Methodenvergleich von UX-Tests im Kontext mobiler Applikationen

Dustin Rauch

eResult GmbH

Zusammenfassung

Die Studie nimmt einen empirischen Vergleich der Evaluationsmethoden Usability-Test im Labor, synchroner Remote Usability-Test (RUT) und asynchroner RUT im Kontext nativer Applikationen vor. Kennzahlen des Vergleichs bilden die Anzahl identifizierter UX-Probleme, deren Schweregrad und die Frequenz.

1 Einleitung

Aufgrund der physikalischen Eigenschaften mobiler Endgeräte (z. B. ein kleines Display oder eine fehlende Tastatur) sowie dem dynamischen Nutzungskontext stellt die UX-Evaluation nativer Applikationen eine besondere Herausforderung dar (Budui 2008). Eine mangelhafte UX führt bei Nutzern häufig zu einer schnellen Deinstallation der Applikation und gegebenenfalls zu einer schlechten Bewertung in den App-Stores (Budui 2008). Die Herausforderung beim Testen nativer Applikationen besteht darin, dass bekannte Strategien von UX-Tests mobiler Websites nicht vollständig auf das Testen nativer Applikationen übertragen werden können (Krannich 2010). Nach heutigem Stand der Technik ist das Testen nativer Applikationen lediglich mit einem hohen technischen Aufwand verbunden (Rauch 2016). Die vorliegende Studie vergleicht zwei innovative Evaluationsmethoden zum Testen nativer Applikationen mit einem klassischen UX-Test im Labor. Das Forschungsinteresse besteht demzufolge darin, die Abhängigkeit der Ergebnisse von der Evaluationsmethode empirisch zu untersuchen.

2 Studiendesign

Ziel der Studie ist ein empirischer Vergleich der Evaluationsmethoden Usability-Test im Labor, synchroner Remote Usability-Test (RUT) und synchroner RUT im Kontext nativer Applikationen. Die Besonderheit eines synchronen RUT besteht durch die direkte Kommunikation zwischen Probanden und Testleiter bei einer parallelen räumlichen Distanz.

Veröffentlicht durch die Gesellschaft für Informatik e.V. und die German UPA e.V. 2016 in S. Hess & H. Fischer (Hrsg.): Mensch und Computer 2016 – Usability Professionals, 4. - 7. September 2016, Aachen.

http://dx.doi.org/10.18420/muc2016-up-0004

Folglich befindet sich der Proband bei einem synchronen RUT in seiner gewohnten Nutzungsumgebung (dem Feld) und führt den Test am eigenen Endgerät durch. Im Kontrast dazu herrscht bei einem asynchronen Remote Usability-Test eine räumliche Distanz zwischen Testleiter und Proband Tests. Weiterhin findet keine direkte Interaktion zwischen Proband und Testleiter statt, da nicht durch einen Testleiter moderiert wird. Neben der räumlichen Distanz sind Testleiter und Probanden demnach auch zeitlich voneinander unabhängig.

Das Testobjekt der Studie ist die App der global agierenden Parfümerie Douglas GmbH. Grundlage des Vergleichs bildeten aufeinanderfolgende Use Cases. Sie teilten den jeweiligen UX-Test in verschiedene Phasen ein, die als Richtlinien für den Ablauf zu interpretieren sind (Sarodnick & Bau 2011). Hierdurch wird gewährleistet, dass keine Fragestellung ausgelassen wird und alle Probanden die identischen Aufgaben bearbeiten.

Entscheidende Messwerte für den Vergleich bilden die relative Häufigkeit, der Schweregrad sowie die Frequenz der identifizierten UX-Probleme. Die Grundlage für die Analyse bildete das laute Denken der Probanden. Weiterhin wurden die Beobachtungen der durchführenden UX-Experten hinzugenommen.

