rapid UX-score: Modulare und Adaptive Messung von User Experience

Marc Busch
rapid user feedback GmbH
Wien, Österreich
marc.busch@user-feedback.at

Christine Kipke, Marco della Schiava
rapid user feedback GmbH
Wien, Österreich
christine.kipke@user-feedback.at,
marco.dellaschiava@user-feedback.at

Benedikt Salzbrunn
Fachhochschule Technikum Wien
Wien, Österreich
benedikt.salzbrunn@technikum-wien.at

ABSTRACT

Mehrere User Experience (UX)-Faktoren im Rahmen von User-Tests zu messen ist nicht trivial: Existierende Messinstrumente sind überwiegend als *Standalone*-Lösungen konzipiert, die aufgrund ihrer unterschiedlichen Item-Konstruktion und Auswerteverfahren nur schwer miteinander vergleichbar sind. Auch für die Benutzer*innen ist die Beantwortung verschiedener Item- und Skalen-Typen kognitiv herausfordernd. Schwerwiegend kommt hinzu, dass die meisten UX-Messinstrumente pro Faktor viele verschiedene Items vorgeben, was zu zeitlichen Problemen in der Durchführung beiträgt.

rapid UX-score soll diese Probleme lösen und ein modulares UX-Instrument entwickeln, bei welchem die Items in adaptiver Weise vorgegeben werden, was in anderen Anwendungsfeldern schon zu stark verkürzten Messinstrumenten geführt hat. Der Einsatz des Rasch-Modells im Bereich der User Experience-Messung ist innovativ, da existierende UX-Messinstrumente nach Prinzipien der Klassischen Testtheorie aufgebaut sind.

Unser Ziel ist ein modularisiertes Messinstrument zur flexiblen und schnellen Messung der Faktoren Usability, Zufriedenheit, Ästhetik, Emotionales- und Flow-Erleben, sowie der subjektiven Preis-Sensitivität. Ziel ist, dass für jeden Faktor nicht mehr als vier Items für eine ausreichend genaue Messung benötigt werden.

Das Messinstrument rapid UX-score soll für die Vorgabe für Business- und Consumer-Produkte konzipiert werden und eine Vergleichsdatenbank mit ähnlichen (digitalen) Produkten enthalten. Die Entwicklung wird auf Open Science- und Open Source-Prinzipien fußen.

KEYWORDS

UX Messung; Modulare UX Messung; Adaptive UX Messung; Rasch-Modell.

1 Evolution der UX-Messung

Seit den 1980er Jahren ist das akademische und angewandte Feld der Mensch-Maschine Interaktion im Wandel: War dieser Fachbereich in den frühen Anfängen von Desktop-Interaktion, überwiegend behavioralen Modellen und Annahmen geprägt, entwickelte sich Mensch-Maschine Interaktion durch insgesamt drei Wellen hin zu einem stark interdisziplinären Feld, welches längst nicht mehr nur auf effiziente und effektive Interaktion fokussiert: Soziale, komplexe Dynamiken menschlicher Interaktion mit neuen Technologien werden mit unterschiedlichen Forschungsmethoden untersucht [3]. Hedonische Aspekte, Werte und Selbstverwirklichung haben eine mechanistische und kognitiv-behaviorale Sichtweise auf technologische Artefakte und interaktive Systeme abgelöst [5].

Seit den 2000er Jahren haben Paradigmen und Praktiken wie User-Centered Design [1] an Beliebtheit gewonnen und halten Einzug in die *Mainstream*-Technologieentwicklung. Zunächst lag ein Schwerpunkt in Wissenschaft und Anwendung auf Gebrauchstauglichkeit eines Systems (der pragmatischen Qualität, auch *Usability*). Nach ISO 9241 meint dies die Effizienz, Effektivität und die Zufriedenheit bei der Benutzung eines interaktiven Systems. Effektivität und Effizienz wurden zunächst überwiegend durch direkte Ansätze erfasst: Beispiele sind die Messung des Erfolgs, d.h. ob eine Benutzerin eine Aufgabe nach je nach Anwendungsfall definierten Kriterien erfolgreich abgeschlossen hat, und die Messung der Zeit die sie für die „Lösung" der Aufgabe benötigt.

