Oliver Sieler

Was kann Crowd-Usability-Testing?

Das Testen von Websites, Apps oder Software wird immer öfter in der Crowd realisiert. Doch wie effektiv ist ein Crowd-Usability-Test (CUT) wirklich im Vergleich mit einem klassischen Usability-Test im Labor? Und welche Rolle spielt dabei der Erfahrungsgrad der teilnehmenden Probanden? Eine Untersuchungsreihe, die beide Testumgebungen und unterschiedlichste Testteilnehmer einbindet, soll etwas mehr Aufschluss darüber geben.

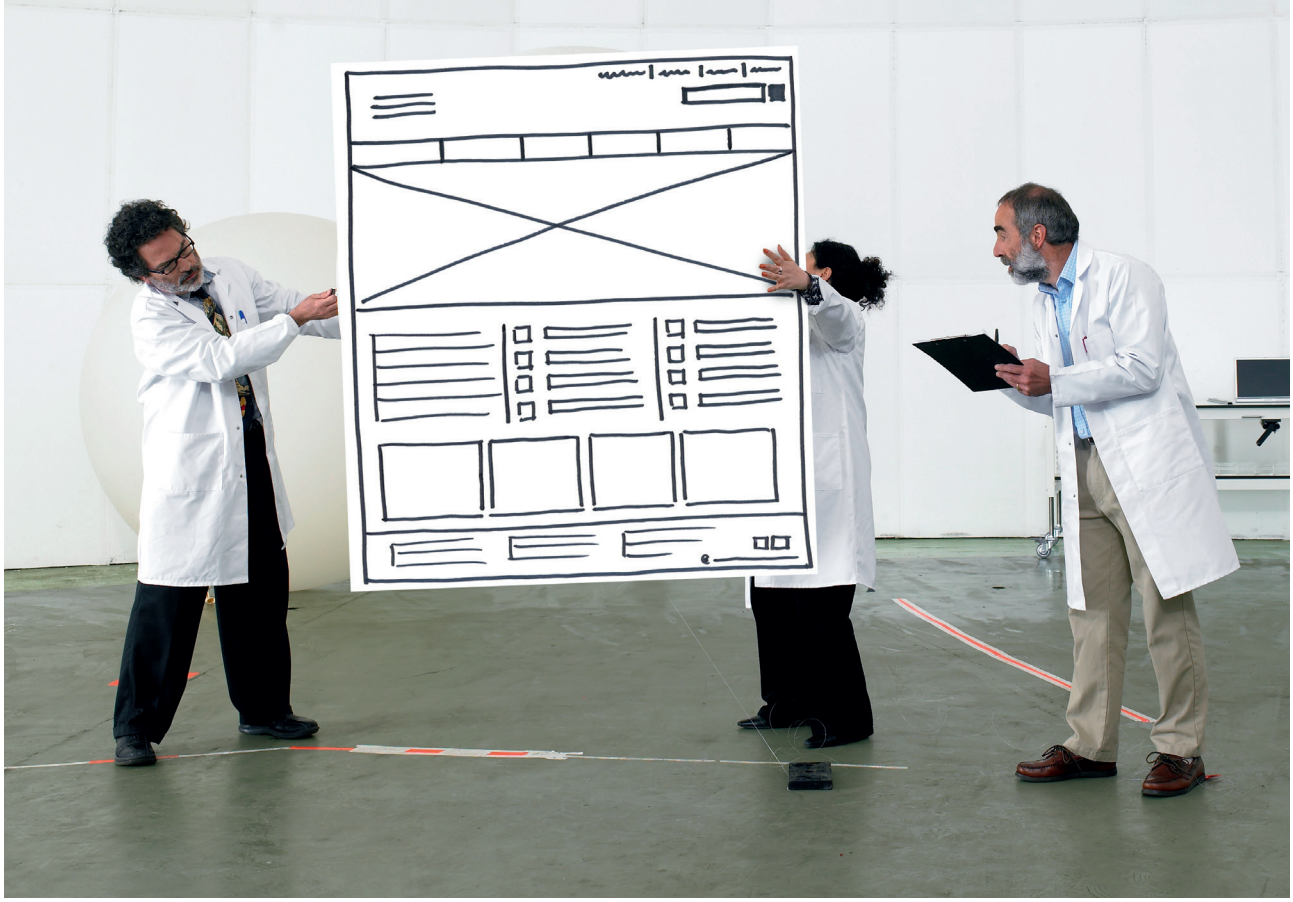

Foto: Michael Blann / thinkstockphotos.de

DER AUTOR

Oliver Sieler ist wissenschaftlicher Mitarbeiter am Zentrum für eLearning der Hochschule Zittau/Görlitz. Zuvor war er Projektmanager bei der Firma test.io.

Für das Evaluieren von Produkten wird in den letzten Jahren seitens der Unternehmen verstärkt auf das Testen mittels einer Crowd gesetzt. Gerade bei kleineren Website-Projekten findet diese Methode Anklang. Dabei verwalten Anbieter von Crowd-Tests eine Tester-Community, die auf Abruf als Probanden-Pool für anstehende Untersuchungen fungiert. Häufig wird Remote-Testing (wie Crowd-Testing auch genannt wird) dazu benutzt, um funktionelle Fehler, also Bugs, auf Webseiten zu finden, die dann protokolliert und an den Auftraggeber weitergeleitet werden können.

