Validierung von Web-Usability-Heuristiken für eine ältere Zielgruppe

Victoria Böhm [1], Andrea Langer[1], Christian Wolff[1]

Lehrstuhl Medieninformatik, Universität Regensburg[1]

Victoria.Boehm@ur.de, Andrea.Langer@stud.uni-regensburg.de, Christian.Wolff@ur.de

Zusammenfassung

Dieser Beitrag stellt die Validierung von spezifischen Web-Usability-Heuristiken für die ältere Zielgruppe vor. Das Heuristiken-Set stammt aus einer Vorarbeit und wurde literarturbasiert abgeleitet. Zur Validierung werden ein Nutzertest mit zehn Probanden aus der Zielgruppe sowie eine heuristische Evaluation mit drei Experten durchgeführt. Die Ergebnisse beider Evaluationsmethoden werden quantitativ und qualitativ miteinander verglichen und etablierte Kennwerte zum Methodenvergleich berechnet. Beide Evaluationen liefern ähnliche Werte bezogen auf Validität und Effektivität. Aufgrund der hohen Anzahl an *unique problems* fällt die Gründlichkeit geringer aus, als bei anderen Methodenvergleichen.

1 Motivation und Zielsetzung

Die heuristische Evaluation ist ein kostengünstiges und etabliertes Verfahren zur Usability-Evaluation auf Grundlage der Bewertung durch mehrere Experten (Molich & Nielsen, 1990). Ein Grund für die Erstellung und Nutzung von Web Heuristiken für Ältere ist die wachsende Zahl an älteren Usern, die das Internet nutzen (Anderson & Perrin, 2018), sowie die erschwerte Rekrutierung für Nutzertests bzw. der besonders herausfordernde Prozess der Nutzerselektion bei Zielgruppen mit hoher Diversität und Heterogenität (Aarhus et al., 2010). Eine Expertenin-spektion mit geeigneten Regeln kann hier einen Nutzertest sinnvoll ergänzen.

Zur Durchführung einer heuristischen Evaluation stehen eine Vielzahl von Heuristiken zur Verfügung. Am verbreitetsten ist das Heuristiken-Set von Nielsen, wobei die Anzahl an Heuristiken, die auf eine spezifische Nutzergruppe oder Domäne zugeschnitten sind, wächst. In Böhm und Nguyen (2017) wurde durch Analyse und Synthese verschiedener existierender Guidelines eine neue Sammlung von Heuristiken für die ältere Zielgruppe erstellt. Nach der

Veröffentlicht durch die Gesellschaft für Informatik e. V. 2018 in
R. Dachselt, G. Weber (Hrsg.):
Mensch und Computer 2018 – Workshopband, 02.–05. September 2018, Dresden.
 https://doi.org/10.18420/muc2018-ws11-0515

Kombination von Regeln unterschiedlicher Herkunft wurden Redundanzen durch ein *card sorting* entfernt. In dieser Arbeit sollen diese Heuristiken validiert und ein Ausblick zur weiteren Verbesserung gegeben werden. Zunächst werden verschiedene Verfahren zur Validierung von Heuristiken diskutiert. Im Anschluss werden das eigene Vorgehen und die daraus gewonnenen Erkenntnisse geschildert.

2 Forschungskontext

2.1 Ansätze zur Validierung von Heuristiken

Aktuelle Literaturanalysen, die Studien zur Entwicklung von Heuristiken betrachten zeigen, dass nur ein kleiner Teil der entwickelten Heuristiken überhaupt validiert wird (Botella et al., 2013; Hermawati & Lawson, 2016; Quinones & Rusu, 2017). Wenn eine Validierung stattfindet, dann werden zumeist folgende Verfahren eingesetzt (Hermawati & Lawson, 2016; Ling & Salvendy, 2005): Der konkurrierende Einsatz mit einer anderen Heuristik, die Durchführung einer *Case Study*, die Durchführung von Nutzerstudien oder die Sammlung von Feedback zu bestimmten Aspekten der Regeln über einen Fragebogen (Botella et al 2013). Nur wenige der von Hermawati & Lawson analysierten Studien ermittelten die für Heuristiken gebräuchlichen Kennzahlen von Hartson et al. (2003). Folgt man den wenigen Studien, die solche Maße berechneten, scheinen spezielle Heuristiken eine höhere Performanz aufzuweisen als generelle Heuristiken (Hermawati & Lawson, 2016, 12f).