Neben Effektivität und Effizienz schließt die ISO 9241-11 Definition für Gebrauchstauglichkeit auch die Zufriedenheit der Benutzer*innen mit ein [12]. Zufriedenheit ist ein subjektives Konstrukt und eine sogenannte latente Variable, welche über manifeste Indikatoren erfasst wird [16]. Ein Ansatz hierzu ist der Einsatz von Likert-Skalen [13,29]. Das Likert-Verfahren wurde zunächst in der Psychologie für die Messung von Einstellungen eingesetzt und wurde dann für andere Anwendungsfelder adaptiert. Eines der ersten – und bis heute verbreitetsten – Verfahren zur Messung von Usability nach dem Likert-Ansatz ist

die *System Usability Scale* (SUS [7]). Siehe Abbildung 1 für ein Beispiel-Item aus der System Usability Scale.

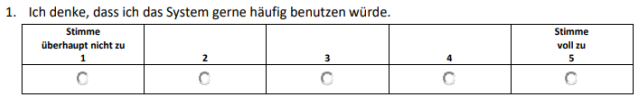

Abbildung 1: Beispiel-Item nach dem Likert-Verfahren aus der System Usability Scale

User Experience (UX) beschreibt das ganzheitliche Erleben und Verhalten vor, während und nach der Interaktion mit digitalen Produkten und inkludiert alle Aspekte der Interaktion mit einem Produkt oder Service [24]. Neben der Messung von Usability ist auch User Experience als psychologisches Konstrukt immer mehr in den Vordergrund gerückt.

Einige weiter verbreitete Instrumente sind neben SUS, AttrakDiff2 noch der *User Experience Questionnaire (UEQ)* [23], UEQ+ [40], VisAWI [42] und meCUE [33], neben weiteren. Diese Instrumente gehören zur Gruppe der Fragebogenverfahren, neben dieser gibt es weitere auf Reaktionsverfahren oder psychophysiologischen Parametern basierende Verfahren. In diesem Paper werden Fragebogenverfahren zur Messung von User Experience und allen darin inkludierten Faktoren und Aspekten kurz *UX-Instrumente* genannt.

UX-Instrumente bauen meist auf zwei Grundannahmen auf: Sie fokussieren meist auf wenige spezielle UX-Faktoren(gruppen). VisAWI bietet beispielsweise ein validiertes Instrument zur Messung der Gruppe der Ästhetik-Faktoren. Zusätzlich sind die Instrumente so konzipiert, dass jedes Item von allen Benutzer*innen bewertet werden muss. Wir stellen die Hypothese auf, dass einige Items für bestimmte Benutzer*innen hinsichtlich des Testwertes („scores") informativer sind, als für andere. Somit zu viele Items vorgegeben werden, als nötig ist.

Instrumente für unterschiedliche Usability oder UX-Faktoren(gruppen) sind auf verschiedene Weisen gestaltet: Dies betrifft sowohl die Konstruktionsweisen der Items (zum Beispiel Statements Ich-Perspektive bei SUS vs. Semantisches Differential bei AttrakDiff2), als auch die Darstellung der Ergebnisse. Sollen mehrere UX-Faktoren(gruppen) gemessen werden, müssen verschiedene UX-Instrumente miteinander kombiniert werden. Dies ist einerseits für die Anwenderin, aber auch für die Versuchsperson zusätzlicher Aufwand und kognitive Belastung. Zudem ergibt sich das Problem der Länge der Instrumente: Zur Messung eines einzelnen UX-Faktors sind oft sehr viele Items notwendig. Um den Faktor Usability mit der System Usability Scale (SUS) zu messen, werden beispielsweise schon 10 Items benötigt.

Die Zeit, die man in User-Tests mit den Personen zur Verfügung hat, soll jedoch meist auch für die Erhebung qualitativer Insights und Barrieren in der Interaktion aufgewendet werden.