Der Ablauf eines klassischen Usability-Tests im Labor

Für einen Labor-Test werden in der Regel fünf bis zehn Probanden eingeladen, die mit dem Produkt konfrontiert werden. Dabei gilt es, verschiedene Aufgaben zu lösen, um unerwünschte Usability-Probleme aufzudecken. Hartnäckig hält sich immer noch das Gerücht, eine geringe Teilnehmerzahl von fünf oder mehr Probanden reiche aus, um nahezu alle Probleme zu identifizieren. Dieses Dogma wurde jedoch bereits im Jahr 2005 von den Usability-Pionieren Rolf Molich und Jakob Nielsen hinreichend

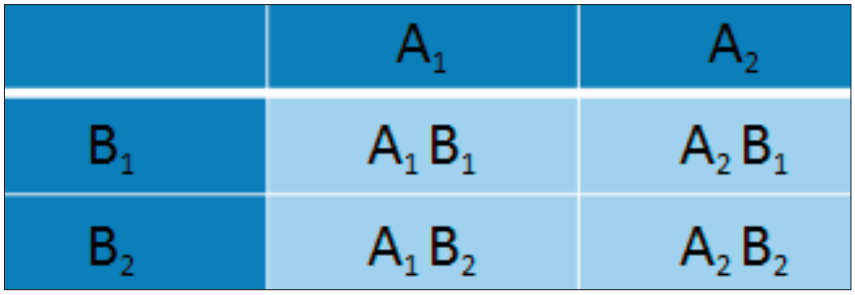

Abb.1: Durch die Einteilung der Probanden in Crowd/Labor und Pro/Newbies entstand ein mehrfaktorielles Design

widerlegt. Während die Probanden versuchen, nutzerrelevante Aufgaben zu lösen, beobachten und protokollieren Mitarbeiter, welche Probleme während der Interaktion mit dem System auftreten. Oft wird dafür ein Venezianischer Spiegel verwendet oder eine Kamera, mit der das Bild in einen anderen Raum übertragen wird. Im Idealfall befinden sich dort ebenfalls Entwickler und Produktmanager des beauftragenden Unternehmens. Während der gesamten Untersuchungsreihe ist ein Versuchsleiter im Aufzeichnungsraum, der den Ablauf überwacht und den Probanden bei etwaigen Problemen unter die Arme greift.