2.2 Quantitative Kennzahlen

Im Folgenden werden die drei Maße definiert, die zur Validierung von Heursitiken verwendet werden. Die *Gründlichkeit* (*thoroughness*) gibt das Verhältnis der Probleme, die durch eine Evaluierungsmethode gefunden wurden, zu allen existierenden Problemen auf der Webseite an (Sears, 1997; Hartson et al., 2013). Die *Validität* (*validity*) gibt den Anteil an echten, relevanten Problemen unter den ermittelten Befunden an (Sears, 1997, 214f.) und die *Effektivität* (*effectiveness*) ist das Produkt aus Validität und Gründlichkeit (Hartson et al., 2003, 165).

- $Gründlichkeit = \frac{Anzahl\ an\ gefundenden\ Problemen}{Anzahl\ an\ existierenden\ Problemen}$
- $Validität = \frac{Echte\ Probleme}{Problemkandidaten}$
- *Effektivität = Gründlichkeit * Validität*

Da einige Kennzahlen die Berechnung von *echten* Usability-Problemen erfordern, diskutieren Hartson et al. verschiedene Ansätze zur Generierung einer Liste an *echten Problemen*. Zum einen wird das *Seeding* von Problemen genannt, also das gezielte Einbauen von Problemen (Hartson et al., 2003, 156). Daneben wird die Durchführung eines Labortests als Gold-Standard betrachtet, der eine definierte Liste an *echten Problemen* liefert. Daneben existiert ein dritter Ansatz, bei welchem zur Ermittlung der *echten Probleme* die Subsets der Methoden vereinigt werden. Dieses Verfahren wird nachfolgend eingesetzt. Die Liste an *echten Problemen*, A (X), wird folgendermaßen definiert und berechnet:

- $A(X) = P(X) \cup Q(X) \cup R(X)$, wobei $P(X)$ das Problem-Set ist, das durch Methode *P* ermittelt wurde, $Q(X)$ das durch Methode *Q* und $R(X)$ das mit Methode *R*. (Hartson et al., 2003, 164)

Setzt man das in die obige Formel für die *Validität* ein, ergibt sich immer der Wert 1; Es ist es daher nicht möglich, die Validität zu berechnen bei Verwendung dieser Methode. (Hartson et al., 2003, 164)

- $Validität = \frac{|P(x) \cap A(x)|}{|P(x)|} = \frac{|P`(x)|}{|P(x)|} = 1$

Um dieses Problem zu umgehen, wurde die Definition von *echten* Befunden modifiziert: Die erzeugten Problemsets wurden wie von Hartson beschrieben vereinigt. Vorher wurde jedes Set jedoch geprüft und Auffälligkeiten entfernt, die nicht zu einem Problem bei der tatsächlichen Nutzung führen würden. Grundlage dieses Urteils bildet eine Beschreibung häufiger Beispiele für Usability-Probleme nach Tullis und Albert (2008, 100).