2 Moderne UX Messung: rapid UX-score

Als Innovation in der UX Messung schlagen wir ein modulares und adaptives Instrument breiter Faktoren-Gruppen vor. Die Innovation liegt vor allem in der Adaptivität: *Computerized Adaptive Testing* ist in berufsbezogenen Anwendungsfeldern wie der Pilot*innenauswahl schon weit fortgeschritten [8], auch gibt es einzelne Beiträge für Messverfahren in der Persönlichkeitspsychologie [36], jedoch fehlt bislang eine Anwendung in der UX-Forschung.

Es gibt noch kein UX-Messinstrument, welches so konstruiert ist, dass die Items adaptiv vorgegeben werden können. Es sollen damit nicht mehr alle Benutzer*innen alle Items bewerten müssen, die Ergebnisse/Werte sollen jedoch trotzdem vergleichbar sein. Durch insgesamt weniger Items können bei gleicher Zeit mehr UX-Faktoren gemessen werden. Unsere Forschung beschäftigt sich mit der Frage, wie man mit höchstens vier Items einen UX-Faktor mit einer Messgenauigkeit erfassen kann, die zumindest ähnlich oder über der Messgenauigkeit publizierter UX-Instrumente liegt.

Im Folgenden wird der modulare Aufbau von rapid UX-score beschrieben, sowie die Konstruktionsprinzipien vorgestellt, welche die adaptive Vorgabe des Instruments und damit eine signifikante Verkürzung von UX-Messung ermöglichen sollen.

2.1 Module und Items von rapid UX-score

Grundlage für die Entwicklung ist eine Literaturrecherche, um das zu messende Konstrukt gut zu verstehen. Auch Anforderungen von Stakeholdern (Produkt- und Innovationsmanagement, Design, Entwicklung, Marketing) hinsichtlich nachgefragter Faktoren wurden erhoben und flossen gemeinsam mit einem Paper Review in die Entwicklung eines UX-Globalkonstruktes ein (siehe Abbildung 2).

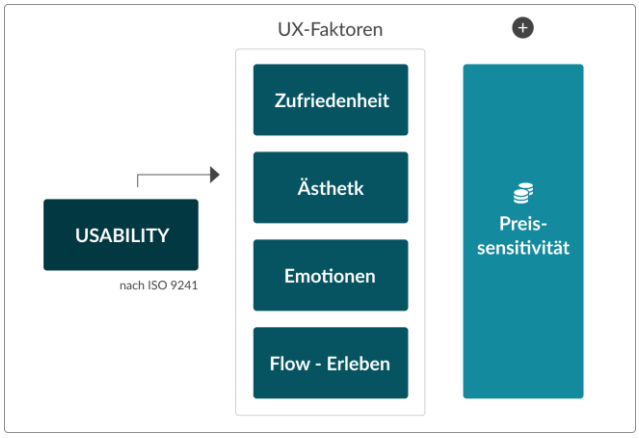

Abbildung 2: UX-Globalkonstrukt als Grundlage für die Item-Entwicklung.

Usability [26,27,39] kommt eine besondere Bedeutung zu: Die subjektive Natur und Messbarkeit dieses Konstrukts wurden

schon sehr gut untersucht. Nach ISO 9241 inkludiert dies die Aspekte Effizienz, Effektivität und Zufriedenheit. Dieser Faktor wird von uns als Grundlage angesehen, um die Ausprägung der anderen Faktoren, die als eigentliche UX-Faktoren zusammengefasst werden, überhaupt zu ermöglichen. Bekannteste Messinstrumente für Usability sind SUS System Usability Scale [7] und UMUX Usability Metric for User Experience [14] bzw. Kurzfassungen wie UMUX-LITE [28].

Zufriedenheit [10,17,20] ist nach ISO 9241 zwar neben Effizienz und Effektivität ein Teil der Usability-Definition, spielt aber in der Marketing- und Verhaltensforschung eine eigenständige Rolle und soll neben Usability als eigenständiger Faktor konzipiert werden. Das verbreitetste Instrument in diesem Bereich ist NPS Net Promoter Score [38].