Der Remote-Test

Für die Untersuchung wurde ein asynchroner Remote-Usability-Test (ARUT) durchgeführt. Das heißt, dass die Teilnehmer ihre Aufzeichnung zeitlich unabhängig und selbstständig zu Hause (oder wo auch immer) durchführen. Die erstellten Videos werden anschließend auf einen Server geladen, von wo die Mitarbeiter der Agentur die Videos abrufen und auswerten können.

Um die Auswertungen der gewonnenen Daten nicht falsch zu interpretieren und zu untermauern, wird häufig die Methode des lauten Denkens (oder auch Thinking Aloud) bei der Aufzeichnung der Untersuchung angewendet. Dabei kommentieren die Probanden während der Interaktion mit dem System ihre Gedankengänge und Entscheidungen. Durch die Verbalisierung ihrer Gedanken sind diese Entscheidungen für den Verantwortlichen besser nachvollziehbar und vereinfachen eine Analyse erheblich. Theoretischer Hintergrund des Verfahrens sind die introspektiven Erhebungsmethoden, deren gemeinsames Merkmal darin besteht, dass die beteiligten Probanden zur Verbalisierung ihrer Gedanken, Wahrnehmungen und Empfindungen aufgefordert werden. Ein großer Vorteil dieser Methode ist die Unmittelbarkeit der Kommentare. Eine nachträgliche Rationalisierung durch den Probanden wird weitgehend vermieden, Fehlinterpretationen durch den Beobachter sind nahezu ausgeschlossen. Die Ergebnisse dieser Methode hängen allerdings stark von der Fähigkeit der Versuchsperson ab, sich zu artikulieren. Die Methode des lauten Denkens wurde sowohl bei den Probanden im Labor als auch bei den Teilnehmern in der Crowd praktiziert.

Vorteile

Die Methode des Crowd-Testings bietet eine Vielzahl von Vorteilen gegenüber dem klassischen Labor-Test. Sie profitiert von einer problemlosen Rekrutierung, da die Tester nicht physisch anwesend sein müssen. Es kann quasi rund um die Uhr getestet werden – Zeitverschiebungen spielen beim Crowd-Testing keine Rolle. Eine frühzeitige Einbindung und Durchführung von Usability-Maßnahmen ist bei der Produktentwicklung unerlässlich. Deswegen ist Crowd-Testing gerade für das iterative Vorgehen in der agilen Softwareentwicklung von großem Vorteil – hier wird nicht nur Zeit, sondern auch Geld gespart. Da das Crowd-Testing neben den Zeitzonen ebenfalls keine Rücksicht auf politische Grenzen oder Sprachbarrieren nimmt, können auch internationale Tests durchgeführt werden, bei denen der kulturelle Hintergrund der Probanden möglicherweise eine Rolle spielt. Nigel Bevan, der Herausgeber von Teil 11 der ISO 9241 (Anforderungen an die Gebrauchstauglichkeit), weist darauf hin, dass es sich bei Usability um eine Qualität der Nutzung handelt. Es lässt sich also nicht anhand der reinen Produkteigenschaften entscheiden, ob ein Produkt gebrauchstauglich ist oder nicht. Das Produkt muss immer in Bezug zum Nutzungskontext betrachtet werden. Da die Tester die Website in einem habituellen Umfeld erkunden, sollte ein wesentlicher Bestandteil des Nutzungskontextes beim Crowd-Testing bereits gegeben sein.

Nachteile

Viele Unternehmen stehen dem Crowd-Testing weiterhin kritisch gegenüber – und das oft zu Recht. Manche Tests sind schon für rund 40 Euro pro Teilnehmer zu haben. Der geringe Kostenaufwand für den Stakeholder (Auftraggeber) ist leider in vielen Fällen auf eine zu knappe und unzureichende Durchführung der Untersuchungen zurückzuführen. Es wird kritisiert, dass Crowd-Tests nur einen kleinen Teil des klassischen Usability-Tests im Labor abbilden – nämlich die Rekrutierung der Probanden sowie die Durchführung und Aufzeichnung der Tests. Bei einigen Crowd-Test-Anbietern müssen die Stakeholder selbst die Szenarios erstellen, was oftmals zu Komplikationen beim Untersuchungsablauf führt. Grund dafür ist das mangelnde Expertenwissen, wodurch ein Szenario erstellt wird, in dem Fehler enthalten sind oder die einzelnen Aufgabenstellungen in einer Sackgasse münden. Da bei der Aufzeichnung der Video-Captures kein

