25.06.2018

Version 1.0

BZgA
Bundeszentrale
für
gesundheitliche
Aufklärung

Expertise Website-Evaluation

Übersicht über bestehende Evaluations-methoden und Entscheidungshilfe für die Evaluation bestehender sowie neu geschaffener Websites

Diese Expertise soll der BZgA eine Übersicht über typische Methoden der Website-Evaluation liefern. Dabei liegt der Fokus auf der begleitenden oder vergleichenden Evaluation bereits bestehender Websites. Dargestellt werden theoretische Hintergründe zur Evaluation im Allgemeinen und im spezifischen Kontext von Websites, eine Entscheidungshilfe zur Website-Evaluation sowie typische Evaluationsverfahren. Die Verfahrensdarstellung geht insbesondere auf Hintergründe, Gütekriterien, notwendige Stichprobenumfänge, Kosten sowie Vor- und Nachteile ein. Dies wird ergänzt um Praxisbeispiele, Exkurse, sowie einen Einblick in kritisch zu bewertende und innovative Verfahren.

Autor: PD Dr. Meinald T. Thielsch
unter Mitarbeit von Prof. Dr. Gerrit Hirschfeld

Expertise Website-Evaluation: Übersicht über bestehende Evaluationsmethoden und Entscheidungshilfe für die Evaluation bestehender sowie neu geschaffener Websites (Version 1.0).
Arbeitsbericht, Köln: Bundeszentrale für gesundheitliche Aufklärung (BZgA).
https://doi.org/10.17623/BZGA:225-EWE-1.0

Abkürzungen und Begriffsklärung

Cronbachs α

Cronbachs α zeigt an, wie gut die Items (Fragen) innerhalb einer Skala zusammenpassen („interne Konsistenz“); dies wird oft als Hinweis auf die ⇒ *Reliabilität* interpretiert. Cronbachs α kann einen Wert zwischen α = 0 und α = 1 annehmen. Werte von α = .70 gelten als noch akzeptabel für Analysen / Untersuchungen auf Gruppenebene, Werte von α = .80 als ausreichend und von α = .90 bis .95 als sehr gut. Werte von α > .95 sind kritisch zu sehen (die verschiedenen Items unterscheiden sich dann meist nicht mehr ausreichend voneinander). Längere Fragebögen haben meistens eine höhere Interne Konsistenz als kurze Fragebögen.

DIN EN ISO

DIN steht für Deutsches Institut für Normung e.V., EN für Europäische Norm und ISO für International Organization for Standardization. Eine DIN EN ISO Norm ist somit in Deutschland, Europa und weltweit anerkannt.

n

Anzahl der BeurteilerInnen / UmfrageteilnehmerInnen

Objektivität

Das Hauptgütekriterium Objektivität ist dann gegeben, wenn die Ergebnisse unabhängig von der Person sind, die eine Befragung *durchführt*, *auswertet* und *interpretiert*. Ein Computerfragebogen mit immer gleicher Instruktion und fester Auswertung ist folglich objektiver als ein freies Interview.

r

Korrelationskoeffizient, Maß für den Zusammenhang zwischen zwei Variablen. Korrelationen können Werte zwischen -1 und 1 annehmen; 0 bedeutet keinen Zusammenhang, ±1 eine perfekte proportionale Übereinstimmung. Zusammenhänge ab ca. r = .30 werden als mittelmäßig ausgeprägt angesehen, ab ca. r = .50 kann ein starker Zusammenhang angenommen werden.

Reliabilität

Reliabilität ist ein weiteres Hauptgütekriterium und befasst sich mit der reinen Messgenauigkeit. Meist wird hier Cronbachs α als Maß angegeben. Ferner lassen sich die Messwerte aus zwei verschiedenen Messzeitpunkten korrelieren um die zeitliche Stabilität eines Instrumentes zu bestimmen („Retest-Reliabilität“).

Validität

Das Hauptgütekriterium Validität fragt, ob ein Fragebogen oder Test wirklich das Merkmal misst, das gemessen werden soll. Validität ist schwieriger zu bestimmen als die Reliabilität. Im Idealfall kommen mehrere Verfahren zur Anwendung: Man vergleicht beispielsweise einen neuen Test mit vorhandenen Verfahren, die das gleiche messen (dann sollte es hoch korrelieren = *konvergente Validität*) oder die etwas ganz anderes messen (dann sollte es wenig bis gar nicht korrelieren = *divergente Validität*); man kann außerdem Expertenurteile oder andere vergleichbare Kriterien heranziehen (= *konkurrente Validität*); führt Experimente durch (= *experimentelle Validität*); testet die Fähigkeit des Instruments, zwischen verschiedenen Zielobjekten zu unterscheiden (= *diskriminante Validität*); oder versucht, die angenommenen Faktoren mittels einer konfirmatorischen Faktorenanalyse zu bestätigen (= *faktorielle Validität*).

Zentrale Konstrukte der Website-Evaluation

Inhalt

Die ISO-Norm DIN EN ISO 9241-151 (ISO, 2006b) definiert Website-Inhalt als die Zusammenstellung von Informationsobjekten, die in Form von Ton, Text oder Video präsentiert werden können. Der Inhalt zeichnet sich aber nicht nur durch objektivierbare Eigenschaften wie syntaktische Struktur, Fehlerfreiheit oder Optimierung für Suchmaschinen, sondern auch durch die subjektive Wahrnehmung der NutzerInnen aus (Thielsch & Hirschfeld, in press). Die Wahrnehmung von Webinhalten wiederum bedingt eine Vielzahl von Aspekten wie beispielsweise die Zufriedenheit der NutzerInnen, deren Präferenz oder Vertrauen in eine Website (für eine Übersicht siehe Thielsch & Hirschfeld, in press).

Usability

In der DIN EN ISO 9241-11 (ISO, 1998) ist Usability als Effektivität, Effizienz und Zufriedenheit beschrieben, mit der NutzerInnen mit einem System vorgegebene Ziele erreichen können. Websites sollten sich also leicht bedienen lassen und ermöglichen, dass NutzerInnen schnell die gewünschten Informationen finden können. Wie der Inhalt ist Usability einerseits objektivierbar (z.B. Ladegeschwindigkeit oder Linktiefe einer Website), hat andererseits aber ebenfalls eine subjektive Komponente (vgl. Hornbæk, 2006; Kurosu & Kashimura, 1995). Usability beeinflusst nicht nur die Zufriedenheit der NutzerInnen, sondern auch viele weitere Aspekte wie zum Beispiel die subjektive Wahrnehmung einer Organisation und das Vertrauen in die Website (bspw. Cober et al., 2003; Flavián, Guinalíu & Gurrea, 2006; Lee & Koubek, 2012). Usability wird im Deutschen auch als Brauchbarkeit, (Be)Nutzerfreundlichkeit oder Gebrauchstauglichkeit bezeichnet.

Ästhetik

Der Begriff Ästhetik bezieht sich auf die Schönheit des Website-Designs. Ästhetik ist in der Forschung zumeist als unmittelbare, angenehme und subjektive Wahrnehmung eines Webobjekts definiert, die wenig durch schlussfolgernde Prozesse beeinflusst ist (vgl. Moshagen & Thielsch, 2010). Da sie sehr schnell wahrgenommen werden kann, etwa innerhalb der ersten halben Sekunde der Website-Nutzung (vgl. Bölte et al., 2017), hat Ästhetik eine besondere Bedeutung für den Ersteindruck (Lindgaard et al., 2006; Thielsch & Hirschfeld, 2012; Tractinsky et al., 2006). Sie beeinflusst aber auch viele weitere Aspekte wie zum Beispiel Zufriedenheit und Vertrauen der NutzerInnen oder Wahrnehmung von Eigenschaften wie der Attraktivität einer Organisation (ein Überblick findet sich bei Moshagen & Thielsch, 2010).

User Experience

Dieser Begriff umfasst alle Erfahrungen der NutzerInnen bei der Interaktion mit einer Website. Damit schließt der Begriff Web User Experience die Wahrnehmung von Inhalt, Usability und Ästhetik einer Website ebenso ein wie auch die Erwartungen hinsichtlich eines zukünftigen Websitebesuchs oder die Bewertung einer früheren Nutzung (vgl. DIN EN ISO 9241-210; ISO, 2009). User Experience wird oftmals mit „UX“ abgekürzt.

Einführung und theoretischer Hintergrund

Einführung

Die Bundeszentrale für gesundheitliche Aufklärung (BZgA) betreibt als Bundesbehörde eine Vielzahl von Websites im Gesundheitsbereich. Die BZgA will aufklären, informieren und gesundheitsfördernde Lebensweisen in der Bevölkerung unterstützen. Zentrales Qualitätsmerkmal einer Website im Anwendungsbereich der BZgA ist daher die hohe inhaltliche Güte. Eine gelungene Website ergänzt diese durch ein zeitgemäßes und leicht zu bedienendes Websitedesign.

Erleben und die Reaktionen der NutzerInnen auf Web-Angebote sind jedoch ohne empirische Testung unklar. Nur begleitet durch maßgeschneiderte Evaluationen können Maßnahmen online optimal umgesetzt und damit eine bestmögliche Wirkung der Inhalte erreicht werden. Die Vielzahl verfügbarer Methoden zur Website-Evaluation, angefangen bei qualitativen Interviews über Expertenverfahren bis hin zu quantitativen Tests, ist unübersichtlich – gleichzeitig unterscheiden sich die verschiedenen Verfahren hinsichtlich ihrer psychometrischen Güte und Aussagekraft. Daneben liefern objektive Verfahren wie beispielsweise Zugriffzahlen einer Website weitere sehr wichtige Hinweise über deren Erfolg.

Aufgrund der Vielzahl möglicher Methoden und Instrumente herrscht hier Bedarf an einer strukturierten Entscheidungshilfe zur Durchführung optimaler Evaluationen. Die vorliegende Expertise soll der BZgA eine Übersicht über typische Methoden und eine gezielte Entscheidungshilfe für die Planung und Durchführung von Website-Evaluationen liefern. Dabei liegt der Fokus klar auf der begleitenden oder vergleichenden Evaluation bereits bestehender Websites. Nicht berücksichtigt ist Entwicklung, Informationsarchitektur und Designkonzeption für von Grund auf neu zu schaffende Websites – daher erfolgt keine Darstellung von Methoden wie agiler Entwicklung, Personas, Prototyping, Scribbles oder ähnlicher Verfahren.
Die vorliegende Expertise fokussiert auf eine Handreichung, die Projektverantwortlichen ermöglichen soll, weitgehend fertig entwickelte, neu geschaffene oder schon länger bestehende Websites zielführend zu evaluieren und notwendige Entscheidungen abzuleiten. Neben der übersichtlichen Darstellung verschiedener Methoden steht daher eine Entscheidungshilfe zur Findung der passenden Evaluationsmethode im Zentrum der Expertise.

Theoretischer Hintergrund

Evaluation bezeichnet die Bewertung eines Sachverhalts, Prozesses, Produkts oder Programms. Anders als in der Wissenschaft, bei der es um den Erkenntnisgewinn geht, sollen Evaluationen in der Praxis Entscheidungsprozesse unterstützen – und hierfür möglichst verlässliche Ergebnisse liefern. Evaluationen helfen Ist-Zustände zu beschreiben und zu bewerten. Nur wenn der gegenwärtige Ist-Zustand treffend beschrieben ist, wird klar ob und welche weiteren Maßnahmen notwendig sind. Ist diese Beschreibung falsch, so besteht die Gefahr, dass Maßnahmen auf falschen Vorannahmen basieren oder Projekte scheitern, da notwendige Schritte nicht erkannt und vollzogen werden.

Beispiel: Warum die Beschreibung des Ist-Zustandes so wichtig ist – oder: Michelle Obama und der Eisbergsalat (aus Gollwitzer & Jäger, 2014, S. 55)

Übergewicht ist ein wichtiges persönliches, aber auch gesellschaftlich relevantes Gesundheitsthema. Die damalige First Lady der USA, Michelle Obama, griff dieses auf als sie Anfang 2010 ihre Kampagne „Let's move" ins Leben rief. Ziel war dem Übergewicht bei Kindern entgegenzuwirken (https://letsmove.obamawhitehouse.archives.gov/). Ein Aspekt der Kampagne bestand darin, Kindern aus ärmeren Familien Zugang zu gesunden Lebensmitteln zu erleichtern. Man war davon ausgegangen, dass es in ärmeren Gegenden schlichtweg zu wenige Supermärkte mit einem entsprechenden Angebot an Obst und Gemüse gäbe. Auch eineinhalb Jahre nach dem Start der Kampagne vermutete Michelle Obama, dass „Menschen die für das Mittagessen ihrer Kinder einen Eisbergsalat [...] oder Obst kaufen wollen, hierzu erst zwei oder drei verschiedene Busse nehmen oder mit dem Taxi fahren müssen" (zitiert nach Gollwitzer & Jäger, 2014, S. 55). Diese Annahme klingt plausibel – ist aber falsch. Die Forschung zeigte, dass es in ärmeren Gegenden sogar mehr Supermarktfilialen gab (Lee, 2012). Mehr noch, Verfügbarkeit von Supermarktfilialen mit Frischobst- und Gemüsetheken und die Essgewohnheiten von (kalifornischen) Kindern standen nicht in Zusammenhang (An & Sturm, 2012). Die mangelnde Verfügbarkeit von Obst und Gemüse war somit kein relevanter Risikofaktor für Übergewicht bei Kindern.

Das Beispiel zeigt, wie zentral die valide Beschreibung des Ist-Zustandes ist. Inhaltlich zu unterscheiden ist dabei zwischen a) dem **Evaluationsgegenstand**, beispielsweise einer Gesundheitswebsite, und b) dem **Evaluationskriterium**. Für einen Evaluationsgegenstand sind oft eine Reihe von verschiedenen Evaluationskriterien denkbar (vgl. Gollwitzer & Jäger, 2014, S. 22); beispielsweise die Akzeptanz einer Online-Intervention, die Qualität einer web-basierten Maßnahme oder der Transfererfolg einer Website. Evaluationen können damit durch sehr verschiedene Zielsetzungen und Rahmenbedingungen geprägt sein. Diese haben einen Einfluss auf das zu verwendende Evaluationsmodell, Evaluationszeitpunkte und Durchführungsmodi. In dieser Expertise wird das Vorgehen in Hinblick auf den Evaluationsgegenstand Websites spezifiziert, mit dem Ziel den verantwortlichen Akteuren eine Entscheidungshilfe zur Findung optimaler Evaluationsdesigns an die Hand zu geben.

Phasen eines Evaluationsprozesses

Das allgemeine Vorgehen in einer Evaluation lässt sich in verschiedene Phasen unterteilen. Angelehnt an Balzer (2005) sollen im Folgenden sieben zentrale Stufen herausgestellt werden (siehe auch Gollwitzer & Jäger, 2014, S. 39):

1. **Klärung des Evaluationsbedarfs**: An erster Stelle steht die Frage, warum eine Evaluation durchgeführt werden soll. Die Festlegung einer klaren Fragestellung und konkreter Evaluationsziele ist entscheidend für den Projekterfolg. Daraus ergibt sich auch, welche Evaluationskriterien herangezogen werden sollten.

2. **Rahmenbedingungen der Evaluation:** Bevor die notwendige Studie zur Beantwortung der Evaluationsfrage angegangen werden kann, sind die Rahmenbedingungen der Evaluation zu klären. Dies umfasst die Information und entsprechende Einbindung der relevanten Entscheider und Beteiligten. Ein partizipatives Vorgehen in der Evaluationsplanung erhöht nicht nur die Akzeptanz auf Seiten der Beteiligten – sie kann auch zu höherwertigeren Studiendesigns führen, da die verschiedenen Sichtweisen auf den Evaluationsgegenstand frühzeitig gehört werden.

3. **Methodische Projektplanung:** Im nächsten Schritt ist konkret zu überlegen, welche Methoden notwendig sind, um die Evaluationsfrage zu beantworten. Die vorliegende Expertise soll insbesondere an dieser zentralen Stelle Informationen über Vorgehensweisen und mögliche Verfahren der Website-Evaluation, sowie deren Stärken und Schwächen geben. Je nach Projektressourcen mag es bereits an dieser Stelle sinnvoll sein einen externen Dienstleister einzubeziehen.

4. **Durchführung der Evaluation:** Hier erfolgt die Umsetzung der geplanten Studie. Pre-Tests und Vorabprüfungen hinsichtlich geltender Standards (siehe nächste Seite) empfehlen sich vor der eigentlichen Datenerhebung. Insbesondere in der Feldphase kommen dann oftmals externe Dienstleister zum Einsatz.

5. **Datenauswertung und Interpretation:** An dieser Stelle ist zu betrachten, ob die Evaluationsfrage anhand der erhobenen Daten vollständig beantwortet werden kann, welche Handlungsempfehlungen sich ergeben – und ob möglicherweise zusätzliche Maßnahmen notwendig sind.

6. **Präsentation und Dissemination der Ergebnisse:** Nur wenn die Ergebnisse einer Evaluation allen relevanten Beteiligten verständlich kommuniziert werden, können Entscheidungen und Konsequenzen optimal abgeleitet werden.

7. **Nutzung der Ergebnisse und abschließende Bewertung der Evaluation:** Eine Evaluation ohne Konsequenzen hat wenig Sinn, aufgezeigte Implikationen sollten zeitnah angegangen werden. Im seltenen Fall, dass die Evaluation keinerlei Handlungsbedarf aufzeigt, ist es an der Zeit den Website-Verantwortlichen für die herausragende Arbeit zu danken.
   Grundsätzlich steht am Ende der Evaluation auch die Frage, was gut gelaufen ist – und was sich im nächsten Evaluationsverfahren verbessern lässt.

Allgemeine Standards und Gütekriterien für Evaluationen

Bereits zu Beginn der 1980er Jahre wurden in den USA verbindliche Standards für Evaluationen von den entsprechenden Fachgesellschaften etabliert. Angelehnt an diese hat in Deutschland die DeGEval – Gesellschaft für Evaluation e.V. entsprechende Standards erstmalig in 2001 veröffentlicht, im Jahr 2016 erfolgte eine Revision (DeGEval, 2016). Hierbei wurden 25 Einzelstandards in vier Gruppen festgelegt (siehe Abbildung 1). Jeder Standard ist in Form eines Satzes formuliert, ergänzt um eine Begründung und Umsetzungshinweise. Bei der Planung und Durchführung einer Evaluation ist es sehr hilfreich sich an diesen Leitlinien zu orientieren und die eigene Arbeit entsprechend kritisch zu hinterfragen.

Nützlichkeit

N 1 Identifizierung der Beteiligten und Betroffenen
N 2 Klärung der Evaluationszwecke
N 3 Kompetenz und Glaubwürdigkeit des Evaluators/der Evaluatorin
N 4 Auswahl und Umfang der Informationen
N 5 Transparenz von Werthaltungen
N 6 Vollständigkeit und Klarheit der Berichterstattung
N 7 Rechtzeitigkeit der Evaluation
N 8 Nutzung und Nutzen der Evaluation

Fairness

F 1 Formale Vereinbarungen
F 2 Schutz individueller Rechte
F 3 Umfassende und faire Prüfung
F 4 Unparteiische Durchführung / Berichterstattung
F 5 Offenlegung von Ergebnissen und Berichten

Durchführbarkeit

D 1 Angemessene Verfahren
D 2 Diplomatisches Vorgehen
D 3 Effizienz von Evaluation

Genauigkeit

G 1 Beschreibung des Evaluationsgegenstandes
G 2 Kontextanalyse
G 3 Beschreibung von Zwecken und Vorgehen
G 4 Angabe von Informationsquellen
G 5 Valide und reliable Informationen
G 6 Systematische Fehlerprüfung
G 7 Angemessene Analyse qualitativer und quantitativer Informationen
G 8 Begründete Bewertungen und Schlussfolgerungen
G 9 Meta-Evaluation

Abbildung 1: Standards für Evaluation der DeGEval – Gesellschaft für Evaluation e.V.

Im Rahmen einer Evaluation kommen qualitative und quantitative Verfahren aus der empirischen Sozialforschung zum Einsatz. Diese sind anhand ihrer Gütekriterien zu bewerten. Die drei zentralen Gütekriterien sind Objektivität, Reliabilität und Validität (siehe bspw. Bühner, 2010; Moosbrugger & Kelava, 2012).

» Verfahren im Kontext Evaluation sind dann **objektiv**, wenn die resultierenden Ergebnisse unabhängig von irrelevanten Randbedingungen sind. Objektivität ist also insbesondere dann gegeben, wenn die Ergebnisse unabhängig von der Person sind, die die Untersuchung *durchführt, auswertet* und *interpretiert.* Ein Computerfragebogen mit immer gleicher Instruktion und fester Auswertung ist also objektiver als ein freies Interview – bei letzterem kann sich die Formulierung von Fragen oder Anweisungen je Interview leicht ändern, es gibt Interviewereffekte (z.B. Auftreten des Interviewers) und meist keine vollständig fixe Auswertungsanweisung.

» Das Gütekriterium der **Reliabilität** bezieht sich auf Messgenauigkeit der Merkmalserfassung. In Tests und Fragebögen prüft man die Reliabilität typischerweise zum einen in Hinblick auf den Zusammenhang (Korrelation) aller Items untereinander (die sogenannte interne Konsistenz), meist wird hier Cronbachs α als Maß angegeben. Zum anderen kann man die Messwerte aus zwei verschiedenen Messzeitpunkten miteinander korrelieren und so die Stabilität bestimmen. Die Korrelationen als Zusammenhangsmaß können zwischen 0 und 1 schwanken, 0 bedeutet keinen Zusammenhang, 1 wäre eine perfekte Übereinstimmung.
Zur Reliabilität im Bereich Website-Evaluation muss grundsätzlich einschränkend gesagt werden, dass Retest-Messungen für Verfahren bisher leider nur selten angewendet werden; Angaben zur Reliabilität beziehen sich daher fast immer auf die interne Konsistenz (Cronbachs α).