Ästhetisches Empfinden [25,35] ist ein in der Psychologie und anderen Disziplinen wie der Mensch-Maschine Interaktion schon gut erforschtes Feld, welches eine große Rolle in der Gestaltung und Evaluierung von User Interfaces spielt. Als gut validiertes Referenzinstrument für visuelle Ästhetik wird von uns VisAWI Visual Aesthetics of Websites Inventory herangezogen [42].

Emotionales Erleben, angewandt auf Produkte und Services [4] bezieht sich auf kurzzeitige, gefühlsartige Reaktionen auf innere oder äußere Reize, einschließlich der Interaktion mit User Interfaces oder der Verarbeitung von Informationen und Erlebnissen mit diesen Produkten und Services. Beispielsweise können die Faktoren Freude, Erregung und Dominanz gemessen werden. Verbreitete, angewandte Messansätze sind Self-Assessment Manikins [6] und PrEmo Product Emotion Measurement Instrument [11].

Flow [9,41] ist ein lange erforschtes Phänomen und bezeichnet einen Zustand in einer optimalen Balance zwischen eigenen Fähigkeiten und Anforderungen externer Aufgaben, jenseits von Unter- oder Überforderung ("optimal experience"). Für die Messung von Flow existieren zahlreiche validierte Messverfahren, zum Beispiel [37].

Weil dies eine Anforderung aus Stakeholder-Interviews ist wird ein Zusatz-Modul für **Preis-Sensitivität** entwickelt. Dies bezeichnet den Effekt, den der Preis eines (digitalen) Produktes oder Services auf das Erleben und Verhalten der Benutzer*innen hat [18]. Ursprünglich hat dieses Konstrukt eine größere Bedeutung bei Produkten für Endkonsument*innen, spielt jedoch gerade bei Software as a Service-Anwendungen eine immer wichtigere Rolle. Ein bekanntes Verfahren zur Messung im Endkonsument*innen-Bereich ist Van Westendorp's *price sensitivity meter* [43].

Es werden mit Expert*innen- Item-Pools auf Basis der im Globalkonstrukt definierten UX-Faktoren, sowie der vorhandenen wissenschaftlichen Literatur zu einzelnen Feldern, konstruiert. Wir planen pro UX-Faktor zumindest 20 Items in hoher Qualität zu generieren.

Die Items werden als kurze Aussagen oder – wann immer möglich – nur als Adjektive formuliert, um eine kurze und prägnante Messung zu ermöglichen.
Das Antwortformat wird binär gewählt: „Eher ja" und „Eher nein". Erstens deshalb, weil Skalen mit einer zweistufigen Ausprägung mit dem Rasch-Modell analysiert werden können. Zweitens, weil eine Studie zeigen konnte, dass binäre Antwortformate kürzere Beantwortungsdauer ermöglichen können [31] und Drittens, weil sie auch mehr-stufigeren Skalen in Reliabilität und Validität nicht unterlegen sein müssen [32].

2.2 Item-Analysen und Extremgruppenvalidierung

Nach der Item-Generation auf Basis des UX-Globalkonstruktes wird das Instrument iterativ validiert. Dabei bauen wir auf mehreren Bausteinen auf, welche die Ansätze der Klassischen Testtheorie (KTT; einen guten Überblick über geeignete Methoden gibt [34]) mit der probabilistischen Testtheorie vereinen, zu der auch das Rasch-Modell zuzuordnen ist.

Konstruktvalidität nach Methoden der KTT. Dies beinhaltet eine Faktorenanalyse und Analyse der Konvergenzvalidität. Es soll mit Methoden der KTT explorativ und konfirmatorisch (mit 6 Faktoren, siehe Abbildung 2) die Faktorenstruktur überprüft werden. Trennschärfe- und Analysen der Inneren Konsistenz (*Cronbachs Alpha*) ergänzen diese erste Einschätzung der Faktorenstruktur und vorläufig geeigneter Items. Je mehr die Werte von Benutzer*innen aus bereits validierten Messinstrumenten für spezifische Faktoren (siehe Kapitel 2.1.) mit den Ergebnissen aus den Modulen von rapid UX-score übereinstimmen, desto besser kann nach dieser Methode die Konvergenzvaldität von rapid UX-score eingeschätzt werden.