Versuchsleiter anwesend ist, müssen bei Unklarheiten in der Aufgabenstellung die Tester dann die Untersuchung abbrechen. Wiederholungen mit neuen Testern kosten dann wieder mehr Zeit und Geld und am Ende fehlt es dem Stakeholder an ausreichenden Resultaten. Auch für die Auswertung der Video-Captures bedarf es ausreichender Expertise im Umgang mit Crowd-Tests. Ebenso ist die auf den ersten Blick scheinbar unkomplizierte und günstige Art der Teilnehmerrekrutierung nur mit Vorsicht zu genießen, da die Rekrutierungskriterien sehr allgemein gehalten sind. Viele Zielgruppen können gar nicht erreicht werden, dafür scheint die Zielgruppe „technikbegeisterter Power-User" und „selbsterklärter Experte" überproportional vertreten. Die Rekrutierung unbelasteter Tester, die den wenig erfahrenen Durchschnittsbesucher von Reiseportalen o. Ä. repräsentieren, wird zwar immer garantiert, doch die Realität sieht meistens anders aus.

Usability & Usability-Problem

Zwischen Usability-Experten und Auftraggebern herrscht oft eine Diskrepanz über die Auffassung der Termini Usability und Usability-Problem. Ein übereinkommendes Verständnis darüber ist jedoch wichtig für den Aufbau der Untersuchung, dafür, was überhaupt gemessen werden soll und welche Ergebnisse für eine Verbesserung oder Weiterentwicklung des Produkts substanziell sind.

Im deutschsprachigen Raum sind synonyme Begriffe für Usability etwa Gebrauchstauglichkeit, Nutzbarkeit, Nützlichkeit, Benutzerfreundlichkeit oder aber auch leichte Handhabung. Usability verfolgt das Ziel, Körper und Geist der Nutzer eines Systems in eben dieses zu integrieren. Usability ist im Gegensatz zur Ergonomie keine eigenständige Disziplin, sondern vielmehr das Ziel der Gestaltung nach den Vorgaben und Erkenntnissen der Software-Ergonomie. Nach Nielsen bildet die Usability (Bedienbarkeit) zusammen mit der Utility (Nützlichkeit) die Usefulness (Brauchbarkeit). Bei Utility steht die Funktionalität im Vordergrund, die für eine bestimmte Aufgabe notwendig ist. Die Usability ebnet den Zugang zu eben diesen Funktionen. Usability wird also mit dem Grad der Einfachheit der Nutzung in Relation gesetzt. Im Gegensatz zu Nielsens Definition bezieht die DIN EN ISO 9241-11 (Anforderungen an die Gebrauchstauglichkeit) ebenfalls den Nutzungskontext mit ein. Dort wird Gebrauchstauglichkeit definiert als „das Ausmaß, in dem ein Produkt durch bestimmte Benutzer in einem bestimmten Nutzungskontext genutzt werden kann, um bestimmte Ziele effektiv, effizient und zufriedenstellend zu erreichen". Um die Gebrauchstauglichkeit einer Website zu ermitteln, empfiehlt es sich, auch die Maße Effektivität, Effizienz und Zufriedenheit weiter zu zerlegen. Es werden für die Untersuchung Ziele und Teilziele identifiziert, der Nutzungskontext- und die Umgebung beschrieben und das Verhältnis von gewünschten und erreichten Zielen während der Nutzung miteinander verglichen. Nach der Untersuchung liegt der Fokus des Evaluators beim Messen der erlebten Nutzungsqualität während der Interaktion mit der zu testenden Website. Dazu gehören sowohl die objektiven Maße (bspw. benötigte Zeit für erledigte Aufgabe) als auch subjektive Maße (benötigte Hilfestellung durch den Versuchsleiter).