» Die **Validität** zeigt an, in welchem Ausmaß ein Verfahren ausschließlich das Merkmal erfasst, dass es erfassen soll. So sollte zum Beispiel ein valider Mathetest nicht von der Lesegeschwindigkeit oder der aktuellen Stimmung eines Schülers beeinflusst sein. Validität ist schwieriger zu bestimmen als die Reliabilität. Hier müssen mehrere Prüfungsmethoden gleichzeitig zur Anwendung kommen: Man vergleicht beispielsweise ein neues Instrument mit vorhandenen Verfahren, die das gleiche messen sollen (dann sind hohe Korrelationen zu erwarten), oder die etwas ganz Anderes messen sollen (dann sind geringe bis gar keine Korrelationen zu erwarten). Man zieht Expertenurteile zu Rate, führt Experimente durch, oder versucht die angenommenen Faktoren mittels einer konfirmatorischen Faktorenanalyse zu bestätigen.

Alle drei Gütekriterien hängen zusammen: Die Objektivität ist eine Voraussetzung für die Reliabilität und die Reliabilität wiederum eine Voraussetzung für die Validität. Im Umkehrschluss heißt dies auch, dass ein hoch valides Verfahren auch sehr reliabel ist.

Zusätzlich lassen sich bei der Bewertung eines Verfahrens verschiedene Nebengütekriterien betrachten, so zum Beispiel:

- **Normierung**: Bei normierten Verfahren liegen Vergleichswerte, zum Beispiel von anderen Testpersonen aus der Zielgruppe, vor.

- **Ökonomie**: Hier geht es schlicht um die Kosten einer Befragung oder Testung. Der Nutzen eines Verfahrens sollte auf jeden Fall höher sein als seine Kosten.

- **Zumutbarkeit**: Die Durchführung eines Evaluationsverfahrens kann für die Testpersonen belastend sein. Hier muss man sich fragen, ob die Anwendung des Verfahrens zumutbar ist und von den Befragten akzeptiert wird.

- **Unverfälschbarkeit**: Im besten Fall ist es für die Getesteten nicht möglich, Werte zu verfälschen und das Ergebnis zu beeinflussen.

Exkurs: Interpretation der zentralen Gütekriterien

Bei der Interpretation der psychometrischen Güte eines Evaluationsverfahrens ist an allererster Stelle die Validität des Verfahrens entscheidend. Ist ein Verfahren nicht valide, heißt das, dass das intendierte Merkmal nicht gemessen wird. Fehlen Angaben zur Validität, so ist dies äußerst kritisch zu sehen – hier ist die Qualität des Verfahrens schlicht unklar.
Zur Beurteilung der Höhe der zentralen Gütekriterien finden sich in der Literatur oftmals folgende Daumenregeln (vgl. Bühner, 2010):

Güterkriterium	niedrig	mittel	hoch
Objektivität (Übereinstimmung zweier Beurteiler)	< .60	.60 - .90	> .90
Reliabilität	< .80	.80 - .90	> .90
Validität (unkorrigierte Korrelation)	< . 40	.40 - .60	> .60

Diese Kennwerte können aber nur als grobe Richtschnur bei der Interpretation der statistischen Kennwerte gelten. Zudem gibt es für die Validität eines Verfahrens keinen einzelnen nummerischen Wert – Validität wird stets in Bezug auf eine Vielzahl von Untersuchungswegen argumentiert. Hier sollten zumindest Analysen zu konvergenten und divergenten Kriterien vorliegen, das heißt ein Verfahren würde dahingehend geprüft, ob es mit verwandten Messungen hoch und mit nicht verwandten niedrig korreliert. Zum Beispiel: Ein Additionstest sollte hoch mit einem anderen Mathetest korrelieren, aber niedrig mit einem Stimmungsfragebogen. Bei standardisierten quantitativen Verfahren (wie beispielsweise Fragebogenverfahren) sind heutzutage eine Vielzahl an weiteren Validierungsprüfungen üblich (u.a. eine statistische Bestimmung und Konfirmierung der Faktorstruktur). Die Gütekriterien eines Verfahren sollten zudem grundsätzlich an ausreichend großen Stichproben aus der Zielpopulation bestimmt worden sein. Die meisten Gütekriterien können dabei ab einer Stichprobengröße von ca. 250 Probanden hinreichend exakt geschätzt werden (vgl. Schönbrodt & Perugini, 2013).

Antworttendenzen und Urteilsfehler

Wie in allen sozialwissenschaftlichen Erhebungsformen muss auch in der Evaluation berücksichtigt werden, dass Antworttendenzen und Urteilsfehler bei der Befragung von Personen auftreten können (vgl. Döring & Bortz, 2016, S. 236 und 250f.). Derartige Abweichungen der Antworten von den wahren Sachverhalten entstehen oft unbewusst und ohne böse Absicht der Befragten. Sie können ihre Ursache in verschiedenen Faktoren haben, wie zum Beispiel dem individuellen Verständnis der Fragen, Erinnerungs- oder Wahrnehmungseffekten oder der sozialen Interaktion zwischen Interviewer und Befragtem. In der Forschung sind viele verschiedene Verzerrungseffekte bekannt, so unter anderem zum Beispiel:

- **Akquieszenz:** Ja-Sage-Tendenz der Befragten unabhängig vom Inhalt der Fragen.
- **Halo-Effekt:** Ein Merkmal überstrahlt andere, in der Website-Evaluation beeinflusst zum Beispiel die Ästhetik einer Website stark die Beurteilung der Usability (Thielsch, Engel & Hirschfeld, 2015).
- **Konsistenzeffekt:** Ähnliche klingende Aussagen werden so beantwortet, dass sie inhaltlich zueinander passen (auch wenn das nicht so einheitlich zutrifft).
- **Positionseffekt:** Je nach Position der Frage erfolgen unterschiedliche Antworten, zum Beispiel durch Verständnisprobleme am Beginn einer Befragung, Ermüdungseffekte am Ende oder Kontrasteffekten zwischen verschiedenen Beurteilungsgegenständen.
- **Soziale Erwünschtheit:** Die Tendenz, Fragen nicht nach der eigenen, persönlichen Sicht zu beantworten, sondern im Einklang mit geltenden sozialen Normen.
- **Tendenz zur Mitte:** Bei mehrstufigen Skalen (z. B. Likert-Skalen) werden systematisch eher die mittleren Skalenpunkte ausgewählt.
- **Tendenz zur Milde/Härte/zu extremen Urteilen:** Die Tendenz von Befragten bei mehrstufigen Antwortmöglichkeiten zu Extremen zu neigen.

Um solchen Fehlern entgegen zu wirken, werden Verfahren standardisiert, Bewertungsobjekte (bspw. verschiedene Websites) randomisiert und ausreichend große Stichproben erhoben, um die Wahrscheinlichkeit für Fehler zu minimieren. Daraus lässt sich erkennen, dass besonders qualitative Verfahren aufgrund ihrer weniger standardisierten Durchführung mit eher kleinen Stichproben anfällig für derartige Fehler sind. Zudem können hier auch weitere sogenannte Interviewer- oder Versuchsleitereffekte auftreten, beispielsweise bedingt durch Kleidung, Eigenschaften, oder Auftreten von Interviewern (Döring & Bortz, 2016, S. 246f.)

Allgemeine Lesetipps zum Thema Evaluation:

DeGEval – Gesellschaft für Evaluation e.V. (2016) (Hrsg.): *Standards für Evaluation: Erste Revision auf Basis der Fassung 2002*. Mainz: DeGEval – Gsellschaft für Evaluation e.V..
Verfügbar via http://www.degeval.de/degeval-standards/

Gollwitzer, M. & Jäger, R. S. (2014). Evaluation kompakt (2. Aufl.). Weinheim: Beltz.

Grundlagen der Website-Evaluation

Die Website-Evaluation hat sich aus der Evaluation von Softwareprodukten, einem Spezialthema der Mensch-Maschine Interaktion (siehe bspw. Gediga & Hamborg, 2002; Salaschek et al., 2007), entwickelt. In diesen Kontexten sind auch für den Bereich gültige ISO-Normen entstanden: Insbesondere verschiedene Teile der DIN EN ISO 9241 sind zentral für die Usability-Evaluation. In Teil 11 wird Usability definiert (ISO, 1998) in Teil 110 der Norm (ISO, 2006a) differenzierte Vorschläge zu Dialoggestaltung gemacht (siehe Appendix I). An diese lehnen sich Software-Evaluationsverfahren wie zum Beispiel der ISONORM-Fragebogen (Prümper, 1997) eng an.

Das World Wide Web ist heutzutage für viele Menschen Teil ihres Alltags – neun von zehn Bundesbürgern nutzen täglich das Internet (Koch & Frees 2017). Websites sind daher ein zentraler Evaluationsgegenstand. Sehr schnell und spontan treffen NutzerInnen eine Auswahl aus der Vielzahl an unterschiedlichen verfügbaren Online-Angeboten. Aus Sicht der NutzerInnen ist das subjektive Erleben einer Website zentral – die sogenannte User Experience. Dieser Begriff umfasst alle Erfahrungen des Users bei der Interaktion mit einer Website, was auch Aspekte der Usability und die Erwartungen hinsichtlich der zukünftigen Benutzung einschließt (vgl. ISO, 2009).

Sowohl in der Forschung als auch in der Praxis werden leider jedoch manchmal Evaluationsverfahren aus dem Bereich Software-Ergonomie oder Produktdesign direkt, ohne spezifische Validierung, auf Websites angewendet. Das World Wide Web unterscheidet sich aber deutlich von Software oder Produkten, insbesondere aufgrund der Funktion des Webs als Informationsmedium. DIN EN ISO 9241-151 (ISO, 2006b) geht hier spezifisch auf Besonderheiten von Web Interfaces aus einer Designperspektive ein. In dieser ISO-Norm wird ein Referenzmodell für die nutzerzentrierte Gestaltung von Websites gegeben, die Norm beinhaltet auch Hinweise zur Gestaltung der Navigation und der Web-Inhalte (siehe Appendix II). Die Rolle des Inhalts ist in diesem Kontext besonders hervorzuheben (vgl. Thielsch & Hirschfeld, in press).

Der Inhalt ist das zentrale Element einer Website – und wird von NutzerInnen als das wichtigste Kriterium für die Beurteilung genannt (Thielsch, Blotenberg & Jaron, 2014). Das World Wide Web ist ein sehr schnelllebiges Medium, in dem Organisationen einem enormen Wettbewerb ausgesetzt sind. Die Aufmerksamkeitsspanne von

Online-NutzerInnen ist enorm kurz, schnell wird über den Besuch eine Website entschieden und die durchschnittliche Verweildauer beträgt oftmals nur wenig mehr als eine Minute (Liu, White, & Dumais, 2010).

Im Vergleich zu klassischen Printmedien verlangt der Hypertext im Web höhere Lesefähigkeiten (Coiro, 2011), gleichzeitig sind die LeserInnen deutlich weniger an eine Quelle gebunden und können anhand leistungsstarker Suchmaschinen schnell und ohne Aufwand zu anderen themenverwandten Websites wechseln. Der Inhalt einer Website muss daher hinsichtlich der Auffindbarkeit in Suchmaschinen und optimal für Zielgruppe selbst aufbereitet sein. DIN EN ISO 9241-151 (ISO, 2006b) macht hier weitere Vorgaben zum Inhaltsdesign und betont unter anderem die Bedeutung von Angemessenheit, Struktur, Vollständigkeit und Zielgruppenanpassung von Web-Inhalten (siehe Appendix II). Grundsätzlich kann daher in bestimmten Situationen sinnvoll sein, bestehende Inhalte nicht nur ins Web zu übertragen, sondern eigens für die Webnutzung neu zu konzeptionieren. In Appendix III findet sich hierfür ein unterstützendes Evaluationsmodell, dass eigens für diesen Zweck zeitlich früher ansetzt als die in dieser Expertise vorrangig beschriebene Evaluation einer existenten Website.

Website-Evaluation widmet sich vor allem der Wahrnehmung und dem Verhalten von Website-BesucherInnen. Dabei ist es sehr hilfreich verschiedene Phasen, in denen NutzerInnen auf verschiedene Aspekte der Webseite reagieren, zu unterscheiden (siehe Abbildung 2): Der Ersteindruck einer Website wird vor allem durch die Ästhetik bestimmt. Diese wird sehr schnell in den ersten hundert Millisekunden der Nutzung wahrgenommen und bewertet (Bölte et al., 2017). Die Bewertung des Inhalts verlangt reflektierte kognitive Prozesse und dauert dementsprechend wahrscheinlich etwas länger. Ersteindrücke hinsichtlich der Glaubwürdigkeit einer Website können beispielsweise nach etwa drei bis vier Sekunden abgegeben werden (Robins & Holmes, 2008). Um die Usability einer Website sinnvoll bewerten zu können, ist eine echte Interaktion mit dieser und damit zusätzlicher Zeitaufwand notwendig (Thielsch et al., 2015). Die Forschung beinhaltet somit eine wichtige Implikation für einen Usability-Test einer Website: Dieser muss ein reales Nutzungsszenario abbilden.

Weiterhin ist zu beachten: Selbst, wenn Usability und Ästhetik einer Website optimal gestaltet sind, verwerfen NutzerInnen diese, wenn sie das Gefühl haben der Inhalt ist nicht hochwertig. Gleichzeitig kann in manchen Fällen ein hochwertiges Design über inhaltliche Mängel hinwegtäuschen (siehe Sillence et al., 2007).

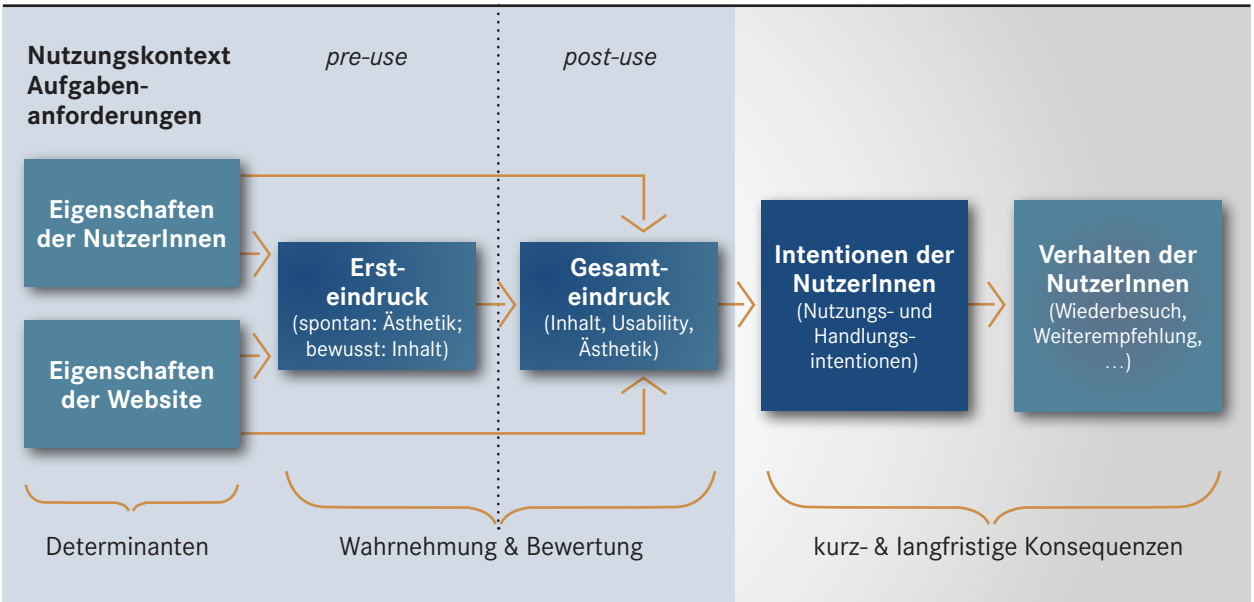


Abbildung 2: Phasen der Website-Nutzung und Bewertung (eigene Abbildung)

Zusammenfassend heißt das: Website-NutzerInnen werden durch eine hohe Ästhetik im Webdesign angezogen und dann mit guten Inhalten, die benutzerfreundlich dargeboten sind gebunden. Für zukünftige Nutzung und Weiterempfehlung ist die Wahrnehmung der Inhaltsqualität zentral, in geringerem Anteil kann dabei die erlebte Ästhetik weiterwirken (siehe Thielsch et al., 2014).

Spezifische methodische Faktoren in der Website-Evaluation

Die zu testende Website stellt den eigentlichen Evaluationsgegenstand dar. Generelle Eigenschaften einer Website sind in ISO 9241-151 (ISO, 2009) beschrieben. Hierzu zählen vor allem die verschiedenen Gestaltungselemente einer Website – insbesondere der Inhalt (definiert über Inhaltsobjekte wie Text, Bild und Multimedia) sowie die Inhaltspräsentation, die Gestaltung von Navigation, Suchfunktion und Anordnung der Inhalte, die sich aus der dahinterliegenden Informationsarchitektur ergibt. Einige dieser Inhaltseigenschaften lassen sich automatisch analysieren und objektiv überprüfen (bspw. Ladegeschwindigkeit oder Linkstruktur einer Website). Entscheidend ist jedoch, wie NutzerInnen auf die Website reagieren. Für eine Evaluation kann es hilfreich sein, Eigenschaften der Website selbst zunächst objektiv zu analysieren, diese

einem Inhaltsbereich zuzuordnen, und dann subjektive Befragungsdaten der NutzerInnen mit vorhandenen Benchmarks und/oder den automatischen Analyseergebnissen abzugleichen (siehe Thielsch & Hirschfeld, in press).

Dabei gibt es jedoch zentrale methodische Faktoren, die eine Website-Evaluation beeinflussen können und daher in der Planung und Durchführung zu berücksichtigen sind. Diese Faktoren lassen sich aus allgemeinen Modellen der Website-Nutzung, User Experience und Mensch-Computer Interaktion ableiten (bspw. Lee & Koubek, 2012; Thielsch, 2017; Thüring & Mahlke, 2007); im Kern sind dies vor allem drei grundlegende Faktoren, die zu beachten sind:

1. Eigenschaften der NutzerInnen
2. Art der Testaufgabe
3. Merkmale der Interaktion / der Testsituation

Bei der Nutzung einer Website sind zunächst Eigenschaften dieser selbst sowie die seiner NutzerInnen und des Kontextes relevant. Gemeinsam prägen sie die subjektive Wahrnehmung der Website, die aber auch von der Art der Testaufgaben und der Testsituation abhängig ist (vgl. Abbildung 2).

Eigenschaften der NutzerInnen

Da verschiedene Eigenschaften der NutzerInnen die User Experience beeinflussen, ist es wichtig, die Zielgruppe genau zu definieren. Empirische Studien finden sich zu Aspekten wie dem biologischen Alter (z.B. Sonderegger, Schmutz & Sauer, 2016; Thielsch, 2017), dem Geschlecht (z.B. Bardzell & Churchill, 2011; Tuch, Bargas-Avila & Opwis, 2010) oder der Persönlichkeit (z.B. Bosnjak, Galesic & Tuten, 2007; Thielsch, 2017). Ebenso können psychische Erkrankungen, wie beispielsweise eine Depression, die Wahrnehmung und Bewertung von Website-NutzerInnen verändern oder zum negativen beeinflussen (Thielsch & Thielsch, 2018). Je spezifischer hier die Anforderungen oder die Zielgruppen einer Website sind, desto mehr Augenmerk ist auf diese Aspekte zu legen. In der gegenwärtigen Forschung ergibt sich aber leider noch kein vollständiges Bild, welche Personeneigenschaften welche Wirkung auf die Website-Wahrnehmung haben. Als erste Erkenntnisse können genannt werden:

- Insbesondere hinsichtlich der Usability finden sich **Alterseffekte**, so zeigen Personen über 50 Jahren im Vergleich zu jüngeren Probanden sowohl eine schlechtere Performance in typischen Testaufgaben als auch subjektiv niedrigere Usability-Bewertungen (z. B. Chadwick-Dias, McNulty, & Tullis, 2003; Sonderegger et al., 2016; Wagner, Hassanein & Head, 2014). Hinsichtlich ästhetischer Präferenzen finden sich bisher keine besonders großen Alterseffekte, Studien kommen dabei teilweise zu widersprüchlichen Ergebnissen – hier bleibt weitere Forschung abzuwarten.

- Auch in typischen Website-Evaluationsinstrumenten wie z.B. Web-CLIC oder VisAWI (Thielsch und Hirschfeld, in press bzw. Moshagen & Thielsch, 2010) zeigen sich keine generellen Alters- oder **Geschlechtseffekte**. Dennoch werden besonders Geschlechtseffekte öfter betrachtet und durchaus gefunden (z.B. Cyr & Bonanni, 2005; Tuch et al., 2010). Die Systematik dieser bleibt jedoch unklar, auch, da viele hier nicht genannte Studien diesen Zusammenhang mit methodisch mangelhaften Studien auf zu kleinen Untersuchungsstichproben fußen.

- Bisher gefundene Zusammenhänge zwischen der **Persönlichkeit** der Web-NutzerInnen im Sinne der Big Five, waren zumeist eher klein (z.B. Bosnjak et al., 2007; Thielsch, 2017) und können daher derzeit in der Planung von Website-Evaluationen eher vernachlässigt werden.

- Generell scheinen sich **Kulturunterschiede** in der Wahrnehmung von Websites zu ergeben, insbesondere hinsichtlich der Einschätzung von Inhalten, Bildern und Farben (z.B. Cyr et al., 2009; Cyr, Head & Larios, 2010; Fletcher, 2006; Robbins & Stylianou, 2003; Zhao et al., 2003)

- Zu Web-NutzerInnen mit **psychischen Erkrankungen** finden sich bisher nur wenige Studien mit eher kleinen (negativen) Effekten von Erkrankungen wie der Depression auf die User Experience (bspw. Rotondi et al., 2007; Thielsch & Thielsch, 2018). Im Handlungsbereich der BZgA könnte dies aber je nach Zielgruppe ein wichtiger Hinweis auf eine notwendige, differenzierte Analyse der Bedürfnisse einer Website-Zielgruppe sein. Rotondi und Kollegen (2015) können beispielsweise zeigen, wie an Schizophrenie Erkrankte von einfachen und klaren Webdesigns profitieren und geben hier eine Reihe von Handlungsempfehlungen für diese spezielle Zielgruppe.

