirm bewerten. Teilnehmer, die mit dem Tablet angefangen haben, geben zwar eine etwas bessere Bewertung des großen Screens ab, insgesamt wird die Qualität des Tablet zum 2. Messzeitpunkt aber *besser* als die des Bildschirms eingeschätzt.

Zusammen mit der geänderten Korrelationsstruktur zwischen Joy und Ease liegt die Interpretation nahe, dass die Versuchsteilnehmer zum zweiten Messzeitpunkt nicht in der Lage waren, die Bewertung der zweiten Interaktion unabhängig von der vorangegangenen ersten

durchzuführen. Obwohl das Tablet deutliche Usability-Probleme hatte, haben die Nutzer bei der Frage nach ihrer Zufriedenheit mit dem Küchenassistenten möglicherweise eine Einschätzung entsprechend ihrer *gesamten* Erfahrung mit dem System getroffen.

Experiment 2

Das zweite Experiment fand im Rahmen einer Untersuchung zu Nutzerfehlern statt (Halbrügge et al., 2015). Es nahmen 20 Versuchspersonen teil, davon 15 Frauen. Das mittlere Alter betrug 32 Jahre (SD=11.6). Als Tablet kam hier ein Samsung GalaxyTab 10.1 zum Einsatz, als Touchscreen ein Asus ET27011 PC mit integriertem 27 Zoll Touchscreen. Die Usability-Probleme des Tablets wurden wie folgt adressiert:

1. Zoom- und Markieren-Gesten wurden im HTML-Code abgeschaltet.
2. Die ursprüngliche Listen-Darstellung der Suchergebnisse wurde durch Wechsel des Browsers wiederhergestellt.
3. Eine vereinfachte Variante der Benutzeroberfläche mit weniger, dafür größeren Schaltflächen pro Bildschirmseite wurde erstellt.

Die ersten beiden Punkte gelten dabei auch für die originale Version des UIs. Um die Eignung der vereinfachten UI-Version zu prüfen, wurden im Experiment alle vier Kombinationen aus Gerät und UI in zufälliger Reihenfolge bearbeitet. Da die Fragebögen in Experiment 1 nicht zwischen den Geräten diskriminieren konnten, ihre Anwendung aber relativ viel Zeit gekostet hatte, wurden sie durch ein Forced-Choice-Paradigma ersetzt. Nach dem letzten Durchgang wurden die Versuchspersonen gefragt, welche Kombination von Gerät und UI ihnen am besten, welche am schlechtesten gefallen hat. Danach sollten sie nochmals von den beiden übrig gebliebenen Kombinationen die bessere wählen, wodurch sich eine vollständige Rangreihe der Gerät-UI-Kombinationen pro Person ergibt.

Die erhaltenen Ränge wurden einer Varianzanalyse mit den Faktoren Gerät (Bildschirm, Tablet) und UI (original, vereinfacht) unterzogen. Die Ergebnisse sind in Tabelle 3 dargestellt. Während die Kombination aus großem Bildschirm und Original-UI insgesamt am besten bewertet wurde, wurde dasselbe UI auf dem Tablet am häufigsten abgelehnt. Die Mittelwerte sind in Abbildung 2 dargestellt, dabei bedeutet 1 den besten und 4 den schlechtesten Rang.

Tab.3: Varianzanalyse der Auswahlränge

	MSE	$F_{1,19}$	η^2	p
Gerät	.64	52.64	.45	<.01
UI	.87	2.81	.06	.11
Interaktion	.68	32.35	.35	<.01

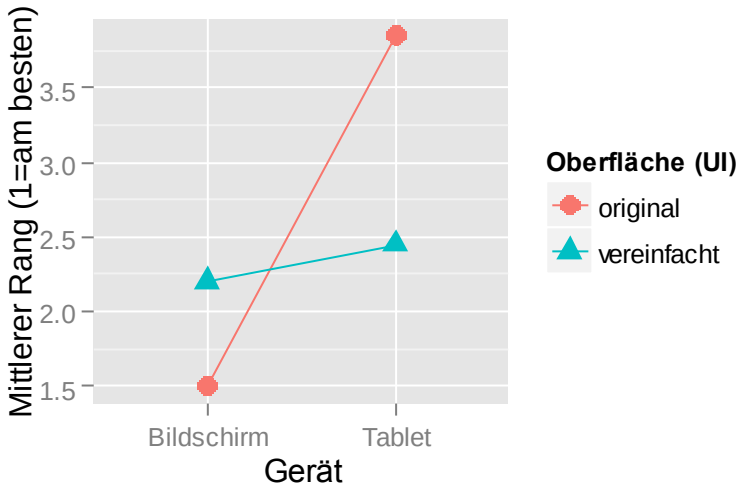

Abb.2: Mittlere Bewertung der vier Kombinationen aus Gerät und UI

Diskussion

Qualitätsurteile hängen einerseits ab von den Eigenschaften der zu beurteilenden Objekte, andererseits von den Erwartungen und der inneren Referenz der urteilenden Personen (Möller, 2010). Unterschiedliche Erwartungen der Versuchsteilnehmer können als Ursache der großen Varianz innerhalb der Gruppen zum ersten Messzeitpunk von Experiment 1 angesehen werden. Grundsätzlich sind daher Messwiederholungs-Designs zu bevorzugen, da die Benutzung ipsativer Werte den Effekt (zeitstabiler) individueller Referenzgrößen nivelliert.