Folgende Hypothesen wurden für den weiteren Verlauf der Studie postuliert:

1. Durch den UX-Test im Labor, den synchronen RUT und den asynchronen RUT werden identische UX-Probleme identifiziert.
2. Die Anzahl an UX-Problemen des Schweregrads "Critical", "Serious" und "Cosmetic' sind weitestgehend unabhängig von der Evaluationsmethode, sodass kein Unterschied in den Ergebnissen festgestellt werden kann.
3. Durch die Möglichkeit der direkten Kommunikation zwischen Testleiter und Proband werden innerhalb der Evaluationsmethoden „UX-Test im Labor“ und „synchrone RUT“ mehr UX-Probleme aufgedeckt als bei der automatisierten asynchronen RUT.
4. Die Frequenz der identifizierten UX-Probleme und die angewandten Evaluationsmethoden sind weitestgehend unabhängig voneinander, sodass kein Unterschied in den Ergebnissen festgestellt werden kann.
5. Die Frequenz der identifizierten UX-Probleme der Schweregrade "Critical", "Serious" und "Cosmetic'‘ differenziert je nach Evaluationsmethode.
6. Im asynchronen RUT ist ein natürlicheres Verhalten der Teilnehmer zu beobachten als im UX-Test im Labor und synchronen RUT.

3 Empirischer Vergleich der Evaluationsmethoden

3.1 Anzahl der identifizierten UX-Probleme

Abbildung 1 zeigt, dass methodenübergreifend 49 verschiedene UX-Probleme identifiziert wurden. Es wird deutlich, dass durch den UX-Test im Labor 25 UX-Probleme festgestellt werden konnten. Im Gegensatz dazu konnten im asynchronen RUT 21 UX-Probleme

aufgedeckt werden. Den höchsten Wert im Vergleich der Evaluationsmethoden lieferte der synchrone RUT, welcher 29 unterschiedliche UX-Probleme identifizierte. Abbildung 2 zeigt, dass 28 der insgesamt 49 identifizierten UX-Probleme lediglich durch eine einzelne Evaluationsmethode identifiziert werden konnten. Demzufolge wurden elf UX-Probleme lediglich durch den synchronen RUT identifiziert, acht allein durch den asynchronen RUT und neun ausschließlich durch den UX-Test im Labor. Die geringste Schnittmenge im Vergleich bilden der UX-Test im Labor und der asynchrone RUT mit lediglich neun gemeinsam identifizierten UX-Problemen. Mit 14 gemeinsam identifizierten UX-Problemen bildeten der synchrone RUT und der UX-Test im Labor die größte gemeinsame Schnittmenge. Lediglich 7 der insgesamt 49 identifizierten UX-Probleme konnten durch alle Evaluationsmethoden identifiziert werden. Anhand der aufgezeigten Erkenntnisse lässt sich bereits eine Tendenz erkennen, dass die Moderation eines Tests einen starken Einfluss auf die Ergebnisse besitzt. Die Analyse der UX-Probleme anhand ihrer Schweregrade soll weiteren Aufschluss geben.