» Grundsätzlich bleibt zu bedenken, dass NutzerInnen Websites potenziell mit verschiedenen **Motiven** aufsuchen, bspw. zur Informationssuche, zu sozialem Austausch oder zu Entertainment-Zwecken (Go et al., 2016). Eine offene Forschungsfrage ist dabei, inwieweit die User Experience einer Website mit der Befriedigung von generellen persönlichen Bedürfnissen, so genannten „human needs" (vgl. Hassenzahl, Diefenbach & Göritz, 2010), zusammenhängt.

Insgesamt verdeutlichen diese ersten Forschungen: In der Planung einer Website-Evaluation ist es wichtig, die Zielgruppe genau zu definieren, zu überlegen welche spezifischen Eigenschaften diese hat, und wie dies die Ergebnisse der Evaluation möglicherweise systematisch beeinflussen könnte.

Art der Testaufgabe

Testaufgaben in einer Website-Evaluation sollen das typische Surfverhalten der NutzerInnen möglichst realistisch abbilden. Oft wird hier zwischen dem freien Explorieren einer Website und gezielten Suchaufgaben unterschieden (Dames et al., under review; Iten, Troendle & Opwis, in press). Die verschiedenen Aufgabentypen haben dabei insbesondere einen Effekt auf die Stärke des Einflusses von Website-Inhalt auf Wiederbesuchs- und Weiterempfehlungsintentionen (Dames et al., under review), sowie auf die Verweildauer auf der Website (längere Verweildauern bei gezielter Suche, siehe Iten et al., in press). In der Studie von Dames und Kollegen (under review) zeigt sich aber weder ein signifikanter Einfluss der Aufgabenart auf den Gesamteindruck, noch eine Veränderung im generellen Ergebnismuster in Abhängigkeit von der Aufgabe. Dementsprechend sollten in einer Website-Evaluation möglichst gut passende Aufgaben gegeben werden – unabhängig vom spezifischen Aufgabentyp.

Beispiele für typische Aufgabentypen in Website-Evaluationen (angepasst entnommen aus eigenen Studien des Autors):

» **Freies Explorieren**: „Wir möchten Sie nun bitten sich mit der Website [XYZ] vertraut zu machen. Bitte schauen Sie sich solange Sie wollen auf der Website um. Die Website kann komplett benutzt und abgesurft werden. Sie können beliebige Unterseiten dieser Website betrachten."

- **Gezielte Suche**: „Wir möchten Sie nun bitten die Website der Organisation [XYZ] zu besuchen und Antworten auf die folgenden Fragen zu finden: a) Was ist das Ziel dieser Organisation? b) Wo finden sich detaillierte Informationen zu Absicht und Zweck dieser Organisation? c) Wer ist zur Zeit Vorsitzende/r dieser Organisation?“
- **Gezielte Suche mit Testcharakter**: „Im Folgenden erwartet Sie ein Text zu einem medizinischen Thema. Es geht um Aphasien. Bitte lesen Sie diesen Text aufmerksam durch. Im Anschluss wird Ihr Textverständnis mithilfe einiger Fragen getestet. Bitte beantworten Sie diese so schnell und so richtig wie möglich. Ihre erreichte Punktzahl hat einen Einfluss auf die Verlosung am Ende der Studie.
  Fragen: a) Welches Brodmann-Areal ist bei der Broca-Aphasie betroffen? b) Wie bezeichnet man das Unvermögen zu lesen? c) Wie können Aphasien erworben werden?“
- Beispiel **Diagnosestellung** anhand von Web-Informationssystem: „Stellen Sie sich vor, Sie sind Arzt und ein Patient stellt sich mit folgenden Symptomen bei Ihnen vor: Der Patient gibt an, sich seit ca. 1 Woche müde und abgeschlagen zu fühlen und leichtes Fieber zu haben. Seit ca. 4 Tagen verspürt er heftige Schmerzen im Bereich des Brustkorbs. Vor zwei Tagen hat er mehrere Bläschen auf der Haut über seinem Brustkorb entdeckt. Welche Krankheit hat dieser Patient?“
- Beispiel **Erinnerungsfrage** nach Darbietung der Website: „Was war auf dem Bild, das rechts neben dem Artikel auf der Startseite abgebildet war, zu sehen? [das Gehirn eines Menschen] [ein Kreuz] [ein Stethoskop] [Nervenzellen] [ein Magnetresonanztomograf] [weiß nicht]“
- Beispiel **Wissensfrage** nach Darbietung der Website: „Welche Farbe betreffen Farbsinnstörungen am häufigsten? [gelb] [rot] [lila] [grün] [blau] [weiß nicht]“

Grundsätzlich sollten Testaufgaben in einer Website-Evaluation möglichst genau an Inhalte der zu evaluierenden Website und Zielgruppe angepasst sein.

Merkmale der Interaktion / der Testsituation

Hier sind zwei Aspekte zu bedenken: Zum einen wirken die anderen Faktoren (NutzerInnen, Aufgabe) zusammen, zum anderen kann die Art der Testung selbst die Daten beeinflussen. Eine Website-Evaluation im Labor kann zu anderen Ergebnissen als eine Studie im Feld oder eine Online-Befragung führen (vgl. Sauer et al., under review). Die Forschungsergebnisse hierzu sind gemischt, die generelle Empfehlung ist aber, innerhalb der Evaluationsstudie die reale Nutzungssituation möglichst nah abzubilden (Sauer

et al., under review). Das heißt auch, dass möglichst funktionierende Versionen einer Website und nicht nur Screenshots getestet werden sollten. Allein mit Screenshots lassen sich Funktionen und typische Navigation einer Website nur schwer abbilden, zudem tendieren Probanden bei Screenshot-Bewertungen dazu, vorrangig die Ästhetik zu bewerten (möglicherweise aufgrund von Halo-Effekten, siehe Thielsch et al., 2015). Damit ist es allein mit Screenshots sehr schwierig Faktoren wie die Usability zu erfassen.

In Website-Evaluationen sind folgende Online-Datenerhebungsformen typisch:

- **Panel-Befragung**: Die Website wird von einem Testkollektiv in einem Befragungspanel evaluiert. Ein Panel besteht aus Personen, die bereit sind regelmäßig an Umfragen teilzunehmen. Die Nutzung von Befragungspanels ist ein üblicher Weg in der Markt- und Meinungsforschung um Stichproben zu generieren.
- **On-site-Befragung**: Personen, die die Ziel-Website nutzen, werden direkt auf dieser gebeten an der Umfrage teilzunehmen.
- **Convenience-Sampling**: Personen aus der Zielgruppe werden nach Verfügbarkeit zufällig eingeladen, beispielsweise über Selbsthilfegruppen und entsprechende Online-Foren.

Es finden sich aber auch Mischformen oder Offline-Datenerhebungen wie beispielsweise:

- Eine Verhaltensbeobachtung im **Labor** (siehe Verfahrensdarstellung unten).
- **Remote-Tests**, bei denen Personen direkt zu Hause befragt werden, entweder über ein Webinterface oder via Telefon.
- In **mixed-mode Erhebungen** kommen verschiedenen Datenerhebungsformen gleichzeitig zum Einsatz. So wird beispielsweise eine Online-Panelumfrage mit einer Telefonumfrage ergänzt.

Allgemeine Lesetipps zum Thema Usability und User Experience Evaluation:

Jacobsen, J. & Meyer, L. (2017). *Praxisbuch Usability und UX*. Bonn: Rheinwerk Verlag.

Sarodnick, F., & Brau, H. (2015). *Methoden der Usability Evaluation: Wissenschaftliche Grundlagen und praktische Anwendung.* Göttingen: Hogrefe.

Website-Evaluation

Entscheidungshilfe zur Website-Evaluation

Auf Basis der zuvor dargestellten Phasen der Evaluation (siehe Seite 11f.) wird im Folgenden anhand von Leitfragen eine Entscheidungshilfe zur Planung und Durchführung von Website-Evaluationen im Handlungsbereich der BZgA vorgeschlagen. Diese ordnet sich im zeitlichen Verlauf eines Evaluationsprojekts. Übergeordnet sollten sich die Projektverantwortlichen drei grundsätzliche Fragen stellen:

1. **Wann anfangen?**
   Evaluationen sollten möglichst früh projektbegleitend eingesetzt werden. Im Optimalfall wird eine Website neu auf Basis einer Anforderungsanalyse geschaffen und bereits dieser Designprozess evaluativ begleitet. Werden Fehler und Probleme erst nach Fertigstellung bewusst, kann dies zeitaufwändige und teure Änderungen nach sich ziehen. In den Projekten und Kampagnen der BZgA sind aber viele Websites bereits erstellt, daher fokussiert diese Expertise auf die Evaluation des Ist-Zustandes bei Websites, die bereits vorhanden sind. Auch hier lohnen sich aber frühzeitige Evaluationen, nicht nur um festzustellen wie es um die Qualität einer Website steht, sondern auch wann und in welchem Ausmaß möglicherweise Veränderungen und Updates notwendig sind.

2. **Wird externe Unterstützung benötigt?**
   Eine wichtige Frage betrifft, wer die einzelnen Schritte der Evaluation durchführen soll. Dies wird in Phase 2 (Rahmenbedingungen der Evaluation) expliziert. Auch wenn externe Dienstleister eingebunden werden sollen, ist es wichtig, sich weiterhin an den sieben generellen Phasen der Evaluation zu orientieren. Diese können zudem genutzt werden, um genau festzulegen, welche Aufgaben von externen und welche von internen Dienstleistern erbracht werden sollen. In der Diskussion sind aktuell insbesondere die Stichprobenkosten von Felddienstleitern – hier hat in den vergangenen Jahren ein Preisverfall stattgefunden, leider auch begleitet durch vereinzelte Betrugsfälle und Datenfälschungen. An dieser Stelle sollte auf eine angemessene Incentivierung der Befragten geachtet werden, um die Qualität und Verlässlichkeit der Daten zu fördern. Bei der methodischen Planung der Evaluationsstudie ist es weiterhin wichtig, dass die Gütekriterien der einzelnen

Verfahren tatsächlich beachtet werden. Das heißt, es werden nur solche Verfahren verwendet, die eine akzeptable Reliabilität und Validität aufweisen (siehe Exkurs zur Interpretation von Gütekriterien auf Seite 16).

3. **Wie wird dokumentiert?**
   Generell ist darauf zu achten, dass alle Arbeitsschritte transparent und nachvollziehbar dokumentiert sind. Eingesetzte Fragebögen sollten ebenso wie Rohdaten und Analyseskripte zur Verfügung gestellt werden. Grundsätzlich ist es sinnvoll in Phase 7 den Evaluationsprozess selbst zu reflektieren. Das bedeutet hinsichtlich externer Dienstleister zu klären, ob der Anbieter in zukünftigen Prozessen erneut eingesetzt werden sollte oder nicht.

Phase 1: Klärung des Evaluationsbedarfs

a) Was ist die Fragestellung der Evaluation, welche Entscheidung steht an?

b) Welche Evaluationsziele werden verfolgt

- Zustandsevaluation (bspw. Stärken-Schwächen-Analyse einer Website, Vergleich mit Benchmarks oder anderen Anbietern im Feld)
- Veränderungsevaluation (bspw. Vergleich mit einer früheren Websiteversion, Monitoring der existierenden Website im Zeitverlauf, Abklärung Bedarf für Neuentwicklung oder Relaunch einer Website)
- Wirksamkeitsevaluation (bspw. zu Folgen und Konsequenzen der Websitenutzung)

c) Welches Evaluationskriterium soll / welche Evaluationskriterien sollen im Rahmen der Website-Evaluation betrachtet werden?

- Qualität der Website (bspw. Ersteindruck, Gesamteindruck, inhaltliche Qualität, technische Qualität und Usability, Designästhetik)
  ⇒ a) hinsichtlich des Ist-Zustandes
  ⇒ b) hinsichtlich des Änderungsbedarfs
- Akzeptanz der Website in der Zielgruppe (bspw. Nutzungsintentionen, Nutzung)
- Wirksamkeit der Website und Transfererfolg (bspw. Verständnis vermittelter Informationen, Verhaltensanpassungen der NutzerInnen)

Phase 2: Rahmenbedingungen der Evaluation

- Wer ist für die Evaluation verantwortlich?
- Welche Entscheidungsträger sind in die Projektplanung einzubinden?
- Welche Mittel und Ressourcen stehen zur Verfügung?
- Wer führt die Evaluation durch (und an welchen Stellen ist externe Unterstützung notwendig):
  - Wer übernimmt die Gessamtkoordination (inkl. Debriefing nach Projektabschluss)?
  - Wer ist für die methodische Planung zuständig?
  - Wer ist für die Datenerhebung zuständig?
  - Wer vollzieht die Datenauswertung und -interpretation?
  - Wer führt die Präsentation und Dissemination der Ergebnisse durch?
  - Wer setzt Konsequenzen aus der Evaluation um?
- Welche weiteren Beteiligten, Betroffenen und Akteure müssen berücksichtigt werden?
- Was ist der allgemeine Zeitplan und was die generellen Meilensteine des Evaluationsprojektes?

Phase 3: Methodische Projektplanung

- Welche Version(en) der Website soll getestet werden? Zu welchem Zeitpunkt/ welchen Zeitpunkten soll getestet werden?
  ⇒ Entscheidung: formative vs. summative Evaluation
  (eine formative Evaluation wird bereits während der Entwicklung oder Umgestaltun einer Website eingesetzt und erlaubt Korrekturen auf Basis von Zwischenergebnissen im laufenden Prozess, eine summative Website-Evaluation hat das Ziel einer abschließenden Bewertung einer fertig gestellten Website)
- Welche Aspekte der Website sollen getestet werden?
  ⇒ Welche(s) Verfahren (siehe Seite 35ff.) sind geeignet um die Evaluationsfrage unter den gegebenen Rahmenbedingungen optimal zu beantworten? Wie lassen sich die Evaluationskriterien methodisch optimal erfassen? Welche Teilaspekte sollen erfasst werden? Welche Verfahren sollen tatsächlich zum Einsatz kommen?

- Welche Zielgruppe soll getestet werden?
  - Was für demographische Eigenschaften hat die Zielgruppe?
  - Was ist die typische Mediennutzung in der Zielgruppe?
  - Wie kann ich die Zielgruppe im Rahmen der Evaluation am besten erreichen (z.B. Befragungspanel, On-site, etc.)?
  - Gibt es für die Evaluation relevante Einschränkungen in der körperlichen oder psychischen Verfassung der Zielgruppe?
  - Welche weiteren besonderen Anforderungen hat die Zielgruppe?
  - Wie stelle ich eine hohe Datenqualität der Befragung dieser Zielgruppe sicher (z.B. durch angemessene Incentives, versierte Felddienstleister, eigene Datenerhebung)?
- Sollen Testaufgaben zum Einsatz kommen? Wenn ja, welche? (bspw. freies Explorieren der Website, gezielte Suchaufgaben, Lernaufgaben, etc.)?
- In welcher Form soll die Evaluationsstudie durchgeführt werden?
  ⇒ Falls die gewählten Verfahren verschiedene Erhebungsmöglichkeiten offen lassen: Entscheidung für Labor vs. Online/Remote-Test vs. Feldstudie vs. mixed-mode Erhebung
  ⇒ Ist eine spezifische Betrachtung bestimmter Endgeräte notwendig (bspw. stationärer Computer vs. Tablet vs. Smartphone)?

Phase 4: Durchführung der Evaluation

- Sind vor dem eigentlichen Studienstart alle notwendigen Vorarbeiten abgeschlossen? Das heißt:
  - **1. Verpflichtende** Durchführung von Pre-Tests der Studie
    - a) mit den beteiligen Verantwortlichen/Experten und
    - b) mit mindestens 3-5 Personen aus der Zielgruppe.

    Ziel: Fehler identifizieren und Studie final optimieren.
  - **2. Kontrolle**, ob die Studie die üblichen Standesregeln und ethischen Vorgaben erfüllt, darunter insbesondere die Evaluationsstandards der DeGEval (siehe http://www.degeval.de/degeval-standards/) sowie gegebenenfalls weitere notwendige Verfahrensspezifische Standards (z.B. für Online-Verfahren siehe http://www.dgof.de/standesregeln/).
- Direkte Kontrolle bei Studienstart mit den ersten Datenerhebungen: Inwieweit funktioniert die Studie wie geplant?

- Kontrolle, wie die Studie sich im Feld entwickelt: Wird die notwendige Beteiligung erreicht, gibt es unerwartete Probleme, ist eine Nachsteuerung nötig? (ggf. diese Aufgabe an Felddienstleister delegieren).
- Sind bis hierhin alle eingesetzten Methoden ausreichend dokumentiert? Im Idealfall erfolgt die Dokumentation parallel zum Feld.

Phase 5: Datenauswertung und Interpretation

- Sind die Rohdaten auf Fehler geprüft worden? Sind die Daten realistisch und vollständig?
  ⇒ Das heißt zum Beispiel:
  - Demographische und andere quantitative Angaben sind in einem realistischen Range
  - Antwortzeiten und Bearbeitungsdauern für Aufgaben sind realistisch
  - In Online-Befragungen sind „Durchklicker“ ausgeschlossen worden
  - Offene Angaben wurden geprüft inwieweit diese Anhaltspunkte für Fehler geben
  - Das Muster der Antwortverteilungen über verschiedene Verfahren ist konsistent
- Falls dies zugesichert war: Sind alle Daten entsprechend anonymisiert? Gibt es Personen, die sich zum Beispiel in offenen Textfeldern aus Versehen selbst deanonymisieren (solche Daten sind zu löschen)?
- Sind alle Auswertungsverfahren angemessen?
- Wurde die Evaluationsfrage vollständig beantwortet? Was sind die Kernaussagen der Evaluationsstudie? Sind weitere Datenerhebungen notwendig?
- Welche Implikationen und Handlungsempfehlungen leiten sich ab?
- Sind Verarbeitungsschritte und Analyse vollständig dokumentiert (inkl. Rohdaten und Auswertungsskripten)?

Phase 6: Präsentation und Dissemination der Ergebnisse

- Wurden die Ergebnisse allen relevanten Beteiligten (siehe, Phase 2) kommuniziert?
- Wurde allen relevanten Beteiligten für ihren Einsatz gedankt?
- Herrscht Einigkeit über Entscheidungen und Konsequenzen aus der Evaluation?

Phase 7: Nutzung der Ergebnisse und abschließende Bewertung der Evaluation

- Wird die Umsetzung der Konsequenzen sinnvoll durchgeführt? Werden Änderungen angegangen bzw. wurden positive Befunde entsprechend belobigt?
- Wurde ein Debriefing zum Evaluationsverfahren selbst durchgeführt (Was ist gut gelaufen? Was kann bei der nächsten Evaluation verbessert werden?)
- Ist die Fragestellung beantwortet worden bzw. sollte ein weiterer Evaluationszyklus angestoßen werden?

Darstellung der Verfahren zur Website-Evaluation

Die Tabelle 1 gibt einen komprimierten Überblick zu den nachfolgend dargestellten Verfahren der Website-Evaluation. Informationen im Detail finden sich bei den Darstellungen der einzelnen Verfahren auf den anschließenden Seiten.

		Evaluationsziel			Evaluationskriterien				Zeitpunkt		typische Kosten	Zeit-aufwand	Erfüllung Messgütekriterien
					Website-Qualität		Website-Inhalte						
		Zustand	Verände-rung	Wirksam-keit	Ist-Zustand	Änderungs-bedarf	Akzep-tanz	Wirksam-keit	formativ	summativ	in €		
qualitative Verfahren	Qualitative Inspektionsmethoden	**	***		***	***	*		***	*	5.000 – 6.000	niedrig bis mittel	schlecht bis mittel
	Qual. Tiefeninterviews mit NutzerInnen	*	**	*	***	**	*	*	***	*	6.000 – 12.000	mittel	schlecht bis mittel
	Fokusgruppen	*	**	*	***	**	*	*	***	*	10.000 – 12.000	hoch	schlecht
quantitative Verfahren	Checklisten und Guidelines	*	*		*	*			***	**	eigene Personalkosten	niedrig	schlecht bis mittel
	Standardisierte Fragebogenverfahren	***	**	**	***	**	***	***	*	***	7.000 – 12.000	mittel	gut bis sehr gut
	Verhaltensbeobachtung im Labor	**	**	*	***	***	*	*	***	**	12.000 – 16.000	hoch	mittel bis gut
	Verhaltensbeobachtung im Feld: Logfile-Analysen	*	**		*	*				**	Installations-/ Serverkosten	niedrig	gut

Tabelle 1: Übersicht über die nachfolgend dargestellten Verfahren.
*** = geeignet, ** = gut geeignet, *** = sehr gut geeignet; Typische Kosten: Bei den hier genannten Werten handelt es sich um übliche Verfahrenskosten. Diese Marktdaten wurden auf Basis der bei den einzelnen Verfahren nachfolgend jeweils angegeben Quellen sowie einer Abfrage über mindestens zwei unabhängige Akteure im Markt (im Februar 2018) geschätzt und können sich von tatsächlich anfallenden Kosten unterscheiden. Details zur Kostenabschätzung (z.B. zur Stichprobengröße) finden sich in den jeweiligen nachfolgenden Verfahrensdarstellungen und müssen entsprechend berücksichtigt werden.**

Kombination verschiedener Verfahren zur Website-Evaluation

In einer Website-Evaluation können verschiedene Verfahren sinnvoll kombiniert werden. Dies erhöht die Menge der verfügbaren Informationen und erlaubt eine reliablere Bewertung einer Website. Tabelle 2 gibt einen komprimierten Überblick darüber, welche Verfahren sich in formativen beziehungsweise summativen Evaluation kombinieren lassen.

		qualitative Verfahren			*quantitative Verfahren*		
		Qualitative Inspektionsmethoden	Qual. Tiefeninterviews mit NutzerInnen	Fokusgruppen	Checklisten und Guidelines	Standardisierte Fragebogenverfahren	Verhaltensbeobachtung im Labor
qualitative Verfahren	Qualitative Inspektionsmethoden		***	***	***	*	*
	Qual. Tiefeninterviews mit NutzerInnen	*		*	***	**	***
	Fokusgruppen	*	–		***	*	–
quantitative Verfahren	Checklisten und Guidelines	*	*	*		–	*
	Standardisierte Fragebogenverfahren	***	**	**	***		**
	Verhaltensbeobachtung im Labor	–	***	–	**	***	
	Verhaltensbeobachtung im Feld: Logfile-Analysen	*	*	*	*	***	***

Tabelle 2: Kombination verschiedener Verfahren in einer formativen (obere Diagonale, blau) bzw. summativen Website-Evaluation (untere Diagonale, grau).
*** = geeignete, ** = gut geeignete, *** = sehr gut geeignete Verfahrenskombination.**
Diese Einschätzung beruht weitgehend auf versuchspraktischen Erwägungen, den resultierenden Daten der jeweiligen Verfahren und dem Aufwand der jeweiligen Verfahren.
Logfile-Analysen sind aus der Darstellung der formativen Evaluation ausgenommen, da diese hier in der Regel nicht sinnvoll durchgeführt werden können.