Im Rahmen der Untersuchungsreihe wurde lediglich die Effektivität von Crowd-Testing und dem Usability-Test im Labor miteinander verglichen (Effektivität bezeichnet die Genauigkeit und Vollständigkeit, mit der ein bestimmtes Ziel erreicht werden kann). Dazu wurde die Anzahl der gelösten Aufgaben während der Untersuchungen festgehalten. Um die Effizienz der Website zu ermitteln, würde man bspw. die Zeiten der einzelnen Probanden für die Erledigung der Aufgaben messen. Die Nutzerzufriedenheit könnte man bspw. mit Fragebögen oder Interviews erfassen.

Für die Auswertung der Ergebnisse ist es wichtig zu definieren, was ein Usability-Problem ist. Nicht jeder misslungene Dialog mit dem zu testenden System ist automatisch auch ein Usability-Problem. Die Probanden der Untersuchung sollten auch ausreichendes Domänenwissen für das zu testende System besitzen. So sollte beispielsweise eine gelernte Einzelhandelskauffrau ohne Bezug zu Internettechnologien nicht als Proband für das Testen von Datenbanksystemen fungieren. Daher muss bereits bei der Planung der Usability-Untersuchung die Repräsentativität der Benutzer eingeplant werden. Ein Usability-Problem ist z. B. alles, was die Aufgabenerledigung verhindert, was ein gewisses Maß an Verwirrung schafft, wenn ein Dialog falsch gedeutet wird oder wenn der Benutzer die Navigation nicht versteht.

Zur Untersuchung

Für die Untersuchungreihe wurden insgesamt 24 Probanden akquiriert. Die Hälfte der Probanden hatte zuvor noch nie an einer Usability-Untersuchung teilgenommen oder dahin gehend eine fachliche Ausbildung erhalten – diese Teilnehmer werden nachfolgend Newbies genannt. Die andere Hälfte (Pros) hingegen war auf diesem Gebiet bereits deutlich erfahren. Die Pros hatten im Vorfeld zwischen 7- und 28-mal an ähnlichen Tests partizipiert und waren teilweise selbst als Usability-Professionals beruflich tätig. Durch die Aufteilung in Pros und Newbies sowie in Crowd- und Laborteilnehmer entstand ein Untersuchungsdesign, wie in Abbildung 1 zu sehen ist. Jeder der Probanden (Labor und Crowd) hatte insgesamt drei Aufgaben zu lösen.

	Labor	Crowd	Gesamt
Pro	66	19	85
Newbie	57	36	93
Gesamt	123	55	178

Abb. 2: Übersicht über die Anzahl aller gefundenen Usability-Probleme

Ergebnisse

Insgesamt wurden 27 unterschiedliche Usability-Probleme von den Probanden aufgedeckt. Zählt man alle Probleme auf, die von den Teilnehmern moniert wurden, so erhalten wir eine Summe von 178 (siehe Abbildung 2). Wie dort ebenfalls zu sehen ist, waren die Pros im Labor die stärkste Gruppe mit 66 entdeckten Usability-Problemen. Auf die wenigsten Usability-Probleme stießen (sehr überraschend) die Pros in der Crowd. Insgesamt kann man sehen, dass Pros und Newbies sich mit 85 zu 93 entdeckten Usability-Problemen ungefähr die Waage halten. Anhand der Tabelle kann eindeutig abgelesen werden, dass das Setting, also die Testumgebung, das entscheidende Kriterium ist. Im Labor traten insgesamt 123 Schwierigkeiten auf, hingegen in der Crowd lediglich 55.