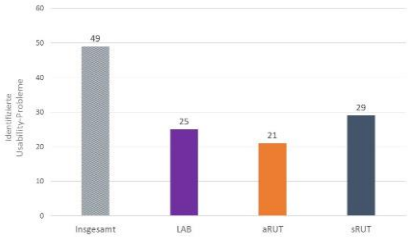

Abbildung 1: Identifizierte UX-Probleme

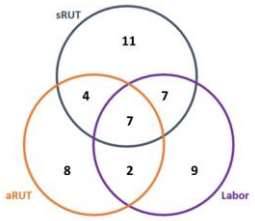

Abbildung 2: Verteilung der UX-Probleme

3.2 Schweregrade aufgefundener UX-Probleme

11 der 49 identifizierten UX-Probleme konnten dem Schweregrad "Critical" zugeordnet werden. Abbildung 3 zeigt, dass der UX-Test im Labor acht der 11 UX-Probleme identifizieren konnte. Einen identischen Wert lieferte der synchrone RUT. Der asynchrone RUT konnte die Ergebnisse der moderierten Verfahren nicht erreichen und identifizierte lediglich drei UX-Probleme des Schweregrades "Critical". Anhand der vorangehenden Ergebnisse lässt sich postulieren, dass sich die Moderation eines UX- Tests positiv auf das Auffinden von UX-Problemen des Schweregrads "Critical" auswirkt (Hypothese 2). Der Einfluss der Moderation lässt sich dahingehend deuten, dass während des asynchronen RUTs ein sehr zögerliches und fehlervermeidendes Verhalten der Probanden zu beobachten war. Das Verhalten der Probanden im asynchronen RUT zielte merklich darauf ab, potentielle Sackgassen des Systems zu vermeiden. Es ließ sich innerhalb der nachträglichen Videobetrachtung beobachten, dass sich die Probanden des asynchronen RUTs in der Verantwortung fühlten, den Test erfolgreich abzuschließen. Folglich waren die Probanden des asynchronen RUTs grundsätzlich weniger dazu geneigt, eine laufende Aufgabe abzubrechen. Durch die geringere Risikobereitschaft der Probanden wurden folglich weniger UX-Probleme des Schweregrads "Critical" identifiziert. Im Kontext der UX-Probleme des Schweregrads "Serious" zeigt sich, dass in der vorliegenden Studie insgesamt 19 UX-Probleme identifiziert worden sind. Abbildung 3 verdeutlicht, dass der synchrone RUT mit 13 identifizierten UX-

Problemen die meisten UX-Probleme dieses Schweregrads identifizieren konnte. Es folgte der asynchrone RUT mit einer Anzahl von zehn identifizierten UX-Problemen. Den geringsten Wert im Vergleich konnte der UX-Test im Labor mit einer Anzahl von neun identifizierten UX-Problemen aufweisen. Im Vergleich zu den als "Critical" eingestuften UX-Problemen gestaltet sich die Verteilung in diesem Fall deutlich homogener. Zwar konnte der synchrone RUT erneut die meisten UX-Probleme identifizieren, doch ist hier die Differenz zwischen den Evaluationsmethoden nicht weitreichend genug, um von einem signifikanten Einfluss zu sprechen (Hypothese 2). Methodenübergreifend wurden insgesamt 19 verschiedene UX-Probleme des Schweregrads "Cosmetic" identifiziert. Abbildung 3 zeigt ein ausgeglichenes Verhältnis in Bezug auf die Anzahl aufgefundener Probleme zwischen UX-Test im Labor, asynchronem RUT und synchronem RUT. Demnach wurden durch jede Evaluationsmethode acht verschiedene UX-Probleme identifiziert. Abbildung 3 ist zu entnehmen, dass ein UX-Test, unabhängig von seiner Form, lediglich 42 % aller identifizierten UX-Probleme des Schweregrads "Cosmetic" aufdeckt. Der subjektive Charakter von UX-Problemen dieses Schweregrads erklärt die insgesamt dennoch hohe Anzahl aufgefundener Probleme. Die Homogenität zwischen UX-Test im Labor, asynchronem RUT und synchronem RUT ist dahingehend zu erklären, dass "Cosmetic"-Probleme dem Probanden zwar optisch oder ergonomisch negativ auffallen, sie jedoch keinen weitreichenden Einfluss auf die Systemnutzung ausüben. Da viele subjektive Meinungen vorherrschen, ist es schwer, Gemeinsamkeiten zu identifizieren.