Qualitative Inspektionsmethoden

Hintergrund

Verfahrenstyp: ***Qualitativ Test aus Expertensicht***
Zeitpunkt: ***formativ***
Zeitaufwand: ***niedrig bis mittel***
Psychometrische Güte: ***gering bis mittel***

Qualitative Inspektionsmethoden sind vor allem in der formativen Evaluation sinnvoll. Zu diesen Inspektionsmethoden zählen Methoden wie *heuristische Evaluationen*, *Experten-Evaluationen* oder so genannte *Walkthrough-Ansätze*. Gemeinsam ist diesen Verfahren, dass in der Regel mehrere Evaluatoren mit hoher Expertise eine Website explorieren. Die Basis der Evaluation ist dabei die Orientierung an vorhandenen Design-Prinzipien, Heuristiken, Richtlinien wie den relevanten ISO-Normen (siehe bspw. Appendix I und II), und der Erfahrungsschatz der beteiligten Experten. Bei Walkthrough-Verfahren wie dem Cognitive Walkthrough evaluiert eine Gruppe von Experten eine (oftmals noch nicht fertige) Website anhand von typischen Nutzungsszenarien und bewertet die notwendigen Handlungsschritte der NutzerInnen (siehe nachfolgendes Praxisbeispiel). Die beteiligten Experten versetzen sich bei all diesen Verfahren in die Rolle der potenziellen NutzerInnen. Hierfür ist sowohl eine domainenspezifische als auch methodische Expertise von Vorteil (vgl. Sarodnick & Brau, 2015; S. 144f.). Das heißt eingesetzte Experten besitzen neben fundierten Kenntnissen im Bereich Website-Evaluation und User Experience beispielsweise zudem eine fachliche Qualifikation im Gesundheitsbereich.

Gütekriterien

Prüfungen der klassischen Gütekriterien für diese verschiedenen qualitativen Inspektionsmethoden werden eher selten vorgenommen, sodass die psychometrische Qualität und insbesondere die Validität der Verfahren oftmals unklar ist (vgl. Mahatody, Sagar & Kolski, 2010; Vermeeren et al. 2010). Zentral ist das Kriterium, wie viele und welche Probleme mit diesen Verfahren identifiziert werden können (vgl. Liljegren, 2006). Ein typisches Ergebnis ist, dass Expertenevaluationen teilweise andere Prob-

leme finden als Nutzertests (z.B. Tan, Liu & Bishu, 2009). Daher wird empfohlen Expertenverfahren mit Nutzertests zu kombinieren (z.B. Jaspers, 2009; Tan et al., 2009) oder letztere im Zweifelsfall vorzuziehen (Liljegren, 2006). Expertenbasierte Verfahren finden sich meist zu frühen Zeitpunkten im Evaluationsverfahren (oft bereits schon in der Kreationsphase einer Website), um einen ersten Eindruck zu gewinnen. Zur Erfassung der Anwendersicht wird dann nachlaufend mindestens ein Verfahren mit Nutzerbefragungen ergänzt (siehe Tabelle 2).

Notwendige Stichprobenumfänge

Expertenbasierte Verfahren werden typischerweise mit wenigen Personen durchgeführt. Aufgrund der mangelnden Erkenntnislage zu den Gütekriterien der Verfahren ist ein Optimum an Personen schwer abzuschätzen. Dieses wird zudem durch Erfahrungsschatz und Qualifikation der spezifisch eingesetzten Experten beeinflusst. Wichtig ist hier zur Erhöhung der Objektivität der Daten nicht nur mit einem einzelnen, sondern mit mehreren Experten zu testen. In der Praxis kommen oft mindestens zwei Experten (im Sinne eines 4-Augen-Prinzips) zum Einsatz.

Kosten

Die Kosten ergeben sich aus den notwendigen Personalkosten. Für externe Experten im Bereich Usability / User Experience kann hier der jährliche Branchenreport der GermanUPA einen Ansatzpunkt für übliche Vergütungen liefern (http://www.germanupa.de/berufsfeld-usabilityux-professionals/branchenreport): In 2017 lagen durchschnittliche Stundensätze für Selbstständige bei 82 €, der mittlere Tagessatz bei 600 € (Tretter et al., 2017). Eine typische Experteninspektion/Walkthrough kostet, inklusive Vorbereitung und Berichtserstellung, insgesamt derzeit ca. 5000-6000 € (Stand: Februar 2018).

Wichtig ist hier aber zwischen den Tagessätzen für den Einsatz von User Experience Experten zum einen und Tagessätzen von Beratern zum anderen zu unterscheiden. Berater haben deutlich höhere Tagessätze, sollten dementsprechend aber weitreichendere Empfehlungen und konzeptionelle Beiträge leisten. Ein User Experience Experte hingegen wendet in der Regel vorhandene Methoden an und wertet diese Ergebnisse

aus, führt aber keine weitreichende Beratung im Rahmen der im Branchenreport der GermanUPA genannten Tagessätze aus.

Bewertung Vor-/Nachteile

Vorteile: Qualitative Expertenansätze sind in frühen Entwicklungsstadien hilfreich, um grundsätzliche Probleme einer Website aufzudecken.
Nachteile: Auf Basis der aktuell verfügbaren Literatur scheinen eindeutig Nutzertests in Ergänzung zu den qualitativen Inspektionsmethoden notwendig zu sein. Nur so kann das Erleben der Zielgruppe der Website umfassend und valide erfasst werden.

Weiterführende Informationen

Eine Darstellung verschiedener Verfahren findet sich bei Sarodnick und Brau (2015; S. 142f.).

Praxisbeispiel: Phasen eines Cognitive Walkthrough

Ein Cognitive Walkthrough ist eine experten-basierte Inspektionsmethode. Ziel der Expertenanalyse ist herauszustellen, inwieweit unerfahrene NutzerInnen vermutlich ein System erlernen und nutzen können. Dabei werden bestimmte Kenntnisse und Fähigkeiten der NutzerInnen angenommen und deren erwartete Interaktionen und Handlungsabfolgen mit einem System (wie einer Website) analysiert. Der Cognitive Walkthrough gliedert sich typischerweise in zwei Phasen (vgl. Sarodnick und Brau, 2015; S. 153f.):

1. **Vorbereitungsphase:** Die Experten bereiten die Grundlagen der Evaluation vor, indem die folgenden Aspekte festgelegt werden:
   - Prototypisches NutzerInnenprofil (Vorkenntnisse, Demographie, Bildung etc.)
   - Auswahl von bedeutsamen und realistischen Aufgaben zu Kernfunktionen der Ziel-Website
   - Festlegung von idealen Handlungsabfolgen zur Bewältigung dieser Aufgaben
   - Falls noch kein Website-Prototyp vorhanden ist: Festlegung der Schnittstelle – Beschreibung dessen, welche Interfaces die NutzerInnen im jeweiligen Handlungsabschnitt zu sehen bekommen

2. **Analysephase**: Die Experten sprechen alle Handlungsschritte zu jeder Aufgabe durch und formulieren plausible Szenarien wie NutzerInnen im jeweiligen Kontext agieren würden.
   Dies kann anhand von vier **Leitfragen** erfolgen:
   a) Werden die NutzerInnen versuchen den gewünschten Effekt zu erzielen?
   b) Werden die NutzerInnen erkennen, dass die korrekte Handlung ausgeführt werden kann?
   c) Werden die NutzerInnen erkennen, dass die korrekte Handlung zum gewünschten Effekt führen wird?
   d) Werden die NutzerInnen den Fortschritt erkennen, wenn sie die gewünschte Handlung ausgeführt haben?

   Beispiel: Wenn auf einer Website beispielsweise Informationsmaterialien angefordert werden können, sollten die NutzerInnen
   a) die Materialien direkt finden und auswählen, dann
   b) das System der Adresseingabe verstehen und
   c) richtig nutzen können sowie
   d) eine Rückmeldung über die erfolgreiche Bestellung erhalten.

Qualitative Tiefeninterviews mit NutzerInnen

Hintergrund

Verfahrenstyp: ***Qualitativ***
Test aus Nutzersicht
Zeitpunkt: ***formativ***
Zeitaufwand: ***mittel***
Psychometrische Güte:
gering bis mittel

Qualitative Einzelinterviewverfahren, auch als Tiefeninterviews bezeichnet, sind sehr flexibel in ihrer Anwendung. Diese können Face-to-Face oder auch via Telefon, Skype oder ähnliche Technologien stattfinden. Als qualitatives Verfahren erlauben sie einen Informationsgewinn durch direktes Nachfragen, können aber auch durch Verhaltensbeobachtungen ergänzt werden (vgl. Krumm, Stenzel & Pauls, 2015). Interviews unterscheiden sich in vor allem hinsichtlich der Freiheitsgrade der Gesprächsgestaltung: a) In standardisierten Interviews sind Fragen und Antwortoptionen vorgegeben (diese Form ähnelt sehr einem schriftlichen Fragebogen, wird in der Website-Evaluation sehr selten verwendet und daher in der folgenden Darstellung auch nicht fokussiert); b) in halbstandardisierten Interviews sind typischerweise die Fragen in einem Leitfaden festgelegt, die Befragten in ihren Antworten aber frei (das präferierte Vorgehen in der Website-Evaluation); während c) in vollständig unstandardisierten Interviews beide Seiten komplett frei in ihrer Gesprächsführung sind.

Interviewverfahren kommen in der Website-Evaluation an verschiedenen Stellen zum Einsatz, beispielsweise bei der Befragung von NutzerInnen oder Experten zu ihrer Meinung zu einer Website oder ergänzend zu einem standardisierten Test im Usability-Labor (siehe unten) oder einem anderem quantitativen Verfahren. Sie stellen vorrangig ein qualitatives Verfahren dar, das insbesondere für die formative Bewertung geeignet ist. Die Interviewfragen werden hierbei zumeist in einem halbstandardisierten Interviewleitfaden spezifisch an den Evaluationsgegenstand angepasst.

Gütekriterien

Die Gütekriterien von Interviews sind nicht immer zufriedenstellend – die Erhebungsform *ermöglicht* objektive, reliable und valide Erhebungen, garantiert diese aber nicht. Hinsichtlich der Objektivität ist eine wirkliche Unabhängigkeit zwischen Interviewer und Befragten kaum möglich. Interviewer sind anfällig für typische Beurteilerfehler (bspw. Halo-Effekte oder Kontrasteffekte zwischen verschiedenen Befragten, siehe Döring & Bortz, 2016 bzw. Seite 17 dieser Expertise) und Interaktionseffekte zwischen beiden Seiten sind bei der Durchführung eines Interviews hochwahrscheinlich. Allerdings können Datenauswertung und Interpretation standardisiert werden, beispielsweise indem Interviewdaten durch zwei Beurteiler ausgewertet werden und deren Übereinstimmung statistisch bestimmt wird (z.B. anhand von Cohens Kappa κ; vgl. Cohen, 1960; Döring & Bortz, 2016). Grundsätzlich kann die Durchführungsobjektivität durch Maßnahmen wie Strukturierung, Aufzeichnung des Interviews oder Durchführung in Teams („4-Augen-Prinzip“) verbessert werden.
Hinsichtlich der Reliabilität ist die Anwendung von Konsistenzmaßen auf Interviews meist inhaltlich nicht sinnvoll, daher liegen kaum entsprechende Analysen vor. Klassische Verfahren der Validierung lassen sich jedoch anwenden, hierbei zeigt sich wieder die Überlegenheit standardisierter Interviews gegenüber nicht-standardisierten Interviews. Insgesamt haben Interviews eine hohe Nützlichkeit, vor allem bei sequentiellen Entscheidungen (bspw. wenn Interviews zu Websites im Rahmen formativer Evaluationen geführt werden) und in der Kombination mit anderen Evaluationsverfahren (Fragebögen, Laborstudien, etc.; siehe Tabelle 2). Interviews sind besonders wertvoll zur ersten Exploration eines Gegenstandes in der formativen Evaluation und wenn zu bestimmten Fragestellungen (bspw. spezifische Reaktionen von Website-NutzerInnen auf bestimmte Inhalte) keinerlei quantitative Verfahren zur Verfügung stehen.

Notwendige Stichprobenumfänge

Wie beschrieben kommen Interviews oft ergänzend zu anderen Verfahren zum Einsatz und sollten sich dann an den dort herrschenden Stichprobenanforderungen orientieren. Als qualitatives Verfahren alleine können Auswertungen bereits mit zehn bis zwanzig Interviewten durchgeführt werden, wenn diese die Breite und Eigenschaften der Zielgruppe exemplarisch abdecken.

Kosten

Die Kosten eines Tiefeninterviews summieren sich aus den Konzeptionskosten, dem Durchführungsaufwand (Personalkosten und Incentives) sowie dem Personaleinsatz bei der Auswertung. Je nach Länge des Interviews schwankt der Aufwand für die Konzeption. Typische Incentives für qualitative Tiefeninterviews liegen bei um die 130 bis 150 € je Proband, können aber insbesondere in medizinischen Studien oder anderen speziellen Themenbereichen deutlich höher ausfallen. Sehr aufwendig kann jedoch der Personaleinsatz bei der qualitativen Auswertung von Interviews sein, insbesondere wenn diese transkribiert und mittels Inhaltsanalyse strukturiert interpretiert werden (siehe bspw. Mayring, 2010). Die Gesamtkosten für Einzelinterviews in typischen Website-Evaluationen mit kleinen Befragungsgruppen liegen im Bereich von 1.200 € je Befragtem (Stand: Februar 2018); weitreichende Anforderungen an die Datenauswertung oder spezielle Zielgruppen können dies aber erhöhen.

Bewertung Vor-/Nachteile

Vorteile: Der große Vorteil von Interviewverfahren liegt in dem flexiblen und offenen Ansatz, in den Möglichkeiten nachzufragen, bei Bedarf gleichzeitig Verhaltensbeobachtungen durchzuführen und somit potenziell sehr tiefgehende Informationen über die Erlebniswelt eines Befragten zu erhalten.
Nachteile: Um Aussagen verlässlich zu quantifizieren, sind andere Befragungsverfahren besser geeignet. Aufwand und mangelnde Standardisierung klassischer Interviewformen erschweren hier die Generierung verlässlicher Repräsentativaussagen.

Weiterführende Informationen

Für allgemeine Konstruktionsregeln zu Fragen und Interviews siehe:

Thielsch, M. T., Lenzner, T. & Melles, T. (2012). Wie gestalte ich gute Items und Interviewfragen? In M. T. Thielsch & T. Brandenburg (Hrsg.), *Praxis der Wirtschaftspsychologie II: Themen und Fallbeispiele für Studium und Anwendung* (S. 221-240). Münster: MV Wissenschaft.

Praxisbeispiel: Inhalte eines typischen Interviewleitfadens

Ein Interviewleitfaden gliedert sich typischerweise in die folgenden Bereiche (siehe auch Jacobsen & Meyer, 2017, S. 188f.):

1. **Einstieg**: Begrüßung; Selbstvorstellung; Klärung der Formalia: Einführung zum Interview, Interviewziele und -verantwortliche, Ablauf des Interviews, Einverständniserklärung des Interviewten zur Datenverwendung und ggf. durchgeführten Audio-/Videoaufzeichnungen

2. **Aufwärmphase**: Einleitende Fragen, z. B. zur Internetnutzung allgemein, bisherigen Erfahrungen im Zielbereich der zu evaluierenden Website, Erwartungshaltung an diese Website

3. **Explorationsphase**: Exploration der Website mit offenen Fragen; Beobachtung durch Interviewer

4. **Aufgabenphase**: Spezifische Fragen zu typischen Nutzungsszenarien inklusive der Durchführung von Aufgaben auf der Website

5. **Bewertungsphase**: Abfrage von Gesamtbewertungen zur Website wie Gesamteindruck (z.B. auf einer Schulnotenskala) oder ggf. auch weiteren quantitativen Maßen (z.B. hinsichtlich Wiederbesuch und Weiterempfehlung)

6. **Abschluss**: Fragen, ob der Interviewer etwas übergangen hat, was dem Befragten noch wichtig gewesen wäre; Klärung von weiteren offenen Fragen auf Seiten der Interviewten; Ausblick zur weiteren Verwendung der Daten; Dank, Incentivierung und Verabschiedung

Fokusgruppen

Hintergrund

Verfahrenstyp: ***Qualitativ***
Test aus Nutzersicht
Zeitpunkt: ***formativ***
Zeitaufwand: ***hoch***
Psychometrische Güte: ***gering***

In einer Fokusgruppe diskutiert eine kleine Gruppe potentieller NutzerInnen über die zu evaluierende Website. Bei solchen Gruppendiskussionen sind neben dem Moderator typischerweise fünf bis zehn Personen anwesend. Der Moderator lenkt das gemeinsame Gespräch auf Basis eines vorher vorbereiteten Leitfadens. Dieses qualitative Verfahren ähnelt methodisch stark dem qualitativen Interview und erlaubt einen Informationsgewinn durch direktes Nachfragen und die angeregte Diskussion der NutzerInnen untereinander. Fokusgruppen sind insbesondere für formative Evaluationen geeignet und werden besonders in frühen Evaluationsphasen eingesetzt.

Gütekriterien

Aufgrund des dynamischen und offenen Formates genügen Fokusgruppen kaum den klassischen Gütekriterien. Ergebnisse der Gruppendiskussion sind von den TeilnehmerInnen und ihrer Interaktion untereinander sowie von Rahmenbedingungen und Moderation abhängig. Objektivität als Unabhängigkeit zwischen Moderator und Befragten ist kaum möglich, typische Beurteilerfehler und insbesondere gruppendynamische Effekte sind zu erwarten (vgl. Döring & Bortz, 2016). Allerdings können wie beim Interview Datenauswertung und -interpretation standardisiert werden, beispielsweise indem die Diskussionsergebnisse transkribiert und durch zwei Beurteiler ausgewertet werden. Deren Übereinstimmung kann dann statistisch bestimmt werden (z.B. anhand von Cohens Kappa κ; vgl. Cohen, 1960; Döring & Bortz, 2016).
Reliabilität und Validität sind als Gütekriterien auf Fokusgruppen schwer anwendbar – die jeweiligen Diskussionen werden unterschiedliche Verläufe nehmen und nicht im klassisch testtheoretischen Sinne replizierbare Ergebnisse liefern. Die Empfehlung ist daher, mehrere Fokusgruppen durchzuführen bis eine sogenannte theoretische Sättigung erreicht ist (das heißt weitere Fokusgruppen generieren keine neuen Ergebnisse mehr).

Notwendige Stichprobenumfänge

Typischerweise werden mindestens zwei Fokusgruppen mit jeweils fünf bis zehn TeilnehmerInnen eingesetzt. Besser ist hier jedoch drei bis vier Fokusgruppen einzuplanen – erfahrungsgemäß können aufgrund von Problemen im Diskussionsverlauf, Daten aus einer Gruppendiskussion durchaus wenig bis gar nicht geeignet sein. Mit drei bis vier Fokusgruppen lässt sich dieses auffangen. Ziel ist eine theoretische Sättigung zu erreichen.

Kosten

Die Kosten einer Fokusgruppe summieren sich aus den Konzeptionskosten, dem Durchführungsaufwand (Personalkosten Moderator und Incentives) sowie dem Personaleinsatz bei der Auswertung. Je nach Länge der Diskussion und Menge der Fokusgruppen schwankt der Aufwand. Sehr aufwendig kann der Personaleinsatz bei der qualitativen Auswertung sein, insbesondere wenn diese transkribiert und mittels Inhaltsanalyse strukturiert interpretiert werden (siehe bspw. Mayring, 2010). Damit liegen die Kosten bei typischen methodischen Aufbaus (Durchführung zweier Fokusgruppen, jeweils sechs bis acht TeilnehmerInnen, Dauer > 1 Stunde) bei insgesamt rund 10.000-12.000 € (Stand: Februar 2018). Wie beschrieben wären methodisch jedoch mehr als zwei Fokusgruppen sinnvoll.

Bewertung Vor-/Nachteile

Vorteile: Der Vorteil von Fokusgruppen liegt in dem flexiblen und offenen Ansatz, in den Möglichkeiten nachzufragen und Themen in die Tiefe zu diskutieren. Fokusgruppen haben vor allem als kreatives Verfahren in der formativen Evaluation ihre Berechtigung.
Nachteile: Fokusgruppen eignen sich keinesfalls zur sicheren und validen Erfassung von Informationen, die repräsentativ für eine Zielgruppe sind.