Extremgruppenvalidierung. Vergleichbar mit dem Validierungsansatz des UMUX sollen pro Faktor von rapid UX-score die UX-Messwerte bei zwei – nur in möglichst einem Faktor sich unterscheidenden – Prototypen miteinander verglichen werden. Anschaulich wird es am Beispiel des Usability-Konstrukts: Eine Studienleiterin wird mit einer Stichprobe Usability-Tests mit zwei Streaming-Diensten durchführt: Mit einem Streaming-Dienst, welcher von Expert*innen anhand definierter Kriterien als gebrauchstauglicher beurteilt wurde als ein anderer Streaming-Dienst. Unsere Hypothese ist, dass rapid UX-score die Ergebnisse dieser heuristischen Evaluierung widerspiegelt.

Rasch-Modell. Nach ersten Analysen mit Methoden aus Klassischer Testtheorie, werden Items ausgeschieden, die einen Faktor nicht unidimensional messen. In weiterer Folge werden Itemschwierigkeitsparameter geschätzt, die zur Entwicklung der Adaptivität des Messinstruments genutzt werden. Mehr dazu in Kapitel 2.3.

Im Einklang mit Forschungsergebnissen [21] werden für die Datenerhebungen pro Analyse relativ große Stichproben (ab n =

500) gewählt, da gerade für Analysen nach der probabilistischen Testtheorie größere Fallzahlen nötig sind, um robuste Ergebnisse und sinnvolle Messmodelle zu erhalten.

Es werden Stichproben aus kommerziellen Online-Panels gezogen. Die Stichproben sollen repräsentativ für Menschen im Alter zwischen 20 und 60 in Österreich hinsichtlich der Merkmale Bildung, Haushaltsnettoeinkommen und Geschlecht sein. Die Personen erhalten marktübliche Aufwandsentschädigungen.

2.3 Adaptivität und Einsatz des Rasch-Modells

Die bisher genannten UX-Instrumente (SUS, UEQ, AttrakDiff2, etc.) haben folgende Gemeinsamkeit: Sie sind nach korrelationsstatistischen Prinzipien der Klassischen Testtheorie konstruiert, was Verfahren wie Faktoren- oder Hauptkomponentenanalysen zur Skalenkonstruktion miteinschließt. Diese Verfahren sind weit verbreitet und werden standardmäßig angewandt, allerdings zeigen Studien, dass die Ergebnisse dieser Klasse von statistischen Verfahren nicht immer robust sind: Sie hängen davon ab, an welcher Stichprobe die Verfahren angewandt werden. Dazu siehe zum Beispiel gesammelte Studien in [22].

Neben „klassischen" Prinzipien wie Faktoren- und Hauptkomponentenanalysen gibt es ergänzende statistische Verfahren zur Entwicklung von Messinstrumenten, die unter dem Begriff Item-Response-Theorie oder probabilistische Testtheorie zusammengefasst werden. Bekannter Vertreter dieser Theorie ist der dänische Statistiker Georg Rasch, der eines der bis heute einflussreichsten Modelle in der Psychometrie (Wissenschaft des Messens psychologischer Konstrukte) entwickelt hat: Das Rasch-Modell [15].