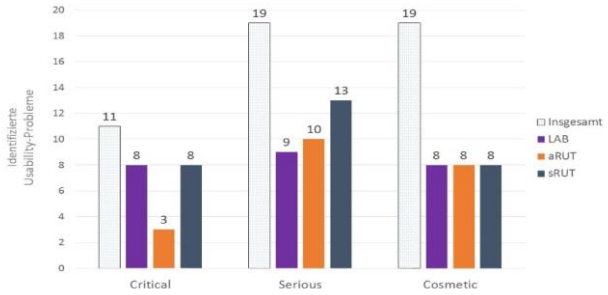

Abbildung 3: Identifizierte UX-Probleme nach Schweregrad

Die vorangehenden Ergebnisse zeigen, dass die durch einen Testleiter begleiteten Methoden mehr UX-Probleme identifizierten als der automatisierte, asynchrone RUT (Hypothese 3). Dies ist dahingehend zu erklären, dass die Probanden durch die Moderation des Testleiters in ihren Handlungsabläufen unterstützt wurden. Durch die Kommunikation mit dem Testleiter fiel es den Probanden leichter, ihre Eindrücke verbal zu äußern. Weiterhin ließ sich im Verlauf der Evaluation erkennen, dass sich die Probanden durch die direkte Interaktion mit einem ausgebildeten Testleiter deutlich stärker in das Testgeschehen eingebunden fühlten. Dies hat zur Folge, dass die Probanden konzentrierter und zielgerichteter bei der Bearbeitung der Aufgaben vorgingen, wodurch die Rückmeldungen quantitativ zunahmen. Durch die direkte Kommunikation zwischen Proband und Testleiter war eine unmittelbare Reaktion auf das Verhalten der Probanden möglich. Sobald der Testleiter Auffälligkeiten in den Aktionen des Probanden feststellte, konnte dieser die Handlungen des Probanden durch gezielte Fragestellungen hinterfragen und so besser nachvollziehen.

3.3 Frequenz der aufgefundenen UX-Probleme

Abbildung 7 und 8 zeigen für alle angewandten Evaluationsmethoden, dass die durchschnittliche Frequenz identifizierter UX-Probleme abnimmt, sofern sich deren Schwere verringert. Weiterhin wird deutlich, dass sich die Moderation eines Tests durch einen erfahrenen UX-Experten positiv auf eine stabile Frequenz von UX-Problemen auswirkt. Die Moderation des UX-Tests durch einen Testleiter bietet das Potential, auf das individuelle Verhalten der Probanden zu reagieren, was sich positiv auf die Frequenz der identifizierten UX-Probleme auswirkt. Dies ist im Kontext eines automatisierten Verfahrens nicht möglich. Handelt es sich bei einem Probanden zum Beispiel um eine eher introvertierte Person, ist anzunehmen, dass sie weniger Eindrücke schildert als eine extrovertierte Person. Ein Testleiter kann im Fall der Moderation auf einen solchen Sachverhalt reagieren und zum Beispiel durch gezielte Fragen Erwartungen und Meinungen der Probanden ergründen. Durch gezieltes Nachfragen des Testleiters ist es möglich, Unklarheiten der Probanden zu identifizieren und dadurch mehr UX-Probleme zu identifizieren.