Weiterführende Informationen

Eine anschauliche Darstellung findet sich in Kapitel 9 bei:
Jacobsen, J. & Meyer, L. (2017). *Praxisbuch Usability und UX*. Bonn: Rheinwerk Verlag.

Ablauf einer Fokusgruppe

Eine Fokusgruppendiskussion gliedert sich typischerweise in drei Teile (vgl. Jacobsen & Meyer, 2017, S. 98f.):

1. **Einleitung:** Moderator und TeilnehmerInnen stellen sich vor, der Moderator führt in das Thema ein, erläutert Ablauf, Zielsetzung und Verhaltensregeln.

2. **Hauptteil:** Hier findet die eigentliche Diskussion statt. Der Moderator führt diese anhand eines vorher erstellten Leitfadens, bei Bedarf werden Stimuli / Websites gezeigt oder spezifische Teilaufgaben gestellt. Die TeilnehmerInnen können gebeten werden, verschiedene Perspektiven einzunehmen (z.B. NutzerIn, AnbieterIn, etc.) oder konkrete Fragen zu beantworten (z.B. wie wirkt Element XY?).

3. **Zusammenfassung:** Abschließend ziehen die TeilnehmerInnen ein persönliches Fazit zu den Ergebnissen der Diskussionen. Der Moderator fasst zusammen, dankt und verabschiedet die TeilnehmerInnen.

Die Auswertung der Ergebnisse erfolgt im Nachgang, optimalerweise liegen hierfür auch Video- oder Audioaufnahmen der Diskussion vor.

Checklisten und Guidelines

Hintergrund

Verfahrenstyp: ***Quantitativ Test aus Expertensicht***
Zeitpunkt: ***formativ oder summativ***
Zeitaufwand: ***niedrig***
Psychometrische Güte: ***gering bis mittel***

Schon frühzeitig wurden in der Praxis und Forschung zur Mensch-Computer Interaktion verschiedene Gestaltungsrichtlinien entwickelt, um die Arbeit von Entwicklern zu unterstützen (vgl. Sarodnick & Brau, 2015, S. 122f.). Für die Website-Evaluation relevant sind dabei vor allem Expertenleitfäden oder Checklisten. Diese können systematisch von Experten oder Verantwortlichen einer Website durchgearbeitet werden und erlauben einen ersten Einblick in die Qualität dieser. Checklisten ermöglichen damit eine erste Quantifizierung, ob und inwieweit geforderte Designprinzipien umgesetzt sind. Checklisten können auch als Guideline für Entwickler und zur Identifikation weiterer Evaluationsbedarfe dienen. Dabei ist zudem eine Anwendung im Kontext vergleichender Evaluationen denkbar, bei der die eigene Website anhand von Checklisten mit anderen themenverwandten Anbietern verglichen wird.

Gütekriterien

Eine Vielzahl von Checklisten ist online verfügbar, ihr Erfolg basiert in der Regel auf der Akzeptanz in der Praxis. Viele Guidelines resultieren allerdings lediglich aus der Meinung einzelner Experten, nennen wenige bis keine Referenzen, und systematische Prüfungen der Gütekriterien finden sich leider selten. Damit sind viele verfügbare Checklisten eher kritisch zu sehen und es bleibt unklar, inwieweit diese vollständig und valide sind. Es gibt aber auch Ausnahmen, wie beispielsweise die „Web Content Accessibility Guidelines (WCAG) 2.0“ zur Barrierefreiheit von Websites (siehe https://www.w3.org/Translations/WCAG20-de/) oder die „HHS Web usability Guidelines“ des US Gesundheitsministeriums (US Department of Health and Human Services, 2006; siehe https://webstandards.hhs.gov/guidelines/ bzw. für eine Buchversion https://www.usability.gov/sites/default/files/documents/guidelines_book.pdf).

Bei beiden genannten Beispielen erfolgten zumindest umfassende Prüfungen durch verschiedene Gutachter und Experten. Empirische Validierungsstudien werden allerdings auch bei diesen Beispielen nicht in den Manualen berichtet.

Notwendige Stichprobenumfänge

Im Prinzip lassen sich Checklisten durch eine Person anwenden. Zur Erhöhung der Objektivität ist jedoch anzuraten, je Evaluation mindestens zwei Personen eine Checkliste anwenden zu lassen und Unterschiede zwischen den Beurteilungen hinsichtlich möglicher systematischer Verzerrungen zu hinterfragen. In der Regel erfolgt die Anwendung von Checklisten durch Experten, Laien sind oftmals mit (technischen) Begrifflichkeiten überfordert und bringen nicht die notwendige Ausbildung im Bereich Evaluation mit.

Kosten

Verschiedene Checklisten zur Website-Evaluation sind online frei verfügbar, allerdings oftmals von zweifelhafter Qualität (siehe oben, Gütekriterien). Empfohlen werden kann aber die Nutzung von geprüften Checklisten wie beispielsweise der oben genannten WCAG und HHS guidelines oder insbesondere der nachfolgend im Praxisbeispiel genannten BaNu. Damit fallen lediglich Personalkosten in der Durchführung, Auswertung und Interpretation an. Für externe Experten kann hier der jährliche Branchenreport der GermanUPA einen Ansatzpunkt für übliche Vergütungen liefern (siehe http://www.germanupa.de/berufsfeld-usabilityux-professionals/branchenreport), in 2017 lagen durchschnittliche Stundensätze für Selbstständige bei 82 €, der mittlere Tagessatz bei 600 € (Tretter et al., 2017).

Bewertung Vor-/Nachteile

Vorteile: Checklisten sind leicht anzuwenden und erlauben eine frühzeitige und einfache Prüfung einer Website.
Nachteile: Checklisten sind oft sehr lang und basieren auf starren, allgemeinen Richtlinien. Damit ist letztendlich keine spezifische Aussage über das Erleben der tatsäch-

lichen Website-NutzerInnen im Anwendungskontext möglich. Die Kombination von Checklisten-Evaluationen mit anderen Evaluationsmethoden, die Nutzertests einschließen, ist daher notwendig.

Weiterführende Informationen

Eine Vielzahl von Standards und Checklisten wurde in der deutschsprachigen Online-Plattform „BaNu – Barrieren finden, Nutzbarkeit sichern“ zusammengeführt. Diese wird vom Informationstechnikzentrum Bund unter http://www.banu.bund.de bereitgestellt.

Praxisbeispiel: BaNu – Barrieren finden, Nutzbarkeit sichern

„BaNu – Barrieren finden, Nutzbarkeit sichern“ wurde im Rahmen des Regierungsprogramms E-Government 2.0 unter dem Projekttitel „Nutzungsfreundlichkeit und Barrierefreiheit“ entwickelt. Ziel des Projektes war, Behörden ein Werkzeug zur Verfügung zu stellen, mit dessen Hilfe sie die Qualität bereits bestehender Webangebote überprüfen können. BaNu stellt einen Prüfkatalog zu Barrierefreiheit und Nutzungsfreundlichkeit für E-Government-Angebote wie Websites oder PDFs zur Verfügung. BaNu fasst über 20 Standards, Verordnungen und Richtlinien zur Barrierefreiheit und Nutzungsfreundlichkeit zu allgemeinverständlichen Fragen und Prüfanleitungen zusammen (darunter auch die oben genannten HHS Web usability Guidelines). Dabei steht die Selbstanalyse im Vordergrund, die Bewertung soll durch die fachlichen MitarbeiterInnen selber möglich sein.

Diese können sich mit BaNu Schritt für Schritt eine individuell auf ihre Bedürfnisse angepasste Checkliste generieren, die nur die Fragen beinhaltet, die für die eigene Prüfung relevant sind. In den Prüfkatalogen sind die zu bewertenden Kriterien in Form von einfachen und verständlichen Fragen hinterlegt. Es sind weitere Informationen zu den Prüffragen, Hilfen und ergänzende Hinweise verfügbar.

Ist die Prüfung einer Website beendet, können verschiedene Auswertungen erstellt werden: Kurzzusammenfassungen, ausführliche Beschreibungen oder auch Überarbeitungspläne, die alle verbesserungswürdigen Punkte beinhalten. Alle Auswertungen können auch als PDF-Dokument heruntergeladen werden. Hauptadressaten der Anwendung sind Mitarbeiterinnen und Mitarbeiter von Behörden, aber auch alle anderen Interessierten, wie beispielsweise Dienstleister und Agenturen, können BaNu nutzen. Für weitere Fragen steht laut Website das ITZBund zur Verfügung.

Quelle dieser Informationen: https://www.itzbund.de/DE/Produkte/BaNu/banu_node.html

Standardisierte Fragebogenverfahren

Hintergrund

Verfahrenstyp: ***Quantitativ Test aus Nutzersicht***
Zeitpunkt: ***summativ***
Zeitaufwand: ***mittel***
Psychometrische Güte: ***gut bis sehr gut***

Standardisierte Fragebögen kommen in der Website-Evaluation häufig zum Einsatz. Sie stellen ein quantitatives Verfahren dar, das insbesondere für die summative Bewertung geeignet ist. Vorhandene Instrumente decken dabei eine Vielzahl von Themen zur Einschätzung von Web-Inhalten und Website-Designs ab. Wichtig ist jedoch darauf zu achten, dass Verfahren spezifisch für den Einsatzbereich Website-Evaluation validiert wurden. Gerade im Bereich Usability werden oftmals ungeprüft Verfahren aus dem Bereich der Software-Evaluation auf Websites angewendet. Aufgrund der unterschiedlichen Anforderungen an Software im Vergleich zu Websites ist dies kritisch zu bewerten. Ein genauer Blick auf die Güte eines Fragebogens ist hier entscheidend.

Gütekriterien

Zentrale Kriterien der psychometrischen Güte eines Fragebogenverfahrens sind Objektivität, Reliabilität und Validität (siehe bspw. Bühner, 2010; Moosbrugger & Kelava, 2012). Jedes standardisierte Verfahren sollte diese Anforderungen erfüllen. Eine Reihe von validen deutschsprachigen Fragebogen-Verfahren der Website-Evaluation ist in der (von der BZgA beauftragten) „Toolbox zur kontinuierlichen Website-Evaluation und Qualitätssicherung“ (Thielsch, 2017) dokumentiert.
Grundsätzlich gilt: Wenn keine Informationen zu Gütekriterien eines Fragebogens vorliegen, ist das Verfahren nicht einzusetzen.

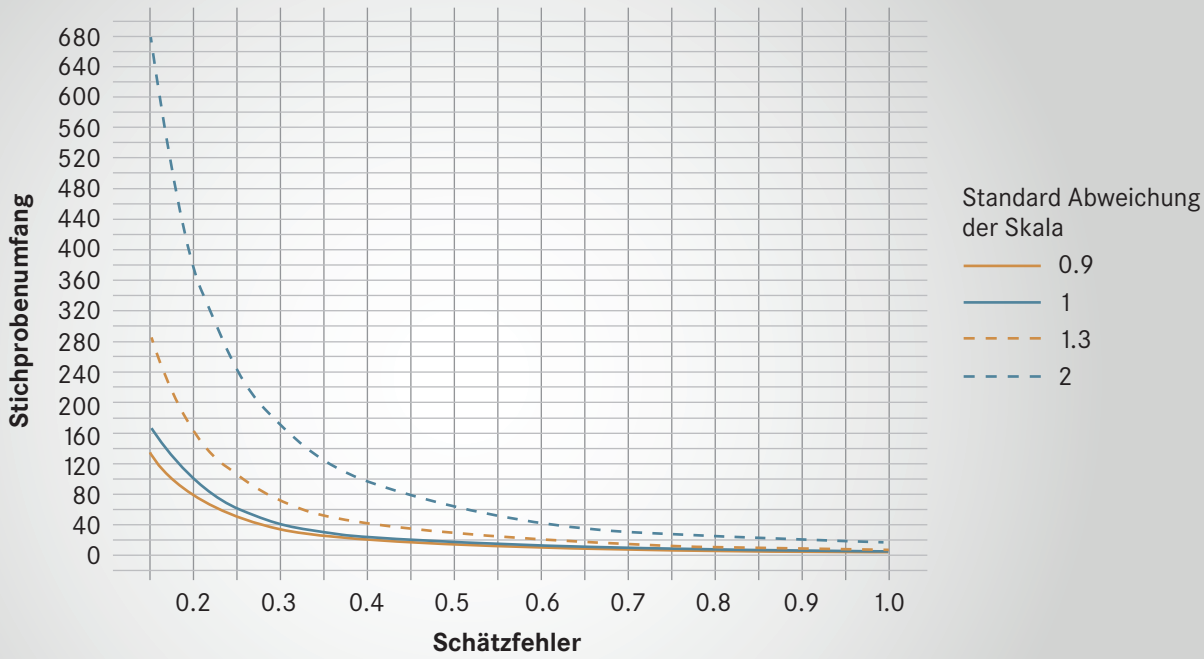


Abbildung 3: Optimale Stichprobenumfänge für standardisierte Fragebogenverfahren (eigene Abbildung)

Notwendige Stichprobenumfänge

Entscheidend sind hier zwei Faktoren: a) eine zufällige und möglichst repräsentative Auswahl von Personen aus der Zielgruppe sowie b) die Stichprobengröße selber. Die notwendige Menge an Befragungspersonen ist dabei davon abhängig, wie genau das Merkmal erfasst werden soll. Je nachdem, wie genau man zum Beispiel den Mittelwert einer Skala schätzen möchte, braucht man unterschiedliche viele Probanden. Abbildung 3 zeigt die benötigten Probanden, um Skalen mit Standardabweichungen von 0,9; 1; 1,3 und 2 unterschiedlich genau zu schätzen. Beispiel: Typische Fragebögen im Bereich Website-Evaluation wie der Web-CLIC (zur Erfassung von Website-Inhalten, Thielsch & Hirschfeld, in press; siehe auch Praxisbeispiel S. 39) haben eine Standardabweichung von um die 1,3. Möchte man die mittlere Inhaltsbewertung der Website-NutzerInnen mit einer Präzision (hier dem 95% Konfidenzintervall) von ± 0,15 Punkten mit dem Web-CLIC erfassen, benötigt man 290 Probanden, reicht es aus dieses Merkmal mit einer Präzision von ± 0,3 Punkten zu erfassen, sind bereits etwa 80 Probanden ausreichend.

Werden verschiedene Websites oder Website-Versionen verglichen, sind bereits 50 Personen ausreichend um mittelstarke Unterschiede zwischen den Websites zu detektieren. Die exakte Größe kann mit gängigen Open Source Programmen zur Stichprobenplanung (beispielsweise G*power, siehe Faul et al., 2007, bzw. http://www.gpower.hhu.de/) berechnet werden. Generell sollten aber immer mindestens 30 Personen je Gruppe getestet werden.
Sind repräsentative Aussagen über bestimmte Zielgruppen oder gar die Bundesbevölkerung angestrebt, werden üblicherweise entsprechend geschichtete Stichproben mit mindestens 1000 Personen befragt (Döring & Bortz, 2016, siehe aber dort insbesondere Seite 398 für eine kritische Bewertung dieser Praxis).
Ausnahme: Bei sehr kleinen Zielgruppen gilt dies natürlich nicht. Gibt es zum Beispiel in einer bestimmten Patienten-Zielgruppe in ganz Deutschland nur wenige tausend Erkrankte, sollten die optimalen Stichprobenumfänge für repräsentative Aussagen wie oben beschrieben in Abhängigkeit von der Fehlertoleranz individuell bestimmt werden.

Kosten

Eine Reihe von validen Fragebogenverfahren zur Website-Evaluation ist frei verfügbar und in der Toolbox zur kontinuierlichen Website-Evaluation und Qualitätssicherung der BZgA dokumentiert (Thielsch, 2017). Dementsprechend ist die kostenpflichtige Lizensierung kommerzieller Verfahren nur dann gerechtfertigt, wenn diese a) eine höhere Güte haben oder b) bei vergleichbarer psychometrischer Güte differenziertere Ergebnisse anbieten.
Neben Personalkosten für Planung, Auswertung und Berichtserstellung sind bei Fragebogenverfahren insbesondere die Kosten der Stichprobe zu bedenken. Den Befragten wird in der Regel eine Aufwandsentschädigung gezahlt, dies führt in üblichen Stichprobenumfängen (n = 300) schnell zu Kosten im vierstelligen Bereich. Feldkosten für typische Befragungen (ca. 15-20 Minuten Dauer, Schichtungen bspw. hinsichtlich der Personenvariablen Alter, Geschlecht und Bildungsgrad) liegen derzeit für Panelstudien bei um die 5 € je Proband. Diese können aber insbesondere in medizinischen Studien deutlich höher ausfallen. Grundsätzliche Hinweise zur Gestaltung von Aufwandsentschädigungen und Incentivierung in Befragungen finden sich bei Pforr (2015), spezifisch für Online-Panels bei Göritz (2014) – bei diesen Quellen ist aber zu bedenken, dass Panelkosten in den letzten Jahren gefallen sind. Typische Umfragen mit Panel-

unterstützung (n = ca. 300) erreichen somit derzeit Gesamtkosten (inkl. Auswertung und Bericht) um die 10.000 bis 12.000 €, bei On-Site Befragungen (mit n = ca. 200-500) sind es um die 7.000 € (da hier klassische Panelkosten entfallen, Stand: Februar 2018). Aber: On-Site Befragungen setzen voraus, a) dass genug Besucher die Ziel-Website frequentieren und b) die technischen Voraussetzungen zur Einbindung der Befragung auf der Website da sind. Letzteres kann zusätzlichen Aufwand und damit deutliche Zusatzkosten mit sich bringen, Panelbefragungen sind daher oft der gewählte Weg.

Bewertung Vor-/Nachteile

Vorteile: Der große Vorteil von Fragebogenverfahren besteht darin, dass sich Aussagen verlässlich quantifizieren lassen. Durch die Analyse relativ großer Zielgruppen sind verlässliche Prognosen möglich.
Nachteile: Nachteilig ist die fehlende Reaktivität und Tiefe derartiger standardisierter Befragungen. Nachfragen ist ebenso wenig möglich wie eine detaillierte und umfassende Betrachtung des individuellen Erlebens einzelner Personen.

Weiterführende Informationen

Eine Vielzahl von Fragebogenverfahren zur Website-Evaluation finden sich dokumentiert in:

Thielsch, M. T. (unter Mitarbeit von Salaschek, M.) (2017). *Toolbox zur kontinuierlichen Website-Evaluation und Qualitätssicherung (Version 2.0)*. Arbeitsbericht, Köln: Bundeszentrale für gesundheitliche Aufklärung (BZgA). http://dx.doi.org/10.17623/BZGA:224-2.0

Auf der nachfolgenden Seite ist aus dieser Toolbox die Übersichtstabelle aller dargestellten Verfahren (Thielsch, 2017, S. 8) inklusive Informationen zu deren Gütekriterien abgebildet (siehe Tabelle 3).

Konstrukt	Instrument	Nr.	# Items	Quelle	Standard-version	Erweiterte Version	Interpreta-tionshilfen	Reliabilität	Validität
1. Ersteindruck	Einzelitems	1.1	4	Thielsch (2017)	√	√	-	*	*
2. Subjektive Inhaltswahrnehmung	Message Credibility Scale	2.1	3	Appelman & Sundar (2016)	√	√	√	*	*
	WWI (Fragebogen zur Wahrnehmung von Website-Inhalten)	2.2	9	Thielsch (2017)	√	√	√	je Skala (*) bis **	(*)
	Trusting Beliefs	2.3	11	McKnight et al. (2002)		(√)[a]	-	**	(*)
3. Subjektive Usability / Nutzerzufriedenheit	UMUX-Lite (Usability Metric for User Experience - Lite)	3.1	2	Lewis et al. (2013)	√		-	*	*
	PWU-G (Perceived Website Usability - German)	3.2	7	Flavián et al. (2006), Thielsch (2017)	√	√	√	**	**
	System Usability Scale (SUS)	3.3	10	Brooke (1996)		√	√	je Studie * bis **	*
4. Visuelle Ästhetik	VisAWI-S (Visual Aesthetics of Websites Inventory - Short)	4.1	4	Moshagen & Thielsch (2013)	√		√	je Studie (*) bis *	**
	VisAWI (Visual Aesthetics of Websites Inventory)	4.2	18	Moshagen & Thielsch (2010)		√	√	Skalen: * Gesamtwert: **	**
5. Emotionale Reaktion: Befindlichkeit, Zustimmung, Zufriedenheit	Smiley-Skala	5.1	1	Jäger (2004)	√	√	√	nb	**
6. Gesamteindruck	Einzelitem Gesamteindruck	6.1	1	Thielsch (2017)	√	√	√	nb	(*)
7. Handlungs- und Nutzungsintentionen	Wiederbesuchs-Skala (Scale assessing the intention to revisit the website)	7.1	4	Moshagen & Thielsch (2010)	√	√	-	**	(*)

Tabelle 3: In die Toolbox (Thielsch, 2017) eingeschlossene Items/Instrumente (insgesamt 35 Items in der Standardversion, bis zu 68 Items in der erweiterten Version) und deren Quellen.

Hinsichtlich der Reliabilität wurden zu den Verfahren die interne Konsistenz (Cronbachs Alpha) betrachtet und wie folgt dargestellt ** = α ≥ .9; * = α ≥ .8; (*) = α ≥ .7; - = α < .7. Bei Einzelitems kann Cronbachs Alpha nicht berechnet werden (dort „nb"), Hinweise zur Retest-Reliabilität liegen bei diesen Items derzeit nicht vor. Hinsichtlich der Validität gilt: ** = umfassende Belege für hohe Validität liegen vor; * = Validitätsbelege liegen vor; (*) = nur wenige Hinweise zur Validität liegen vor bzw. diese könnte eingeschränkt sein. [a] Die Trusting Beliefs-Skala ist in der Forschung sehr weit verbreitet; da notwendige Validitätsprüfungen noch ausstehen, ist ihr Einsatz im Einzelfall zu prüfen.