Bei der wiederholten Messung der Qualität eines Kochassistenten mit Fragebögen (Experiment 1) zeigte sich allerdings, dass die Teilnehmer zunehmend unspezifisch antworteten und sich die erwarteten Unterschiede auf Gruppenebene sogar ins Gegenteil verkehrten. Es bleibt unklar, ob es sich hierbei lediglich um eine Vermischung der Bewertungen der beiden Varianten handelt, oder ob sich durch die Interaktion mit einer der Varianten auch die interne Referenzgröße für die Qualitätsbeurteilung der anderen Variante geändert hat.

Der Wechsel zu einem Forced-Choice-Paradigma in Experiment 2 ergab dagegen sehr deutliche Unterschiede zwischen sogar vier Varianten des Kochassistenten, obwohl mehrere der Usability-Probleme des Assistenten im zwischenzeitlichen Entwicklungszyklus behoben worden waren. Die einfache Frage nach der besten und schlechtesten Variante benötigte auch deutlich weniger Zeit als das Ausfüllen der Fragebögen. Ein Nachteil dieser Vorgehensweise ist allerdings die Eindimensionalität der Ergebnisse, während Fragebögen grundsätzlich Auswertungen von Subskalen oder Einzelfragen erlauben. Eine Kombination mit offenen Fragen („Wieso hat Ihnen diese Variante am besten gefallen“) bietet sich daher an. Ein weiterer Nachteil der Präferenz-Methode ist, dass keine numerischen Aussagen über die Qualität der Benutzeroberfläche möglich sind (siehe auch Annett, 2002). Dies kann problematisch sein, wenn alle untersuchten Varianten nicht gebrauchstauglich sind und daher ein neues Design erforderlich ist.

Fazit

Der Vergleich von nur leicht unterschiedlichen Varianten von Benutzeroberflächen erzeugt ein versuchsplanerisches Dilemma. Lässt man die Nutzer nur mit einer Variante interagieren, so benötigt man sehr große Fallzahlen für belastbare Aussagen. Zeigt man den Nutzern mehrere Varianten (Messwiederholung), so fällt es den Versuchsteilnehmern möglicherweise schwer, die Varianten zu unterscheiden, insbesondere bei der mehrfachen Anwendung von

Standardfragebögen. Einen Ausweg aus diesem Dilemma stellt die Benutzung eines Forced-Choice-Paradigmas dar. In unseren Experimenten konnten die Versuchsteilnehmer hierbei sehr gut zwischen UI-Varianten diskriminieren, die sich vorher mit Fragebögen nicht unterscheiden ließen.

Danksagung. Die Arbeit entstand im Rahmen des durch die Deutsche Forschungsgemeinschaft (DFG) geförderten Projektes „Automatische Usability-Evaluierung modellbasierter Interaktionssysteme für Ambient Assisted Living“ (MO 1038/18-1).

Literatur

Annett, J. (2002). Subjective rating scales: science or art? *Ergonomics 45*(14), 966-987.

Bortz, J. (2005). *Statistik: Für Human-und Sozialwissenschaftler*. Berlin: Springer-Verlag.

Halbrügge, M., Quade, M. & Engelbrecht, K.-P. (2015). A Predictive Model of Human Error based on User Interface Development Models and a Cognitive Architecture . InTaatgen, N. A.; van Vugt, M. K.; Borst, J. P. & Mehlhorn, K. (Eds.) *Proceedings of the 13th International Conference on Cognitive Modeling*, 238-243. Groningen, NL: University of Groningen

Hassenzahl, M. & Monk, A. (2010). The inference of perceived usability from beauty. *Human-Computer Interaction, 25*, 235-260

Marcotte, E. (2011). *Responsive web design.* New York: A Book Apart

Möller, S. (2010). Quality Engineering. Heidelberg: Springer

Olejnik, S. & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological methods, 8*, 434-447

Park, Y. S. & Han, S. H. (2010). Touch key design for one-handed thumb interaction with a mobile phone: Effects of touch key size and touch key location. *International journal of industrial ergonomics, 40*, 68-76

Quade, M., Halbrügge, M., Engelbrecht, K.-P., Albayrak, S. & Möller, S. (2014). Predicting Task Execution Times by Deriving Enhanced Cognitive Models from User Interface Development Models. In *Proceedings of the 2014 ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, 139-148

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Wien. http://www.R-project.org/.

Wechsung, I. (2014). *An Evaluation Framework for Multimodal Interaction.* Berlin: Springer