Woher rührt diese Diskrepanz? Anbieter von Crowd-Tests werben gerne mit erfahrenen Testern. Doch genau dort liegt ein immanentes Problem. Erfahrene Tester, die teilweise auch ausgewiesene Experten sind, nehmen bei vielen Interaktionen Barrieren gar nicht wahr. Wo eine Hausfrau die Suchfunktion oder die Navigation nicht versteht, sieht der Pro keine Probleme. Diese Problematik wird auch Expert-Effect genannt. Und in der Crowd wird dieser Effekt verstärkt. Viele der routinierten Tester nehmen häufiger an funktionellen Tests teil, in denen sie Bugs finden müssen, für die sie dann bezahlt werden. Mehr Bugs, mehr Geld. Bei Usability-Tests kann dann beobachtet werden, dass diese Routiniers Usability-Probleme regelrecht suchen. Ungeübte Probanden stoßen auf Probleme, obwohl sie die Untersuchung eigentlich möglichst ohne Hindernisse überstehen wollen.

Bei den Pros wurde der Mangel an Unerfahrenheit dadurch kompensiert, dass sie versuchten, mittels Perspektivübernahme Verbesserungsvorschläge abzugeben. Aufgrund ihrer Erfahrung und Expertise gaben sie überdurchschnittlich viele Kommentare hinsichtlich Layout, Farben, Schriftgrößen etc. ab, die jedoch keine Usability-Probleme darstellen und somit nicht in die Auswertung einflossen. Die Probanden aus der Crowd, speziell die Erfahrenen, benannten überwiegend Probleme, die nicht besonders kritisch sind und für die Interaktion der Website mit repräsentativen Nutzern nicht besonders relevant gewesen wären.

Ein weiterer Aspekt ist die geringe Menge an Informationen, die bei der Auswertung der Crowd-Untersuchung zur Verfügung standen, da lediglich der Screen der Probanden aufgezeichnet wurde. Eine direkte Beobachtung der Probanden oder wenigstens ein Video-Capture von Gesicht und Oberkörper wäre für die Interpretation der Interaktionen äußerst nützlich. Und hierzu bedarf es auch jemandes, der darin geschult ist, Mimik, Gestik und Körperhaltung des Probanden professionell einzuschätzen.

Bei der Labor-Untersuchung ist der Versuchsleiter fester Bestandteil des Settings und durch ihn können weit mehr Informationen generiert werden, gerade weil er involviert ist. Bei Problemen, die während der Untersuchung auftreten, kann er Hilfestellung geben und gezielt nachfragen, warum etwas schiefgelaufen ist, solange beim Probanden die Informationen noch aktuell und leicht abrufbar sind.

Warum also Crowd-Testing?

Die Ergebnisse der Untersuchungsreihe sollten nicht generalisiert werden. Obwohl der klassische Usability-Test im Labor als empirische Methode immer noch den Königsweg darstellt, ist Crowd-Testing eine sinnvolle und nützliche Ergänzung, in manchen Fällen auch Alternative. Gerade während der Entwicklungsphase ist ein Test via Crowd zu empfehlen. Er stellt ein kostengünstiges und zeitunaufwendiges Pendant zum Usability-Test dar und kann auch spontan realisiert werden. Des Weiteren lässt er sich sinnvoll mit Fragebögen, Klickpfadanalyse oder einer Expertenevaluation kombinieren. Um noch aussagekräftigere Ergebnisse zu erzielen, ist es denkbar, dass ein Versuchsleiter den Probanden online begleitet. Bei Komplikationen kann er Hilfestellung geben oder den Tester während einer anschließenden Videokonfrontation zu bestimmten Problemen befragen.

Inzwischen gibt es einige Anbieter von Crowd-Tests, die über ausreichendes Know-how verfügen und mit exzellenten Usability-Professionals aufwarten. Wer sich hier im Vorfeld über potenzielle Agenturen informiert, kann, egal ob junger Entrepreneur oder gewiefter Marketing-Manager, mithilfe von Crowd-Testing wichtige Erkenntnisse erlangen, um sein Projekt zu realisieren. ¶