Praxisbeispiel: Der Web-CLIC Fragebogen

Der Web-CLIC ist ein Fragebogen mit dem NutzerInnen den Inhalt einer Website bewerten können (Thielsch & Hirschfeld, in press). Insgesamt 12 Fragen decken dabei die vier Bereiche Verständlichkeit, Gefallen, Informationsgehalt und Glaubwürdigkeit ab (die Abkürzung Web-CLIC steht für „Website-Clarity, Likeability, Informativeness, Credibility"). Zudem kann ein Gesamtwert berechnet werden, dieser spiegelt das subjektive Erleben des Website-Inhalts insgesamt wider (siehe Abbildung 4). Beispielhafte Items des Web-CLICs sind

- „Die Inhalte sind anschaulich aufbereitet." (Skala: Verständlichkeit)
- „Ich lese diese Website gerne." (Skala: Gefallen)
- „Die Website ist informativ." (Skala: Informationsgehalt)
- „Ich kann den Informationen auf der Website vertrauen." (Skala: Glaubwürdigkeit)

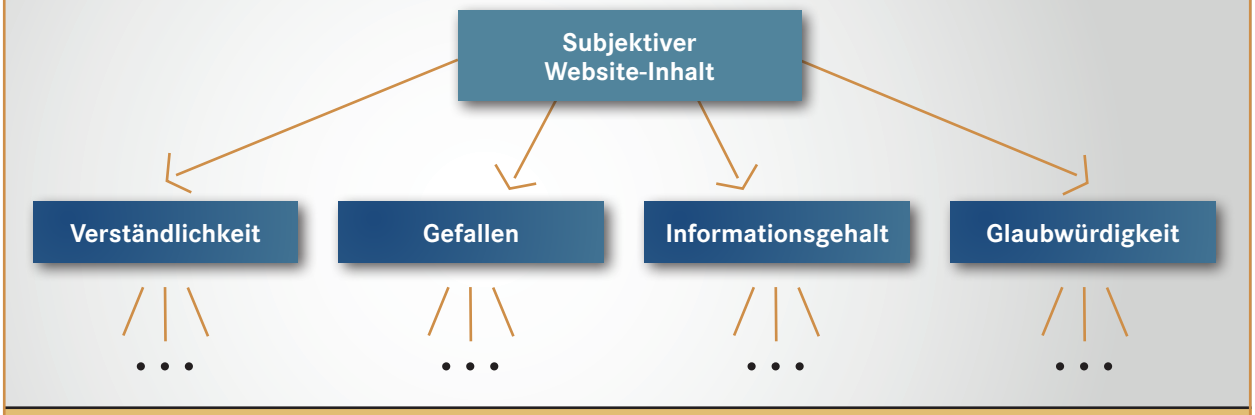


Abbildung 4: Strukturmodell des Web-CLICs (Thielsch & Hirschfeld, in press)

Konstruktion und Validierung des Web-CLICs basieren auf sechs Studien mit insgesamt n = 3106 Befragten und m = 60 getesteten Websites (Thielsch & Hirschfeld, in press). Der Fragebogen erweist sich als reliabel: Die interne Konsistenz (Cronbachs α) ist mindestens > .80 für die Skalen und > .90 für den Gesamtwert. Die Zeitstabilität (Retest-Relibalität) über einen Zeitraum von 2 Wochen liegt für die Skalen im Bereich von .69 ≤ r ≤ .81, bzw. bei r = .84 für den Gesamtwert. Hinsichtlich der Validität findet sich eine umfassende empirische Evidenz, geprüft wurden faktorielle, konvergente, divergente, diskriminative, konkurrente, experimentelle und prädiktive Validität. In den Studien der Autoren erfasste der Web-CLIC beispielsweise spezifisch und zielgenau experimentelle Variationen der Glaubwürdigkeit einer Gesundheitsinformationswebsite und war darüber hinaus in der Lage das Spendenverhalten der Befragten für gemeinnützige Organisationen vorherzusagen.

Als weitere Auswertungshilfen liegen für den Web-CLIC optimale Schwellenwerte und Benchmarks für zehn verschiedene Website-Kategorien auf Basis von 7379 Bewertungen von m = 120 Websites vor. Weitere Informationen finden sich unter www.WebCLIC.de.

Verhaltensbeobachtung im Labor

Hintergrund

Die Verhaltensbeobachtung im Labor wird im Kontext Website-Evaluation oft auch schlicht als „Usability-Test“, „Usability-Lab(or)“ oder „UX Lab(or)“ bezeichnet. Im Rahmen einer solchen Labortestung werden Personen aus der Zielgruppe einer Website bei der Nutzung dieser beobachtet, dabei können verschiedene Aufgaben gestellt und auf unterschiedliche Weise Daten erhoben werden. Neben der Beobachtung (mittels Videoaufzeichnung und/oder durch einen Einwegspiegel) sind viele verschiedene Messungen denkbar, so unter anderem:

Verfahrenstyp: ***Qualitativ und quantitativ***

Test aus Nutzersicht

Zeitpunkt: ***formativ oder summativ***

Zeitaufwand: ***hoch***

Psychometrische Güte: ***mittel bis gut***

- Fehlerquoten und Zeitaufwand für gegebene Aufgaben
- Strukturierte qualitative Interviews anhand eines Leitfadens, gezieltes Nachfragen zu gegeben Aufgaben
- Thinking-aloud (Methode des lauten Denkens): Die Probanden sollen während der Nutzung der Website alles was sie denken verbalisieren und laut aussprechen
- Mouse-Tracking und Screen-Capture: Mausbewegungen und Bildschirmeingaben werden aufgezeichnet
- Eye-Tracking (Blickverlaufsmessung): Es wird erfasst welche Teile der Website wann und wie lange betrachtet werden

Die Verhaltensbeobachtung im Labor kann somit eher qualitativ oder eher quantitativ ausgerichtet sein sowie Methoden aus beiden Bereichen kombinieren. Verschiedene Verfahren können hier in eine sehr leistungsstarke Testbatterie zusammengestellt werden.

Gütekriterien

Die Güte eines Labor-Tests ist abhängig von den individuell eingesetzten Methoden. Kommen eher qualitative Methoden zum Einsatz, so sind aufgrund der Probleme in der Standardisierung die Ergebnisse möglicherweise weniger reliabel und valide. Kommen quantitative Verfahren mit ausreichend großen Stichproben zum Einsatz, können die Testergebnisse im Labor ausreichend reliabel und valide sein. Stets bleibt aber kritisch zu hinterfragen, inwieweit die Übertragbarkeit der Ergebnisse in die reale Nutzungssituation möglicherweise durch die Labortestung an sich eingeschränkt ist.

Notwendige Stichprobenumfänge

Oftmals hört man in der Praxis die von Jacob Nielsen geprägte Daumenregel fünf Testpersonen seien ausreichend (Nielsen, 1993). Faulkner (2003) zeigte, dass bei nur fünf Testpersonen aber substanzielle Probleme übersehen werden können und empfahl die Zahl auf mindestens 10, für sichere Datengrundlagen auf 20 anzusetzen. Basierend auf Macefield (2009) kann dies weiter differenziert werden: Für explorative Studien sind 10 bis 20 Probanden optimal, für Studien mit einem statistisch vergleichenden Ansatz (z.B. Vergleich zweier Website-Versionen) ist mit (mindestens) 15 bis 25 Probanden pro Versuchsgruppe zu planen. Aufgrund des hohen Anteils qualitativer Methoden ist letzteres aber eine eher seltenere Form der methodischen Umsetzung.

Kosten

Die Kosten einer Verhaltensbeobachtung im Labor sind sehr abhängig von den eingesetzten Methoden. Aufgrund des hohen Aufwands sind hier in typischen methodischen Settings (12 bis 16 Testpersonen; Dauer jeweils mind. 1 bis 1,5 Stunden) Kosten von 1.000 € je Proband, das heißt in Summe 12.000-16.000 € zu erwarten (Stand: Februar 2018).

Bewertung Vor-/Nachteile

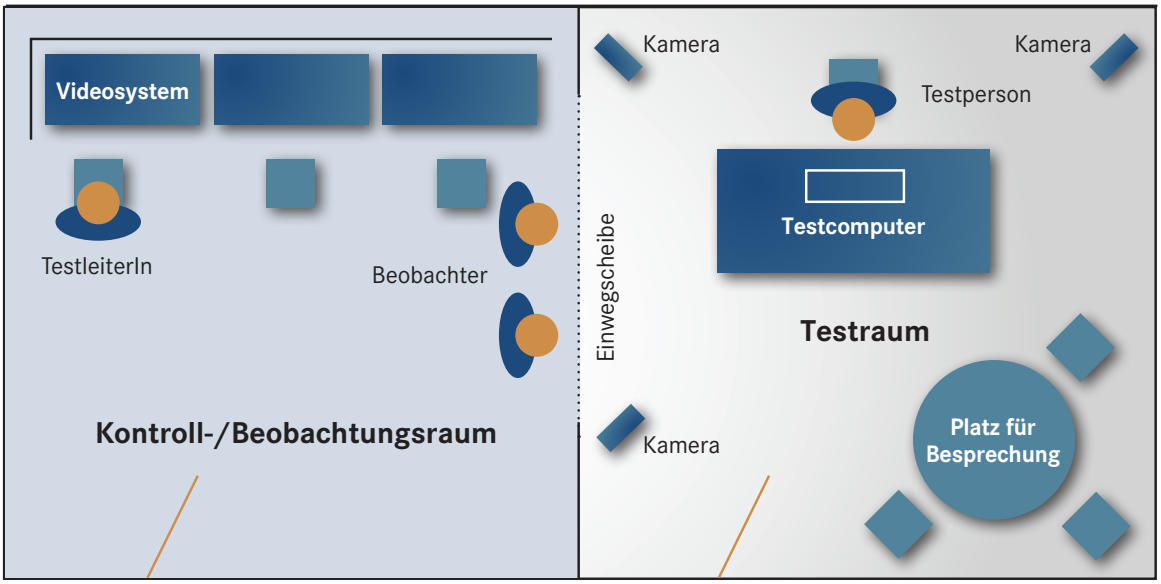
Vorteile: Die Verhaltensbeobachtung im Labor liefert einen umfassenden Einblick in das Verhalten der Probanden mit einer Website und ist damit ein sehr leistungsstarkes Verfahren. Ein wichtiger Pluspunkt ist die Verhaltensorientierung; Antwort- und Verzerrungstendenzen, wie sie in klassischen Einstellungsmessungen auftreten, sind hier weitgehend minimiert.
Nachteile: Neben den hohen Kosten ist ein zentraler Nachteil, dass die Übertragbarkeit der Ergebnisse in die reale Nutzungssituation überprüft werden muss. Hierzu eignen sich Interviews im realen Nutzungskontext, standardisierte Fragebogenverfahren oder Verhaltensbeobachtungen im Feld (Logfile-Analysen), siehe dazu auch Tabelle 2.

Weiterführende Informationen

Praktische Hinweise finden sich in Kapitel 17 bei:
Jacobsen, J. & Meyer, L. (2017). *Praxisbuch Usability und UX*. Bonn: Rheinwerk Verlag.

Aufbau eines Usability Labors

Schematische Darstellung eines typischen Aufbaus nach Sarodnik und Brau (2015, S. 168):



Verhaltensbeobachtung im Feld: Logfile-Analysen

Hintergrund

Verfahrenstyp: ***Quantitativ***
Beobachtung Nutzerverhalten
Zeitpunkt: ***summativ***
Zeitaufwand: ***niedrig***
Psychometrische Güte: ***gut***

Beim Surfen im Internet erzeugen NutzerInnen vielfältige Verhaltensspuren. Diese können im Rahmen von Logfile-Analysen ausgewertet werden. Dabei ist zu unterscheiden, wie detailliert ein Webserver das Verhalten von NutzerInnen speichert. Im einfachsten Fall werden tatsächlich nur Zugriffszahlen über die Zeit erfasst. Sobald man aber spezialisierte Software verwendet und diese auf den einzelnen Seiten einbindet, kann auch die Verweildauer auf einzelnen Webseiten gespeichert werden. Zusätzlich werden häufig Informationen über die NutzerInnen gesammelt, beispielsweise der Ort an dem diese sich befinden, das verwendete Betriebssystem, der Browser, oder die Webseite, die vor der Zielwebseite besucht wurde. Zwei große Anbieter sind Google Analytics und Matomo (vormals als Piwik bekannt). Aufgrund der in Deutschland geltenden Datenschutzregelungen sollte jedoch nur Letzteres eingesetzt werden. Um dieses System zu verwenden, muss man im Prinzip auf jeder einzelnen Webseite ein Skript einbinden, das automatisch aufgerufen wird und die entsprechenden Verhaltensdaten speichert.

Gütekriterien

Es ist sehr schwierig, die Gütekriterien für Logfile-Analysen anzugeben. Einerseits stellen diese ein sehr direktes Maß der Websitenutzung dar, das heißt Logfiles haben eine maximale Validität um dieses Merkmal abzubilden. Andererseits kann es recht schwierig sein, auf Basis solcher Verhaltensdaten allein auf spezifische zugrundeliegende Probleme hinsichtlich Inhalt, Usability oder Ästhetik einer Website zu schließen.

Notwendige Stichprobenumfänge

Streng genommen handelt es sich bei der Betrachtung von Logfiles nicht um eine Stichprobenziehung, da ja alle NutzerInnen erfasst werden. Dennoch ist klar, dass man auch bei Logfile-Analysen darauf achten sollte, dass die zu betrachtenden Maße (z.B. die Abbruchquote einer einzelnen Webseite), für eine ausreichend große Zahl von Personen berechnet wird. Es gelten ansonsten die gleichen Zusammenhänge zwischen der gewünschten Präzision und der Stichprobengröße, die auch schon oben für standardisierte Fragebogenverfahren beschrieben wurden.

Kosten

Um Logfile-Analysen zur Evaluation von Websites einzusetzen, muss wie oben beschrieben eine spezielle Software (z.B. Matomo) auf dem Server installiert und mit den einzelnen Unterseiten der Website verknüpft werden. Diese Installation ist mit Kosten verbunden. Ferner müssen Methoden entwickelt werden, um diese Logfiles auszuwerten (beziehungsweise Personen entsprechend beauftragt werden). Software wie Matomo liefert hierfür in der Standardinstallation schon viele automatische Auswertungen – beispielsweise Abbruchquoten pro Seite oder eine Übersicht der häufigsten verlinkten Webseiten.

Bewertung Vor-/Nachteile

Vorteile: Der Einsatz von Logfile-Analysen hat den Vorteil, dass sie relevante Kriterien der Website-Evaluation (die tatsächliche Nutzung) direkt abbilden können. Als nicht reaktives Maß sind diese Analysen (anders als persönliche Befragungen) außerdem besonders robust gegenüber Verzerrungen und jede Form von Antworttendenzen.
Nachteile: Es entsteht durch den Einsatz von Logfile-Analysen ein zusätzlicher technischer Aufwand und es müssen besondere datenschutzrechtliche Aspekte beachtet werden.

Weiterführende Informationen

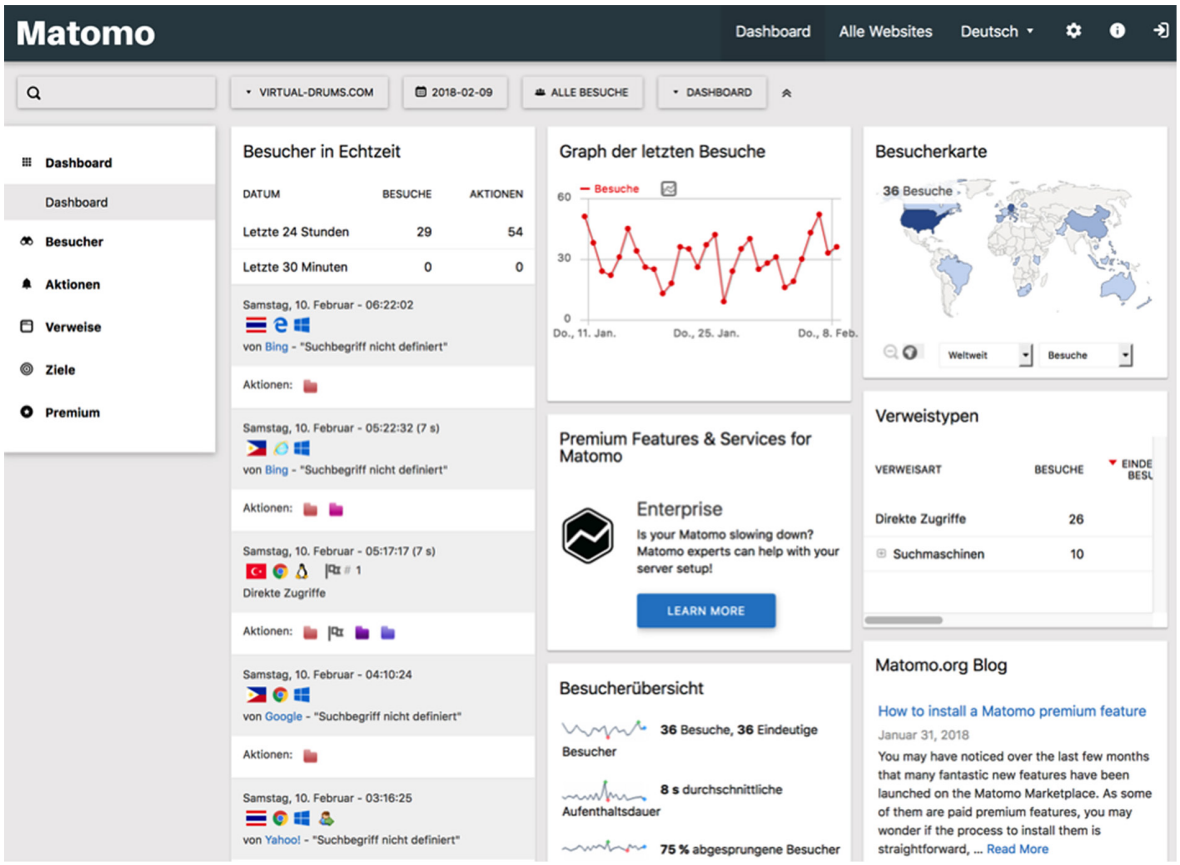
Praktische Hinweise finden sich in Kapitel 22 bei:

Jacobsen, J. & Meyer, L. (2017). *Praxisbuch Usability und UX*. Bonn: Rheinwerk Verlag.

Eine gute Übersicht über die technischen Aspekte von Matomo bietet https://t3n.de/news/piwik-starterg-guide-626254/.

Beispiel: Screenshot einer automatischen Auswertung in Matomo

Eine Online-Demoversion von Matomo (vormals als Piwik bekannt) findet sich unter https://demo.matomo.org/ - hier können die grundlegenden Funktionen betrachtet werden, wie der nachfolgende Screenshot zeigt:



Kritisch zu bewertende Methoden

Verschiedene Methoden, die in der Praxis zur Website-Evaluation angewendet werden, sind aus wissenschaftlicher Sicht kritisch zu sehen. Problematisch sind meistens die mangelnde Reliabilität und Validität der eingesetzten Verfahren oder das Vorhandensein deutlich geeigneterer Ansätze. Diese kritischen Verfahren lassen sich grob in drei Gruppen einteilen:

1. **Ein-Item-Messungen**
   Messungen mit nur einzelnen Fragen sind beliebt. Es kann durchaus seinen Sinn haben, eine Website mit einer einzelnen Frage wie zum Beispiel einer Gesamtnote bewerten zu lassen. In manchen Bereichen mag es zudem schwer fallen verschiedene Fragen zu einem eng umrissenen Themenkreis zu formulieren. Zum Beispiel kann die Weiterempfehlung einer Website oft nur generell mit einer Frage erfasst werden. Allerdings sind solche Einzelitems schlicht nicht so messgenau wie Skalen, die aus mehreren Fragen bestehen (Schmidt & Hunter, 1996) und weniger geeignet um komplexe Konstrukte zu erfassen (Baumgartner & Homburg, 1996). Es entsteht der trügerische Schein, eine Website sehr schnell und umfassend mit nur einer Frage bewerten zu können. Ein typisches Beispiel ist hier der sogenannte „Net Promoter Score“ (NPS; Reichheld, 2003). Der NPS besteht aus einem Einzelitem zur Empfehlung auf einer 11-stufigen Skala. Die Antworten werden jedoch nur in drei Stufen ausgewertet, dabei gelten nur Antworten auf den beiden höchsten Skalenpunkten als gut. In einer derartig vereinfachten Auswertung gehen Informationen verloren und der NPS wird dementsprechend in der Forschung hinsichtlich Güte und Einsetzbarkeit stark kritisiert (siehe z.B. Grisaffe, 2007; Keiningham et al., 2007; Sharp, 2008). Auch lässt die sehr allgemeine Angabe zur Weiterempfehlung im NPS keine sicheren Rückschlüsse auf Aspekte des Designs einer Website zu (siehe Dames et al., under review).

2. **Unstrukturierte Verfahren**
   Manchmal ist schlicht wenig Zeit für Evaluation eingeplant und es wird einfach drauf los getestet. In sogenannten „Guerilla-(Usability)-Tests“ wird dabei teilweise bewusst gegen typische Testregeln verstoßen. Dann werden beispielsweise nur wenige Testpersonen rekrutiert, diese stammen nicht aus der Zielgruppe oder haben andere Eigenschaften als durchschnittliche NutzerInnen der zu evaluierenden Website. Die Argumentation solcher unstrukturierter Verfahren ist, dass „irgendein Test immer noch besser ist als gar keiner“. Solche Verfahren können durchaus hilfreiche Anregungen in frühen Phasen einer Website-Entwicklung liefern – sie können aber keinesfalls strukturierte Evaluationen ersetzen und werden auch nicht die Breite und Tiefe an Ergebnissen liefern wie strukturierte Verfahren.