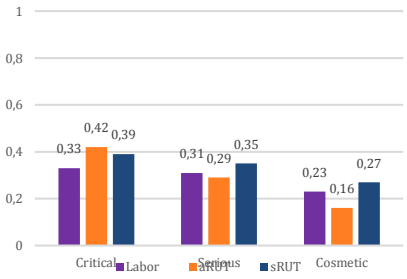

Abb. 7: Frequenz nach Schweregrad

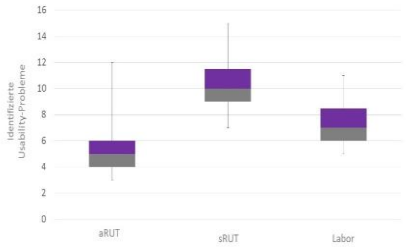

Abb. 8: UX-Probleme pro Proband

4 Fazit

Im Verlauf der Analyse konnte die Erkenntnis gewonnen werden, dass die durch einen Testleiter begleiteten Methoden mehr UX-Probleme identifizierten als ein automatisiertes Verfahren. Der Sachverhalt ist dahingehend zu deuten, dass die Probanden durch die Moderation des Testleiters in ihren Handlungsabläufen unterstützt wurden. Des Weiteren ließ erkennen, dass sich die Probanden durch die direkte Interaktion mit einem professionellen Testleiter deutlich stärker in das Testgeschehen einbinden ließen. So wirkten die Probanden konzentrierter und zielgerichteter bei der Bearbeitung der Aufgaben, wodurch die Rückmeldungen sowohl qualitativ als auch quantitativ zunahmen. Durch die direkte Kommunikation zwischen Proband und Testleiter war eine unmittelbare Reaktion auf das Verhalten der Probanden möglich. Dadurch fiel es den Probanden demnach leichter, ihre Eindrücke verbal zu äußern. Ein gegenteiliges Verhalten ließ sich im Verhalten der Probanden des asynchronen RUTs beobachten, da die Probanden eher zurückhaltend wirkten. Folglich wurden im Kontext der moderierten Evaluationsmethoden durchschnittlich mehr UX-Probleme identifiziert als im automatisierten Verfahren.

Bei der mobilen UX-Evaluation ist zu beachten, dass bis dato keine Methode an die Standards aus dem Desktop-Bereich heranreicht, unabhängig davon, ob synchron oder asynchron. Das synchrone Remote-UX-Testing im mobilen Bereich liefert sehr gute Ergebnisse, ist jedoch mit einem hohen technischen Aufwand verbunden. Asynchrone Verfahren stehen vor der Herausforderung, die automatisierte Moderation in das zu testende Objekt einzubetten. Ohne diese Funktion kann es besonders im Kontext schwerwiegender UX-Probleme zu Schwierigkeiten kommen. Es gilt bis heute, dass die alleinige Erhebung von Nutzungs- und Befragungsdaten mithilfe asynchroner Tools (meist) nicht ausreicht, um eine Website optimieren zu können. Beide Remote-Verfahren bieten dennoch eine Möglichkeit, um zum Beispiel eine schwer zu rekrutierende Zielgruppe zu testen. Der Mehraufwand in der technischen Infrastruktur oder Abstriche bei den Ergebnissen müssen jedoch beachtet werden.

Literaturverzeichnis

Albert, W. & Tullis, T. (2013). *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Elsevier Science.* New York: Interactive Technologies

Budui, R. (2015) *Mobile User Experience: Limitations and Strengths.* Online. Verfügbar unter: https://www.nngroup.com/articles/mobile-ux/- Letzter Zugriff: 09.04.2016

Heinsen, S. (2003) *Usability praktisch umsetzen: Handbuch für Software, Web, Mobile Devices und andere interaktive Produkte*. München: Hanser

Krannich, D. (2010). *Mobile System Design*:. Norderstedt: Books on Demand

Molich, R. (2000). *User-Friendly Web Design.* Kopenhagen: Ingenioren Book

Rauch, D. (2016) *Mobile Remote-Usability-Tests – Eine kritische Beleuchtung der aktuellen UX-Tool-Landschaft.* Online. Verfügbar unter: http://www.usabilityblog.de/2016/03/mobile-remote-usability-tests-eine-kritische-beleuchtung-der-aktuellen-ux-toollandschaft/ Letzter Zugriff: 09.04.2016

Sarodnick, F. & Brau, H. (2011) *Methoden der Usability Evaluation: wissenschaftliche Grundlagen und praktische Anwendung.* München: Huber

Tulathimutte, T. & Bolt, N.: *Remote Research.* Brooklyn: Rosenfeld Media

Autor

Rauch, Dustin

Dustin Rauch ist seit Oktober 2015 als Junior User Experience Consultant bei eResult in Göttingen tätig. Er studierte Internationales Informationsmanagement an der Universität Hildesheim und spezialisierte sich in seinem Masterstudium auf die Konzeption von Websites und Applikationen.