2.3 Heuristische Evaluation

Zur Validierung wurde eine heuristische Evaluation von drei Experten durchgeführt, von denen alle drei Erfahrung mit der Methode selbst haben und zwei davon mit User Interface Design für ältere User. Als Testobjekt für die Evaluation diente die Online-Apotheke **shop.apotal.de,** wobei auf ein konsistentes Vorgehen bei der Bewertung geachtet wurde. Alle Experten verwendeten die Heuristiken in gedruckter Form während sie die Seite inspizierten. Um zu gewährleisten, dass alle wichtigen Bereiche der Apotheke betrachtet werden, wurden vorab folgende Bereiche für die Inspektion festgelegt: Die Suchfunktion, die Detailansicht eines Produkts, Browsen über Kategorien sowie der Warenkorb. Diese wurden in Abstimmung mit den Aufgaben für den Nutzertest festgelegt, sodass sich möglichst keine Unterschiede in den resultierenden Problemen durch andere Bereiche, die betrachtet bzw. benutzt werden, ergeben. Auch die Dauer der heuristischen Evaluation wurde vorab auf zwei Stunden festgelegt. Jedes identifizierte Problem wurde in einen Protokollbogen eingetragen und der Schweregrad bewertet, wobei drei Kriterien bewertet wurden: *Häufigkeit*, *Auswirkung* und *Persistenz*. Abschließend konnten die Experten Auffälligkeiten zur Verständlichkeit und Vollständigkeit der Regeln als Freitext formulieren und Kürzungsvorschläge nennen.

2.4 Nutzertest

Um die Effektivität der heuristischen Evaluation ermitteln zu können, wurde ein Nutzertest mit derselben Apotheke mit zehn Probanden durchgeführt. Andere Studien, die Heuristiken mit einem Labortest vergleichen, verwenden eine höhere Anzahl an Studienteilnehmern, da dort die Probleme des Nutzertests allein als Standardliste an *echten* Problemen für die Berechnung der Kennwerte herangezogen werden. In dieser Studie waren weniger Probanden ausreichend, da die Methode der Kombination von Listen für eine gute Produktivität der Evaluationen sorgt. Der eigentliche Test bestand aus sechs vorgegebenen Aufgaben, die alle problemorientiert formuliert waren. Nach dem Aufgabenteil füllte jeder Proband einen Fragebogen aus. Dieser enthielt alle soziodemografischen Angaben sowie Fragen zur Vorerfahrung. Alle

Tests fanden bei den Probanden zuhause statt, was bei der Forschung mit Älteren empfohlen wird (Grönvall & Kyng, 2013).

3 Auswertung und Ergebnisse

3.1 Aufbereitung der Befunde

Beim Nutzertest nahmen insgesamt zehn Probanden mit einer Altersspanne von 58 bis 80 Jahren. und einem Durchschnittsalter von 65,5 Jahren teil (SD=7,6). Im Schnitt nutzten die Probanden seit 14,8 Jahren das Internet (SD=9,1) und gaben an, den Computer 12 Stunden (SD=11,9) und das Internet 3,8 Stunden pro Woche zu nutzen (SD=2,9). Vorerfahrung mit einer Online-Apotheke besitzen vier der zehn Probanden. Zunächst wurde die heuristische Evaluation von drei Experten durchgeführt, die dabei Usability-Befunde in einem Problem-Template festhielten. Es wurden dabei von jedem Experten Kontext, Problembeschreibung, Schweregrad und die verletzte Heuristik angegeben. Danach wurde der Nutzertest durchgeführt und die Befunde ebenfalls im gleichen Template dokumentiert. Ausgewertet wurden die Befunde der einzelnen Methoden, indem zunächst Doppelungen innerhalb der Ergebnisliste einer Methode entfernt wurden. Hierdurch wird für beide Methoden eine Liste an Problemkandidaten erzeugt, siehe Spalte A in Tabelle 1. Anschließend wurden die beiden Listen auf *Echtheit* geprüft, um die Validität später ermitteln zu können. Damit ergab sich eine Liste an *echten* Problemen für die heuristische Evaluation und eine für den Nutzertest, siehe Spalte B. Im letzten Schritt wurden die Listen beider Methoden kombiniert, wiederum Doppelungen entfernt, sodass ein überschneidungsfreies Set an Problemen ermittelt wurde, siehe Spalte C.