3. **„Off-label-use“ – ungeprüfte Verwendung von Verfahren aus anderen Kontexten**
   Es kommt vor, dass Verfahren die in einem Kontext gut funktionieren, ungeprüft zur Website-Evaluation angewendet werden. Hier ist dann möglicherweise das Instrument nicht für den Evaluationsgegenstand Website geeignet. Wird zum Beispiel ein klassisches Verfahren aus der Software-Ergonomie (wie z.B. der ISONORM-Fragebogen; Prümper, 1997) auf Websites angewendet, werden nicht alle Fragen sinnvoll zu beantworten sein. Aufgrund der technischen Nähe von Websites zu Software mag dies zwar einige verwertbare Ergebnisse bringen – dennoch sollte ein besser passendes, spezifisches und entsprechend validiertes Instrument verwendet werden. Noch weniger Sinn würde es machen, ein psychometrisches Verfahren aus der Persönlichkeitsmessung (z.B. ein Big Five Fragebogen) zur Bestimmung der „Persönlichkeit“ einer Website zu verwenden.

Innovative Methoden

Die in dieser Expertise vorgestellten Verfahren stellen den Standardkanon üblicher Verfahren im Bereich Website-Evaluation dar. Natürlich gibt es in diesem Feld auch eine Reihe von innovativen Methoden, die derzeit erprobt werden. Dabei bleibt allerdings noch abzuwarten, was sich als weitere Standard-Methodik zukünftig etablieren wird. Drei Verfahrenstypen sind hier vielversprechend und sollen daher im Folgenden kurz Erwähnung finden:

1. **Big Data:** Dieser Trend ist seit einiger Zeit in der Diskussion. Eine gestiegene Leistungsfähigkeit von Computern und intelligente Software erlauben, in großen Datenmengen automatisiert nach Mustern zu suchen. In der Website-Evaluation ist hier denkbar, umfassende Nutzungsdaten der User automatisch zu erfassen und zu analysieren. Dies geht dann weit über klassische Logfile-Analysen hinaus, da dabei eine Kombination aller verfügbaren Daten (z.B. Log-Files, Befragungsdaten, personenbezogene Daten, etc.) angestrebt wird. Derartige Analysen haben jedoch zwei zentrale Nachteile: a) Formell sind hier je nach gewählter Analysetiefe Datenschutzbedenken vorzubringen sobald personenbezogene Auswertungen erfolgen, b) inhaltlich muss betont werden, dass Big Data Analysen zwar Datenmuster erkennen können – diese müssen aber sinnvoll interpretiert werden. Die alleinige Existenz eines statistischen Musters erklärt noch lange nicht die inhaltliche Kausalität. Ein bekanntes Beispiel für Scheinkorrelationen (d.h. statistische Zusammenhänge ohne inhaltlich sinnvollen Zusammenhang) ist die signifikante Korrelation zwischen der Zahl der Kindergeburten und der Zahl der Storchenpaare in verschiedenen europäischen Ländern ($r = .62$, $p = .008$, siehe Matthews, 2000).

2. **Experience Sampling:** Klassische Tagebuchmethoden haben aufgrund der technischen Entwicklung eine Renaissance erlebt. Experience Sampling (auch bezeichnet als „ecological momentary assessment“, EMA) erlaubt dabei das Erleben der Probanden in alltäglichen Situationen zu erfassen. Es wird also nicht mehr in der Rückschau ein Gegenstand bewertet oder – wie im Usability-Lab – eine künstliche Testsituation geschaffen. Die Befragten können stattdessen beispielsweise über ein Smartphone direkt Fragen beantworten, sobald eine entsprechende Situation im Alltag auftritt. Denkbar ist zudem im Experience Sampling direkt typische Sensordaten zu nutzen, das heißt beispielsweise den Aufenthaltsort der Befragten über das Smartphone zu erfassen. Weiterhin wird im Gesundheitsbereich überlegt, inwieweit Smartphones (oder andere sogenannte Wearables) auch direkt Patientendaten wie beispielsweise Puls oder Körpertemperatur messen könnten. Datenschutzbedenken sind hier offensichtlich. Dennoch gibt erste Anbieter, die es ermöglichen, Personen (die dem zustimmen) in Evaluationsstudien auch direkt im Alltag zu befragen.

3. **Physiologische Messungen:** Seit langer Zeit werden immer wieder auch physiologische Messungen in der Forschung zur Mensch-Computer Interaktion genutzt. So verwenden beispielsweise Thüring und Mahlke (2007) in ihrer Forschung zum CUE Modell der User Experience Herzratendaten, sowie Daten zur Hautleitfähigkeit (elektrodermale Aktivität, EDA) und Muskelaktivität (Elektromyographie, EMG). Eine direkte Beobachtung derartiger körperlicher Vorgänge ist aus Sicht der Forschung hochinteressant und diese Daten können als Marker für erlebte psychische Vorgänge wie Stress oder Aufregung dienen. In der Website-Evaluation haben sich diese jedoch noch nicht etabliert. Ein zentrales Problem ist hier die Interpretation der Daten insbesondere in Kombination mit subjektiven Befragungsdaten. Oder um es als einfaches Beispiel zu illustrieren: Das körperliche Aktivitätsniveau beispielsweise in der Evaluation der Arbeitstätigkeit eines Rettungssanitäters bietet deutlich mehr Variabilität in den Daten und extremere körperliche Reaktionen als die Rezeption einer Website am Bildschirm. Möglicherweise ergeben sich aber zukünftig interessante Evaluationsverfahren, wenn die Messinstrumente sensibler oder aber die Daten durch Big Data Ansätze sinnvoller aufbereitet werden können. Massive Fortschritte der Forschung im Bereich der Gesichtserkennung zeigen hier weitere mögliche zukünftige Entwicklungen auf, wie beispielsweise die direkte Erkennung von Emotionen der NutzerInnen.

Appendix

Appendix I: Dialogprinzipien nach DIN EN ISO 9210-110

1. Aufgabenangemessenheit

Ein Dialog ist aufgabenangemessen, wenn er den Benutzer unterstützt, seine Arbeitsaufgabe effektiv und effizient zu erledigen.

Umsetzungsbeispiele:

- Ein Eingabefeld erkennt eine fehlerhafte Eingabe automatisch, der Cursor wird direkt in das zu korrigierende Feld gesetzt.
- Auf einer Website sind Ansprechpartner zu finden, so dass BenutzerInnen bei Bedarf eine E-Mail persönlich adressieren kann.
- Zwischenergebnisse einer längeren Online-Transaktion können gespeichert werden.

2. Selbstbeschreibungsfähigkeit

Ein Dialog ist selbstbeschreibungsfähig, wenn jeder einzelne Dialogschritt durch Rückmeldung des Dialogsystems unmittelbar verständlich ist oder dem Benutzer auf Anfrage erklärt wird.

Umsetzungsbeispiele:

- Links sind so formuliert, dass man sicher vorhersagen kann, wohin sie führen.
- Eine Web-Applikation hat eine Online-Hilfe, die kontextspezifische Bedienhinweise gibt.
- Nachdem eine Anfrage an eine Datenbank gesendet wurde, erscheint eine Meldung „Anfrage wird bearbeitet, bitte warten“.

3. Steuerbarkeit

Ein Dialog ist steuerbar, wenn der Benutzer in der Lage ist, den Dialogablauf zu starten sowie seine Richtung und Geschwindigkeit zu beeinflussen, bis das Ziel erreicht ist.

Umsetzungsbeispiele:

- Eine Tabelle hat Buttons, mit deren Hilfe die Informationen spaltenweise sortiert werden können.
- Eine Suchmaschine bietet die Möglichkeit, die Zahl der auf einer Seite angezeigten Treffer einzustellen.
- In einem Eingabefeld gibt es die Möglichkeit, die letzte Eingabe rückgängig zu machen.

4. Erwartungskonformität

Ein Dialog ist erwartungskonform, wenn er konsistent ist und den Merkmalen des Benutzers entspricht, z.B. seinen Kenntnissen aus dem Arbeitsgebiet, seiner Ausbildung und seiner Erfahrung sowie den allgemein anerkannten Konventionen.

Umsetzungsbeispiele:

- Der Link zur Startseite ist unter dem Firmenlogo oben links platziert.
- Der „Warenkorb“ in einem Online-Shop heißt immer und in allen Zusammenhängen „Warenkorb“.
- Unterstrichene Wörter sind immer Hypertext-Links.

5. Fehlertoleranz

Ein Dialog ist fehlertolerant, wenn das beabsichtigte Arbeitsergebnis trotz erkennbar fehlerhafter Eingaben entweder mit keinem oder mit minimalem Korrekturaufwand seitens des Benutzers erreicht werden kann.

Umsetzungsbeispiele:

- Beim Rückwärtsbrowsen in einer Web-Applikation mit der Back-Taste wird die Information immer aktualisiert, sodass nicht fälschlicherweise der Eindruck entsteht, Bearbeitungsschritte seien rückgängig gemacht worden.
- Über ein Skript werden die Daten eines Formulars auf Plausibilität, fehlende oder unvollständige Eingaben geprüft, bevor Sie abgesendet werden.
- Fehlermeldungen werden nicht technisch verklausuliert oder als Nummer angezeigt, sondern in der Sprache der BenutzerInnen formuliert.

6. Individualisierbarkeit

Ein Dialog ist individualisierbar, wenn das Dialogsystem Anpassungen an die Erfordernisse der Arbeitsaufgabe sowie an die individuellen Fähigkeiten und Vorlieben des Benutzers zulässt.

Umsetzungsbeispiele:

- In einem personalisierten Web-Portal kann man festlegen, welche Informationen angezeigt werden.
- Ein editierbares Profil ermöglicht es anzugeben, welche News man in einer Mailing-Liste lesen möchte.
- Kunden eines Online-Shops werden von System erkannt und entsprechende Formularfelder automatisch anhand vorhandener Informationen vorausgefüllt.

7. Lernförderlichkeit

Ein Dialog ist lernförderlich, wenn er den Benutzer beim Erlernen des Dialogsystems unterstützt und anleitet.

Umsetzungsbeispiele:

- In einer „Guided Tour“ werden die Benutzer mit Besonderheiten in der Bedienung einer Website vertraut gemacht.
- Im Buchungs-System eines Reiseanbieters besteht die Möglichkeit eine Probebuchung vorzunehmen.
- In einer Sitemap kann man sich ansehen, nach welcher Logik eine Website strukturiert ist.

Appendix II: Gestaltungsaspekte nach DIN EN ISO 9210-151

DIN EN ISO 9210-151 (ISO, 2006b) orientiert sich an einem Referenzmodel für mensch-zentriertes Design von Web User Interfaces (siehe Abbildung 5). Hierin werden Design-, Prozess- und Evaluationsdomäne in Beziehung gesetzt. Die DIN EN ISO 9210-151 konzentriert sich auf die Designdomäne, nennt aber jeweilige Verknüpfungen zu relevanten anderen ISO-Normen aus den anderen beiden Domänen.
In der Designdomäne unterscheidet die DIN EN ISO 9210-151 fünf relevante Bereiche für das Design einer Website und gibt hier jeweils konkrete Empfehlungen welche Aspekte zu bedenken sind. Dabei werden allerdings in den meisten Fällen die relevanten Teilaspekte zwar benannt (zur Veranschaulichung siehe die folgende Seite), allerdings werden nur vereinzelt Beispiele oder konkrete Handlungsempfehlungen genannt.

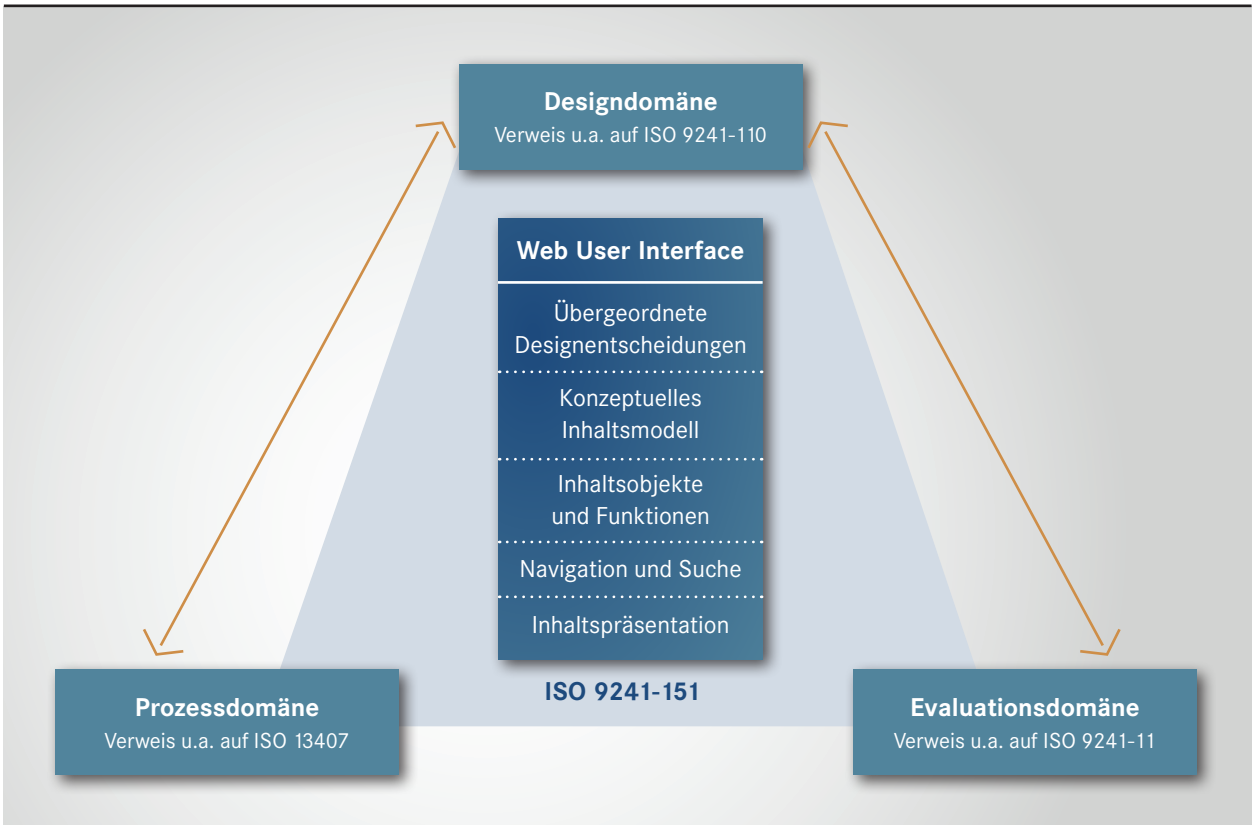


Abbildung 5: Referenzmodel für mensch-zentriertes Design von Web User Interfaces aus DIN EN ISO 9241-151

Level 1: Übergeordnete Designentscheidungen

- Zielfestlegung der Website
- Analyse der Zielgruppe
- Analyse der Ziele und Aufgaben der NutzerInnen
- Übereinstimmung von Websitezielen und Nutzerzielen
- Zielpriorisierung
- ...

Level 2: Konzeptuelles Inhaltsmodell

- Erstellung eines konzeptuellen Inhaltsmodells
- Angemessenheit von Inhalten für die Zielgruppe
- Vollständigkeit der Inhalte
- Angemessene Strukturierung der Inhalte
- Detaillevel der Inhalte
- ...

Level 3: Inhaltsobjekte und Funktionen

- Unabhängigkeit von Inhalt, Website-Struktur und Präsentationsformat
- Auswahl angemessener medialer Objekte
- Aktualität
- Barrierefreiheit
- Feedbackfunktionen
- Datenschutz / Privacy
- Adaptierbarkeit
- ...

Level 4: Navigation und Suche

- Unterstützung der Navigation
- Navigationsstrukur
- Komponenten der Navigation
- Anforderungen an Suchfunktionen
- ...

Level 5: Inhaltspräsentation

- Beachtung grundlegender Wahrnehmungsprinzipien
- Seitendesign (u.a. Platzierung von Elementen, Länge, Farben, ...)
- Gestaltung von Links
- Interaktive Objekte
- Textdesign (u.a. Lesbarkeit, Schreibstil, Qualität, ...)
- Allgemeine Designaspekte (u.a. Sprachversionen, Hilfefunktionen, Geschwindigkeit)
- ...

Appendix III: Evaluationsmodell für neu zu schaffende Website-Inhalte

Website-Inhalte sind für die NutzerInnen zentral – und werden vermutlich kognitiv anders verarbeitet als Designfaktoren wie Usability oder Ästhetik (vgl. Thielsch & Hirschfeld, in press). Falls für eine BZgA-Website Inhalte vollständig neu konzeptioniert und konfiguriert werden, bietet sich an diese in einem eigenen Evaluationsprozess zu prüfen. Hierzu kann das folgende fünfstufige Evaluationsmodell dienen (adaptiert nach Modellierungen von Gollwitzer & Jäger, 2014, Kirkpatrick & Kirkpatrick, 2006 sowie Schenkel, 2000).

Stufe 1: Konzeptions- und Produktionsebene
Wie bewerten Experten die inhaltlichen Entwürfe während der Erstellungsphase, wie bewerten sie das finale Produkt?

Stufe 2: Reaktionsebene
Wie reagieren die Website-NutzerInnen auf die Informationen?

Stufe 3: Lernebene
Findet eine Verbesserung der Kenntnisse der Website-NutzerInnen statt?

Stufe 4: Handlungsebene
Verändert sich das Verhalten der Personen?

Stufe 5: Erfolgsebene
Welche Wirkung hat die Kampagne in der Bevölkerung? Stehen Kosten und Nutzen in einer zufriedenstellenden Relation? Welche Ergebnisse werden insgesamt erzielt?

Auf Stufe 1 findet typischerweise eine begleitende, formative Evaluation mit qualitativen Verfahren statt. Stufe 2 und 3 lassen sich leicht mit verschiedenen qualitativen und quantitativen Verfahren (wie sie in dieser Expertise beschrieben sind) erfassen. Die Evaluation auf höheren Stufen ist deutlich schwerer und wird daher nur selten vorgenommen (vgl. Kirkpatrick & Kirkpatrick, 2006; Schenkel, 2000). Grund ist, dass die Wirkungszusammenhänge nur schwer zu bestimmen sind, da eine Vielzahl anderer Variablen Einfluss nimmt. Sind keine empirisch belastbaren Beweise für Wirkzusammenhänge verfügbar, so sind auf diesen Stufen die vorhandenen Hinweise in verfügbaren Daten zu interpretieren, dabei ist eine entsprechende Vorsicht in der Interpretation zu wahren. Zumindest auf Stufe 4, der Handlungsebene, sind aber durchaus belastbare Aussagen möglich, zum Beispiel durch Evaluationsdesigns im Längsschnitt auf Basis von wiederholten Messungen.

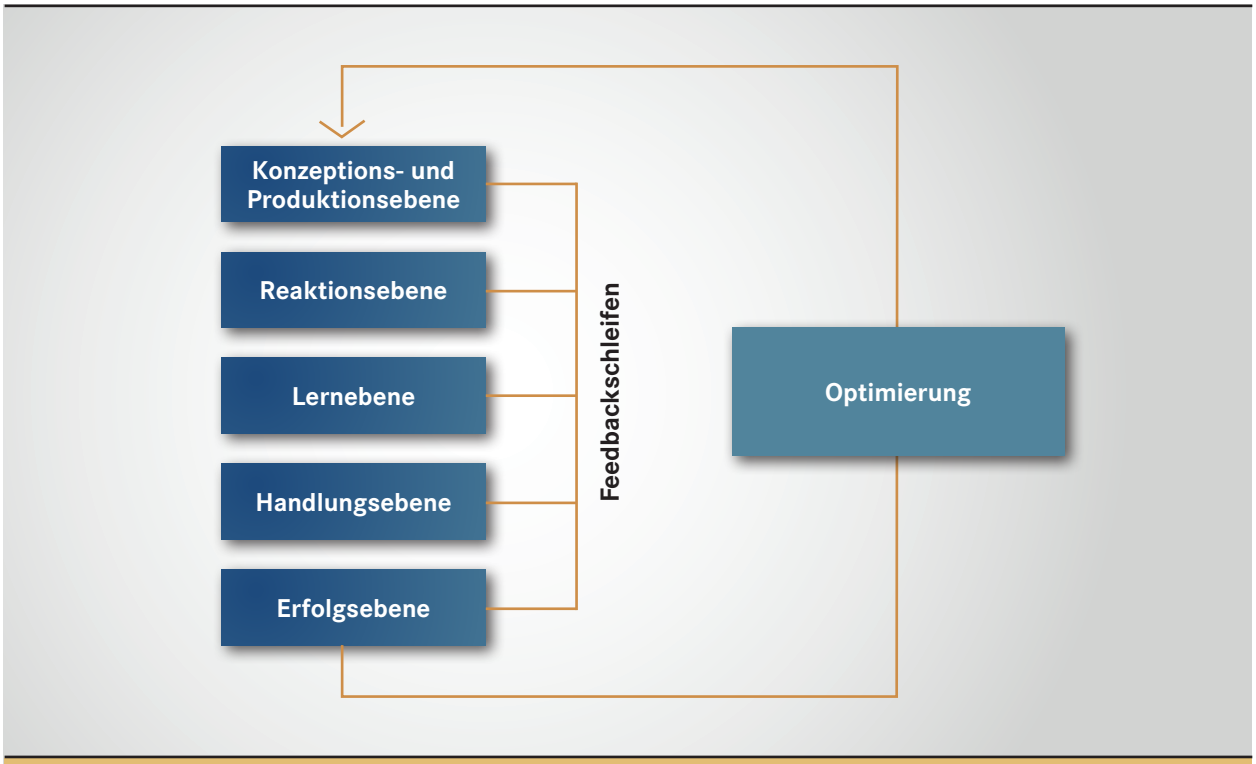


Abbildung 6: Angepasstes Evaluationsmodell für neu zu schaffende Website-Inhalte

Jede Stufe kann für sich evaluiert werden – und damit ergeben sich in jeder Stufe möglicherweise auch Ideen für Optimierungen und Verbesserungen. Berühren dabei die Ergebnisse aus höheren Stufen auch die Konzeptionsebene, so ist eine Überarbeitung der Web-Inhalte notwendig. Aus einer solchen Überarbeitung kann dabei die Notwendigkeit einer erneuten Evaluation resultieren.