Unter der Item-Response-Theorie wird der Zusammenhang zwischen der Beantwortung eines Items in einem Instrument und der dahinterliegenden, latenten Faktor explizit getestet. Es wird hierzu die Beziehung des Antwortverhaltens der Person und des Faktors modelliert. Es ergibt sich hierdurch ein entscheidender Vorteil des Einsatzes der Item-Response-Theorie zur Skalenkonstruktion: Um zu einem Score/Wert pro Person zu gelangen, müssen nun nicht mehr alle Items von einer Person beantwortet werden. Es ist möglich, nach Beantwortung eines einzelnen Items einen vorläufigen Wert zu errechnen, den sog. Personenparameter einer Person (θ_v). Durch im Vornhinein erhobene Itemschwierigkeitsparameter (zur Benennung siehe weiter unten) eines Items (σi) kann das jeweils nächste Item, welches der Person vorgegeben wird, so gewählt werden, dass der Informationsgewinn bei der Beantwortung maximiert wird. Items, die nur wenig Informationsgehalt haben, müssen so nicht vorgegeben werden.

Der Begriff Itemschwierigkeitsparameter meint in diesem Fall die Zustimmungs- bzw. Ablehnungstendenz zu einzelnen Items in einer Stichprobe. Die Item-Response-Theorie fand bisher überwiegend Anwendung für berufsbezogene psychometrische Anwendung, weshalb hier von „Schwierigkeit" der Items gesprochen wird, es gilt meist Aufgaben nach rein objektiven Kriterien zu lösen.

Die Wahrscheinlichkeit der Antwort von einer Person n_v bei Item x_i kann durch die Modellgleichung des Rasch-Modells bestimmt werden. Löst Person n_v das Item x_i, nimmt X_{vi} den Wert 1 an, löst sie das Item nicht, nimmt X_{vi} den Wert 0 an. exp ist die natürliche Exponentialfunktion:

$$p(X_{vi} = 1) = \frac{\exp(\theta_v - \sigma_i)}{1 + \exp(\theta_v - \sigma_i)}$$

Die unbekannten Parameter θ_v und σi werden über einen Maximum-Likelihood-Ansatz geschätzt. Im Rasch-Modell kann die Schätzung des Personenparameters dann unabhängig von der Schätzung des Itemparameters erfolgen, was ein Vorteil bei diesem Ansatz ist.

3 Umsetzung von rapid UX-score

Die Entwicklung von rapid UX-score hat im Mai 2020 begonnen und ist geplanter Weise im April 2022 abgeschlossen. Datenanalysen werden in R durchgeführt, dezidierte Pakete wie *eRm extended Rasch Modeling* [30] werden zur statistischen Modellbildung verwendet. Das UX-Messinstrument soll als Web-Anwendung zur Verfügung gestellt werden. Die Entwicklung der Web-Anwendung wird mit dem *shiny*-Ökosystem durchgeführt.

Fokus bei der Entwicklung liegt auf sehr guter Anwendbarkeit: Gespräche mit Stakeholder*innen haben gezeigt, dass die Anwendung von UX-Messinstrumenten derzeit sehr aufwändig ist und noch viel manuelle Arbeit erfordert. Messinstrumente und Auswertungsanleitungen liegen oft als Text- bzw. Spreadsheet-Datei vor, der Einsatz erfordert Vorbereitung und Fehler können leicht passieren.

Die Web-Anwendung rapid UX-score soll dabei alle Funktionen im *UX Measurement Life Cycle* übernehmen: Eine einfache und konsistente Auswahl der zu messenden Faktoren (Module), adaptive und damit Items, die vorgegeben werden, sowie Auswertung und Reporting mittels interaktiver Visualisierungen. Wir orientieren uns an etablierten Report-Formaten [2], werden diese durch Inputs von Stakeholder*innen und Expert*innen jedoch entsprechend dem Stand der Technik anpassen.

Die anonymisierten Forschungsdaten sowie der Analyse-Code werden während und nach der Entwicklung auf einem frei zugänglichen GitHub-Repository der Öffentlichkeit zur Verfügung gestellt. Es wird eine Open Source-Lizenz gewählt.

ACKNOWLEDGMENTS

Dieses Projekt wird im Programm Innovation/18 -21+ unter dem Titel "rapid UX-score – Schnelle und zuverlässige Messung von User Experience" (ID: 3132106) von der Wirtschaftsagentur Wien. Ein Fonds der Stadt Wien gefördert. Wir bedanken uns bei der Wirtschaftsagentur Wien und der Jury. Wir danken Astrid Meisslitzer und den Studierenden für die Mitarbeit am Projekt.