3.2 Quantitative Ergebnisse

Betrachtet man die absolute Anzahl an ermittelten Problemen, schneiden beide Evaluationen ähnlich ab. Insgesamt konnten durch den Nutzertest 46 Problemkandidaten ermittelt werden, von denen 45 *echte* Probleme darstellten. Die heuristische Evaluation lieferte 44 Befunde, von denen 41 als *echt* zu bewerten sind. Die Vereinigung der *echten* Probleme wird als die Anzahl aller existierenden echten Probleme angesehen und umfasst 75 Probleme. Die *unique problems*, das sind Probleme, die jeweils nur mit einer Evaluationsmethode gefunden werden, sind bei beiden Evaluation sehr hoch: Der Nutzertest lieferte 34 *unique problems* und die heuristische Evaluation 30 *unique problems*. Elf Probleme wurden mit beiden Methoden gefunden.

Methode	(A) Ermittelte Problemkandidaten	(B) Ermittelte *echte Probleme*	(C) *Unique problems*	(D) Anzahl existierender Probleme	Gründlichkeit	Validität	Effektivität
Nutzertest	46	45	34	75	0,6	0,98	0,59
Heuristik	44	41	30	75	0,55	0,93	0,51

Tabelle 1: Ermittelte Kennwerte

Der Nutzertest schneidet mit einer ermittelten Gründlichkeit von 0,6 gerinfügig besser ab als die heuristiche Evaluation mit 0,55. Erklären lässt sich dieser Wert zunächst durch die kleine Schnittmenge von Problemen, also solchen, die in beiden Methoden identifiziert werden. Bezogen auf die Validiät ergeben sich für beide Methoden hohe Werte: Es werden hauptsächlich *echte Probleme* ermittelt. Die Validität für den Nutzertest beträgt 45/46= 0,98 und für die heuristischen Evaluation 41/44= 0,93. Bezogen auf die Effektivität der beiden Methoden ergaben sich folgende Werte: Der Nutzertest weist eine Effektivität von 0,6 x 0,98 = 0,59 auf und die heuristische Evaluation 0,55 x 0,93 = 0,51. Somit schnitt der Nutzertest gerinfügig besser ab als die Heuristische Evaluation.

3.3 Qualitative Ergebnisse

Da sich sehr unterschiedliche Befunde durch die zwei Methoden ergaben, soll zusätzlich betrachtet werden, um welche Art von Problemen es sich handelt. Beim Nutzertest fällt ins Auge, dass viele der Probleme entstehen, weil die Probanden eine Funktion *übersehen* bzw. *nicht erkennen*. Beispiele hierfür sind das Übersehen des Warenkorbs, das sechs Mal auftrat, sowie das Übersehen der Produktbewertungen, was fünf Mal auftrat. Bezogen auf die Gesamtzahl an Problemen aus dem Nutzertest, machen diese Probleme fast ein Viertel aus mit 10 von 45 Problemen, die durch Übersehen oder unzureichende Auffindbarkeit hervorgerufen werden.

Analysiert man die Ergebnisse der heuristischen Evaluation nach zugrundeliegender Heuristik, so fällt auf, dass vorwiegend Befunde zum „Informationsdesign“ gefunden werden.16 Probleme von den insgesamt 41 wurden durch eine Heuristik aus dieser Kategorie ermittelt, „Navigation“ und „Target Design“ liefern jeweils sechs beziehungsweise sieben Probleme.

Das Feedback der Experten zu den Heuristiken selbst ergab drei Ansätze zur weiteren Verbesserung. Zum einen wurde vorgeschlagen, die Anzahl an Regeln pro Kategorie zu reduzieren und zu vereinheitlichen auf maximal sieben. Daneben sollte der Umfang reduziert werden. Bisher konnte nur eine Regel zur Kürzung identifiziert werden: Regel 4.4 *„Use warm and harmonic colour schemes”*. Darüber hinaus sollten die wichtigsten inhaltstragenden Schlüsselwörter hervorgehoben werden, um die Regeln leichter lesen zu können.