Anhang

Literaturverzeichnis

An, R., & Sturm, R. (2012). School and Residential Neighborhood Food Environment and Dietary Intake among California Children and Adolescents. *Amercian Journal of Preventive Medicine*, *42*(2), 129–135. http://doi.org/10.1016/j.amepre.2011.10.012.School

Appelman, A., & Sundar, S. S. (2016). Measuring Message Credibility: Construction and Validation of an Exclusive Scale. *Journalism & Mass Communication Quarterly*, *93*(1), 59–79. http://doi.org/10.1177/1077699015606057

Balzer, L. (2005). *Wie werden Evaluationsprojekte erfolgreich? Ein integrierender theoretischer Ansatz und eine empirische Studie zum Evaluationsprozess*. Landau: Verlag Empirische Pädagogik.

Bardzell, S., & Churchill, E. F. (2011). IwC special issue "Feminism and HCI: new perspectives" Special Issue Editors' introduction. *Interacting with Computers, 23*(5), iii-xi. http://doi.org/10.1016/S0953-5438(11)00089-0

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International journal of Research in Marketing, 13*(2), 139-161.

Bölte, J., Hösker, T., Hirschfeld, G. & Thielsch, M. T. (2017). Electrophysiological correlates of aesthetic processing of webpages: A comparison of experts and laypersons. *PeerJ*, 5:e3440. http://dx.doi.org/10.7717/peerj.3440

Bosnjak, M., Galesic, M., & Tuten, T. (2007). Personality determinants of online shopping: Explaining online purchase intentions using a hierarchical approach. *Journal of Business Research*, *60*(6), 597–605. http://doi.org/10.1016/j.jbusres.2006.06.008

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4-7. http://doi.org/10.1002/hbm.20701

Bühner, M. (2010). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München: Pearson Studium.

Chadwick-Dias, A., McNulty, M., & Tullis, T. (2003, November). Web usability and age: how design changes can improve performance. In *ACM SIGCAPH Computers and the Physically Handicapped* (No. 73-74, pp. 30-37). ACM.

Cober, R.T., Brown, D.A., Levy, P.E., Cober, A.B. & Keeping, L.M. (2003). Organizational web sites: Web site content and style as determinants of organizational attraction. *International Journal of Selection and Assessment, 11*(2/3), 158-169. http://doi.org/10.1111/1468-2389.00239

Cohen, J. (1960). A coefficient of a for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Coiro, J. (2011). Predicting Reading Comprehension on the Internet. *Journal of Literacy Research*, *43*(4), 352–392. doi:10.1177/1086296X11421979

Cyr, D., & Bonanni, C. (2005). Gender and website design in e-business. *International Journal of Electronic Business*, *3*(6), 565-582.

Cyr, D., Head, M., Larios, H., & Pan, B. (2009). Exploring Human Images in Website Design: A Multi-Method Approach. *MIS Quarterly, 33*(3), 539–566.

Cyr, D., Head, M., & Larios, H. (2010). Colour appeal in website design within and across cultures: A multi-method evaluation. *International Journal of Human-Computer Studies, 68*(1–2), 1–21. http://doi.org/10.1016/j.ijhcs.2009.08.005

Dames, H., Hirschfeld, G., Sackmann, T. & Thielsch, M. T. (under review). Browsing vs. Searching - Exploring the influence of consumers' goal directedness on website evaluation.

DeGEval – Gesellschaft für Evaluation e.V. (2016) (Hrsg.): *Standards für Evaluation: Erste Revision auf Basis der Fassung 2002*. Mainz: DeGEval – Gesellschaft für Evaluation e.V.. Verfügbar via http://www.degeval.de/degeval-standards/

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation* (5. Aufl.). Heidelberg: Springer.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of\nincreased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, *35*(3), 379–383.

Flavián, C., Guinalíu, M., & Gurrea, R. (2006). The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Information & Management*, *43*(1), 1–14. http://doi.org/10.1016/j.im.2005.01.002

Fletcher, R. (2006). The impact of culture on web site content, design, and structure: An international and a multicultural perspective. *Journal of Communication Management*, *10*(3), 259–273. http://doi.org/10.1108/13632540610681158

Gediga, G., & Hamborg, K. (2002). Evaluation in der Software-Ergonomie: Methoden und Modelle im Software-Entwicklungsprozess. *Zeitschrift Für Psychologie*, *210*(1), 40–57. http://doi.org/10.1026//0044-3409.210.1.40

Go, E., You, K. H., Jung, E., & Shim, H. (2016). Why do we use different types of websites and assign them different levels of credibility? Structural relations among users' motives, types of websites, information credibility, and trust in the press. *Computers in Human Behavior*, *54*, 231–239. http://doi.org/10.1016/j.chb.2015.07.046

Göritz, A. (2014). Online-Panels. In: M. Welker, M. Taddicken, J. H. Schmidt & N. Jackob (Hrsg.). *Handbuch Online-Forschung. Sozialwissenschaftliche Datengewinnung und -auswertung in digitalen Netzen* (S. 104–122). Köln: Halem.

Gollwitzer, M. & Jäger, R. S. (2014). *Evaluation kompakt* (2. Aufl.). Weinheim: Beltz.

Grisaffe, D. B. (2007). Questions about the ultimate question: conceptual considerations in evaluating Reichheld's net promoter score (NPS). *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, *20*, 36. http://doi.org/10.1016/j.jcps.2014.06.001

Hassenzahl, M., Diefenbach, S., & Göritz, A. (2010). Needs, affect, and interactive products–Facets of user experience. *Interacting with computers, 22*(5), 353-362.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, *64*(2), 79–102. http://doi.org/10.1016/j.ijhcs.2005.06.002

ISO (1998). *ISO 9241: Ergonomic requirements for office work with visual display terminals, VDTS) – Part 11: Guidance on usability.* Geneva: International Organization for Standardization.

ISO (2006a). *ISO 9241: Ergonomics of Human-System Interaction – Part 110: Dialogue principles.* Geneva: International Organization for Standardization.

ISO (2006b). *ISO 9241: Ergonomics of Human-System Interaction – Part 151: Guidance on World Wide Web Interfaces.* Geneva: International Organization for Standardization.

ISO (2009). *9241-210: 2010. Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems.* Geneva: International Organization for Standardization.

Iten, G. H., Troendle, A., & Opwis, K. (in press). Aesthetics in Context–The Role of Aesthetics and Usage Mode for a Website's Success. *Interacting with Computers*.

Jacobsen, J. & Meyer, L. (2017). *Praxisbuch Usability und UX*. Bonn: Rheinwerk Verlag.

Jäger, R. (2004). Konstruktion einer Ratingskala mit Smilies als symbolische Marken. *Diagnostica*, *50*(1), 31–38. http://doi.org/10.1026/0012-1924.50.1.31

Jaspers, M. W. M. (2009). A comparison of usability methods for testing interactive health technologies: Methodological aspects and empirical evidence. *International Journal of Medical Informatics*, *78*(5), 340–353. http://doi.org/10.1016/j.ijmedinf.2008.10.002

Keiningham, T. L., Cooil, B., Andreassen, T. W., & Aksoy, L. (2007). A longitudinal examination of net promoter and firm revenue growth. *Journal of Marketing*, *71*(3), 39-51. http://doi.org/10.2307/30163980

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programms: The four levels* (3rd ed.) an overview. San Francisco, CA: Berret-Koehler.

Koch, W., & Frees, B. (2017). ARD/ZDF-Onlinestudie 2017: Neun von zehn Deutschen online. *Media Perspektiven*, 9/2017, 434–446.

Krumm, S., Stenzel, N. & Pauls, C. A. (2015). Diagnostische Interviews: Planung, Gesprächsführung und Auswertung. In G. Stemmler & J. Margraf-Sticksrud (Hrsg.), *Lehrbuch Psychologische Diagnostik* (S. 77-155). Berlin: Springer.

Kurosu, M., & Kashimura, K. (1995). Apparent usability vs. inherent usability: experimental analysis on the determinants of the apparent usability. In *Conference companion on Human factors in computing systems* (pp. 292–293). ACM.

Lee, H. (2012). The role of local food availability in explaining obesity risk among young school-aged children. *Social Science and Medicine*, *74*(8), 1193–1203. http://doi.org/10.1016/j.socscimed.2011.12.036

Lee, S., & Koubek, R. J. (2012). Users' perceptions of usability and aesthetics as criteria of pre- and post-use preferences. *European Journal of Industrial Engineering*, *6*(1), 87–117. http://doi.org/10.1504/EJIE.2012.044812

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013, April). UMUX-LITE: when there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2099-2102). ACM.

Liljegren, E. (2006). Usability in a medical technology context assessment of methods for usability evaluation of medical equipment. *International Journal of Industrial Ergonomics*, *36*(4), 345–352. http://doi.org/10.1016/j.ergon.2005.10.004

Lindgaard, G., Fernandes, G., Dudek, C., & Browñ, J. (2006). Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & Information Technology*, *25*(2), 115–126. http://doi.org/10.1080/01449290500330448

Liu, C., White, R. W., & Dumais, S. (2010). Understanding web browsing behaviors through Weibull analysis of dwell time. *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '10* (379-386. http://dx.doi.org/10.1145/1835449.1835513

Macefield, R. (2009). How To Specify the Participant Group Size for Usability Studies: A Practitioner 's Guide. *Journal of Usability Studies*, *5*(1), 34–45. Retrieved from http://www.usabilityprofessionals.org/upa_publications/jus/2009november/JUS_Macefield_Nov2009.pdf

Mahatody, T., Sagar, M., & Kolski, C. (2010). State of the art on the cognitive walkthrough method, its variants and evolutions. *International Journal of Human-Computer Interaction*, *26*(8), 741–785. http://doi.org/10.1080/10447311003781409

Matthews, R. (2000). Storks deliver babies (p= 0.008). *Teaching Statistics*, *22*(2), 36–38. http://doi.org/10.1111/1467-9639.00013

Mayring, P. (2010). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (11. Aufl.). Weinheim: Beltz.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334–359. http://doi.org/10.1287/isre.13.3.334.81

Moosbrugger, H. & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion* (2. Aufl.). Heidelberg: Springer.

Moshagen, M. & Thielsch, M. T. (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies, 68*(10), 689-709. http://dx.doi.org/10.1016/j.ijhcs.2010.05.006

Moshagen, M. & Thielsch, M. T. (2013). A short version of the visual aesthetics of websites inventory. *Behaviour & Information Technology, 32*(12), 1305-1311. http://dx.doi.org/10.1080/0144929X.2012.694910

Nielsen, J. (1993). *Usability engineering*. Boston: AP Professional.

Pforr, K. (2015). Incentives. Mannheim, GESIS – Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines). http://dx.doi.org/10.15465/gesis-sg_001

Prümper, J. (1997). Der Benutzungsfragebogen ISONORM 9241/10: Ergebnisse zur Reliabilität und Validität. In R. Liskowsky, B.M. Velichkovsky & W. Wünschmann (Hrsg.). *Software-Ergonomie '97-Usability Engineering: Integration von Mensch-Computer-Interaktion und Software-Entwicklung* (S. 253-262). Berlin: Springer. Verfügbar via http://people.f3.htw-berlin.de/Professoren/Pruemper/instrumente.html

Reichheld, F. F. (2003). The One Number You Need to Grow. *Harvard Business Review*, *81*(12), 46–54+124. http://doi.org/10.1111/j.1467-8616.2008.00516.x

Robbins, S., & Stylianou, A. (2003). Global corporate websites: An empirical investigation of content and design. *Information & Management*, *7206*(40), 205–212. http://doi.org/10.1016/S0378-7206(02)00002-2

Robins, D., & Holmes, J. (2008). Aesthetics and credibility in web site design. *Information Processing & Management*, *44*(1), 386–399. http://dx.doi.org/10.1016/j.ipm.2007.02.003

Rotondi, A. J., Sinkule, J., Haas, G. L., Spring, M. B., Litschge, C. M., Newhill, C. E., … Anderson, C. M. (2007). Designing Websites for Persons With Cognitive Deficits: Design and Usability of a Psychoeducational Intervention for Persons With Severe Mental Illness. *Psychological Services*, *4*(3), 202–224. http://doi.org/10.1037/1541-1559.4.3.202

Rotondi, A. J., Eack, S. M., Hanusa, B. H., Spring, M. B., & Haas, G. L. (2015). Critical Design Elements of E-Health Applications for Users with Severe Mental Illness: Singular Focus, Simple Architecture, Prominent Contents, Explicit Navigation, and Inclusive Hyperlinks. *Schizophrenia Bulletin*, *41*(2), 440–448. http://doi.org/10.1093/schbul/sbt194

Salaschek, M., Holling, H., Freund, P. A., & Kuhn, J.-T. (2007). Benutzbarkeit von Software: Vor- und Nachteile verschiedener Methoden und Verfahren. *Zeitschrift Für Evaluation*, *6*(2), 247–276.

Sarodnick, F., & Brau, H. (2015). *Methoden der Usability Evaluation: Wissenschaftliche Grundlagen und praktische Anwendung.* Göttingen: Hogrefe.

Sauer, J., Sonderegger, A., Heyden, K., Biller, J., Klotz, J. & Uebelbacher, A. (under review). Extra-laboratorial usability tests: an empirical comparison of remote and classical field testing with lab testing.

Schenkel, P. (2000). Ebenen und Prozesse der Evaluation. In: P. Schenkel, S.-O. Tergan & A. Lottmann (Hrsg.) *Qualitätsbeurteilung multimedialer Lern-und Informationssysteme. Reihe multimediales Lernen in der Berufsbildung* (S. 52-74). Nürnberg: BW Bildung und Wissen Verlag und Software GmbH.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*(2), 199.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. http://doi.org/10.1016/j.jrp.2013.05.009

Sharp, B. (2008). Net promoter score fails the test. *Marketing research*, *20*(4), 28-30.

Sillence, E., Briggs, P., Harris, P. R., & Fishwick, L. (2007). How do patients evaluate and make use of online health information? *Social Science & Medicine*, *64*(9), 1853–62. http://doi.org/10.1016/j.socscimed.2007.01.012

Sonderegger, A., Schmutz, S., & Sauer, J. (2016). The influence of age in usability testing. *Applied Ergonomics, 52,* 291–300. http://doi.org/10.1016/j.apergo.2015.06.012

Tan, W.-S., Liu, D., & Bishu, R. (2009). Web evaluation: Heuristic evaluation vs. user testing. *International Journal of Industrial Ergonomics*, *39*(4), 621–627. http://doi.org/10.1016/j.ergon.2008.02.012

Thielsch, M. T. (2017). *Ästhetik von Websites. Wahrnehmung von Ästhetik und deren Beziehung zu Inhalt, Usability und Persönlichkeitsmerkmalen* (2. Aufl.). Lengerich: Pabst Science Publisher.

Thielsch, M. T. (unter Mitarbeit von Salaschek, M.) (2017). *Toolbox zur kontinuierlichen Website-Evaluation und Qualitätssicherung (Version 2.0)*. Arbeitsbericht, Köln: Bundeszentrale für gesundheitliche Aufklärung (BZgA). http://dx.doi.org/10.17623/BZGA:224-2.0

Thielsch, M. T., Blotenberg, I. & Jaron, R. (2014). User evaluation of websites: From first impression to recommendation. *Interacting with Computers, 26* (1), 89-102. http://dx.doi.org/10.1093/iwc/iwt033

Thielsch, M. T., Engel, R. & Hirschfeld, G. (2015). Expected usability is not a valid indicator of experienced usability. *PeerJ Computer Science*, 1:e19. http://dx.doi.org/10.7717/peerj-cs.19

Thielsch, M. T. & Hirschfeld, G. (2012). Spatial frequencies in aesthetic website evaluations – explaining how ultra-rapid evaluations are formed. *Ergonomics, 55* (7), 731-742. http://dx.doi.org/10.1080/00140139.2012.665496

Thielsch, M. T. & Hirschfeld, G. (in press). Facets of website content. *Human-Computer Interaction.* http://dx.doi.org/10.1080/07370024.2017.1421954

Thielsch, M. T. & Thielsch, C. (2018). Depressive symptoms and web user experience. *PeerJ* 6:e4439. http://dx.doi.org/10.7717/peerj.4439

Thüring, M., & Mahlke, S. (2007). Usability, aesthetics and emotions in human–technology interaction. *International Journal of Psychology*, *42*(4), 253–264. http://doi.org/10.1080/00207590701396674

Tractinsky, N., Cokhavi, A., Kirschenbaum, M., & Sharfi, T. (2006). Evaluating the consistency of immediate aesthetic perceptions of web pages. *International Journal of Human - Computer Studies*, *64*(11), 1071–1083.

Tretter, S., Diefenbach, S., Ullrich, D. & Gerber, N., (2017). Branchenreport UX/Usability 2017. In: Hess, S. & Fischer, H. (Hrsg.), Mensch und Computer 2017 - Usability Professionals. Regensburg: Gesellschaft für Informatik e.V.. https://doi.org/10.18420/muc2017-up-0207

Tuch, A. N., Bargas-Avila, J. A., & Opwis, K. (2010). Symmetry and aesthetics in website design: It's a man's business. *Computers in Human Behavior, 26*(6), 1831–1837. http://doi.org/10.1016/j.chb.2010.07.016

US Department of Health and Human Services (2006). The research-based web design & usability guidelines. Available online at https://webstandards.hhs.gov/guidelines/, Print version at https://www.usability.gov/sites/default/files/documents/guidelines_book.pdf

Vermeeren, A., Law, E., Roto, V., Obrist, Marianna Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: current state and development needs. *Proceedings: NordiCHI 2010*, 521–530. http://doi.org/10.1145/1868914.1868973

Wagner, N., Hassanein, K., & Head, M. (2014). The impact of age on website usability. *Computers in Human Behavior*, *37*, 270-282.

Zhao, W., Massey, B., Murphy, J., & Fang, L. (2003). Cultural Dimensions of Website Design and Content. *Prometheus*, *21*(1), 74–84. http://doi.org/10.1080/0810902032000051027

Autorenprofile

PD Dr. **Meinald T. Thielsch**, Dipl.-Psych.; seit 2004 Mitarbeiter am Institut für Psychologie der Westfälischen Wilhelms-Universität Münster; dort 2008 Promotion zur „Ästhetik von Websites“ (Hauptfach Psychologie, Nebenfach Wirtschaftsinformatik) und 2013 Habilitation im Themenfeld „Mensch-Computer Interaktion“. Seit 2014 Akademischer Rat in der Organisations- und Wirtschaftspsychologie an der Universität Münster. Forschungsschwerpunkte liegen in den Bereichen Human-Computer Interaction und User Experience, Wirtschaftspsychologie, Evaluation und Online-Forschung. Parallel zur wissenschaftlichen Arbeit seit 2005 verschiedene nebenberufliche Tätigkeiten vor allem im Bereich User-Experience und Online-Forschung; seit 2014 Mitglied im Vorstand der Deutschen Gesellschaft für Online-Forschung e.V., seit 2016 wissenschaftliche Beratung der BZgA. Weitere Informationen finden sich unter www.meinald.de.

Prof. Dr. **Gerrit Hirschfeld**, Dipl.-Psych.; seit 2014 Professor für Quantitative Methoden an der Hochschule Osnabrück. Befasst sich im Rahmen von drittmittel-geförderten (BMBF, DFG, VW) Projekten mit Methoden um Schwellenwerte für diagnostische Verfahren in Psychologie und Psychosomatik zu entwickeln und so deren Interpretierbarkeit zu verbessern. Daneben testet er die querschnittliche und längsschnittliche Stabilität von psychologischen Instrumenten um deren Vergleichbarkeit zu untersuchen. Davor war er drei Jahre Post-Doc am Deutschen Kinderschmerzzentrum an der Kinderklinik Datteln. Dort plante und supervidierte er klinische und diagnostische Studien zu chronischen Schmerzen bei Kindern. Neben seiner wissenschaftlichen Arbeit gestaltet er seit 2003 Schulungen zur Anwendung statistischer Verfahren für Psychologen, Mediziner, Gesundheitswissenschaftler, Hebammen, Agrar- und Medienwissenschaftler. Gemeinsam mit Meinald Thielsch hat er das Blog surefoss.org initiiert, das versucht den Einsatz von Open Source Software in der Forschung zu erleichtern.

Haftungsausschluss

Die Informationen in dieser Expertise wurden nach besten Wissen und dem aktuell verfügbaren Stand der Forschung zusammengestellt. Marktdaten, insbesondere zu üblichen Verfahrenskosten, wurden zum einen auf Basis der jeweils angegeben Quellen benannt. Zum anderen erfolgte hierzu eine Abfrage über mindestens zwei unabhängige Akteure im Markt (im Februar 2018). Soweit zum Zeitpunkt der Erstellung bekannt, wurden Hinweise auf weitere relevante Entwicklungen gegeben. Zukünftige Veränderungen und methodische Innovationen sind im Bereich der Website-Evaluation / User Experience zu erwarten. Eine Gewährleistung für die inhaltliche Richtigkeit der Dokumente, die dem Auftraggeber übergeben wurden, sowie für die Eignung der Instrumente für den von NutzerInnen intendierten Zweck wird nicht übernommen. NutzerInnen dieser Expertise stellen den Auftragnehmer / die Autoren von der Haftung für Ansprüche Dritter frei, die aufgrund einer fahrlässig oder vorsätzlich erfolgten unzulässigen Nutzung des Dokuments geltend gemacht werden.