4 Interpretation und Diskussion

Die Heuristiken erreichen in dieser Studie fast die gleichen Kennwerte wie der Nutzertest. Die Validität der gefundenen Probleme ist mit 0,93 sehr hoch, das heißt, es werden nur wenige *false positives* ermittelt. Auf der anderen Seite ist die Gründlichkeit mit 0,55 nicht so hoch. Das liegt daran, dass bei beiden Evaluationstechniken viele *unique problems* gefunden werden. Bei der Validierungsmethode mit vereinten Listen führt das zu niedrigen Werten in der Gründlichkeit. Insgesamt deuten die Erkenntnisse darauf hin, dass die Heuristiken eine valide Methode zur Evaluation darstellen. Für eine weitere Optimierung und Kürzung ist eine Evaluation mit einer höheren Nutzeranzahl geplant und die anschließende Bewertung von ermittelten Problemen aus Sicht der Nutzer.

Literaturverzeichnis

Aarhus, R., Grönvall, E., Kyng, M (2010). Challenges in participation: Users and their roles in the development of home-based Pervasive Healthcare applications. In the proceedings of Pervasive Health 2010, Munich, Germany, 22-25 March, 2010.

Anderson, M. & Perrin, A. (2018) Technology use among seniors. Retrieved from http://www.pewinternet.org/2017/05/17/technology-use-among-seniors/

Böhm, V., & Nguyen, A. (2017). Entwicklung eines Web-Usability- Heuristiken-Sets für die ältere Zielgruppe. In: Burghardt, M., Wimmer, R., Wolff, C. & Womser-Hacker, C. (Hrsg.), Mensch und Computer 2017- Workshopband. Regensburg: Gesellschaft für Informatik e.V.. https://doi.org/10.18420/muc2017-ws17-0408

Botella, F., Alarcon, E., & Peñalver, A. (2013). A new proposal for improving heuristic evaluation reports performed by novice evaluators. In Proceedings of the 2013 Chilean Conference on Human - Computer Interaction - ChileCHI '13 (S. 72–75). New York: ACM Press. https://doi.org/10.1145/2535597.2535601

Grönvall, E., and Kyng, M., (2011). Beyond Utopia: reflections on participatory design in home-based healthcare with weak users. *European Conference on Cognitive Ergonomics*, ACM, pp. 189–196.

Hartson, H. R., Andre, T. S., & Williges, R. C. (2003). for Evaluating usability evaluation methods. International Journal of Human-Computer Interaction, 15(1), 145–181. doi:10.1207/S15327590IJHC1501_13

Hermawati, S., & Lawson, G. (2016). Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus? Applied Ergonomics, 56, 34–51. https://doi.org/10.1016/j.apergo.2015.11.016

Ling, C., Salvendy, G. 2005. Extension of Heuristic Evaluation Method: a Review and Reappraisal. Ergonomia. An International Journal of Ergonomics and Human Factors 27 (3). 179-197. Retrieved from https://pdfs.semanticscholar.org/9cd1/916541b886da7f224ed8cfa630a1c05d29a2.pdf

Nielsen, J., & Molich, R. (1990). Heuristic Evaluation of User Interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (S. 249–256). New York: ACM. doi:10.1145/97243.97281

Rusu, C., Roncagliolo, S., Rusu, V. & Collazos, C. (2011). A Methodology to Establish Usability Heuristics. *The Fourth International Conference on Advances in Computer-Human Interactions.* Zuletzt abgerufen am 23.02.2018, unter https://www.researchgate.net/publication/229040164_A_ Methodology_ to_establish_usability_heuristics

Sears, A. (1997). Heuristic Walkthroughs: Finding the Problems without the Noise. International Journal of Human-Computer Interaction, 9(3), 213–234. https://doi.org/10.1207/s15327590ijhc0903_2