REFERENCES

[1] Abras, C., Maloney-Krichmar, D., & Preece, J. (2004). User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*, *37*(4), 445-456.

[2] ANSI. (2001). Common Industry Format for Usability Test Reports.

[3] Bannon, L. J. (1995). From human factors to human actors: The role of psychology and human-computer interaction studies in system design. In *Readings in human–computer interaction* (pp. 205-214). Morgan Kaufmann.

[4] Bosch, C., Schiel, S., & Winder, T. (2006). Emotionsmessung. *Emotionen im Marketing: Verstehen—Messen—Nutzen*, 171-212.

[5] Bødker, S. (2015). Third-wave HCI, 10 years later---participation and sharing. *interactions*, *22*(5), 24-31.

[6] Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49-59.

[7] Brooke, J. (1986). System usability scale (SUS): a quick-and-dirty method of system evaluation user information. *Reading, UK: Digital Equipment Co Ltd*, *43*.

[8] Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*(1), 1-20.

[9] Csikszentmihalyi, M., & Csikszentmihalyi, I. S. (Eds.). (1992). *Optimal experience: Psychological studies of flow in consciousness*. Cambridge university press.

[10] Czeplel, J. A., & Rosenberg, L. J. (1977). Consumer satisfaction: concept and measurement. *Journal of the academy of Marketing Science*, *5*(3), 403-411.

[11] Desmet, P. (2003). Measuring emotion: Development and application of an instrument to measure emotional responses to products. In *Funology* (pp. 111-123). Springer, Dordrecht.

[12] DIN, E. (2016). 9241-11-Ergonomics of human-system interaction-Part 11: Usability: Definitions and concepts.

[13] Edmondson, D. R. (2005, April). Likert scales: A history. In *Proceedings of the 12th conference on historical analysis and research in marketing (CHARM)* (pp. 127-133).

[14] Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323-327.

[15] Fischer, G. H., & Molenaar, I. W. (Eds.). (2012). *Rasch models: Foundations, recent developments, and applications*. Springer Science & Business Media.

[16] Fonseca, J. R. (2009). Customer satisfaction study via a latent segment model. *Journal of Retailing and Consumer Services*, *16*(5), 352-359.

[17] Giese, J. L., & Cote, J. A. (2000). Defining consumer satisfaction. *Academy of marketing science review*, *1*(1), 1-22.

[18] Han, S., Gupta, S., & Lehmann, D. R. (2001). Consumer price sensitivity and price thresholds. *Journal of retailing*, *77*(4), 435-456.

[19] Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & computer 2003* (pp. 187-196). Vieweg+ Teubner Verlag.

[20] Huber, F., Herrmann, A., & Wricke, M. (2001). Customer satisfaction as an antecedent of price acceptance: results of an empirical study. *Journal of Product & Brand Management*.

[21] Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in psychology*, *7*, 109.

[22] Kubinger, K. D. (2006). *Psychologische diagnostik: Theorie und praxis psychologischen diagnostizierens*. Hogrefe Verlag.

[23] Laugwitz, B., Schubert, U., Ilmberger, W., Tamm, N., Held, T., & Schrepp, M. (2009). Subjektive Benutzerzufriedenheit quantitativ erfassen: Erfahrungen mit dem User Experience Questionnaire UEQ. *Tagungsband UP09*.

[24] Law, E. L. C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009, April). Understanding, scoping and defining user experience: a survey approach. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 719-728).

[25] Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British journal of psychology*, *95*(4), 489-508.

[26] Lewis, J. R. (2018). Measuring perceived usability: The CSUQ, SUS, and UMUX. *International Journal of Human–Computer Interaction*, *34*(12), 1148-1156.

[27] Lewis, J. R., & Sauro, J. (2009, July). The factor structure of the system usability scale. In *International conference on human centered design* (pp. 94-103). Springer, Berlin, Heidelberg.

[28] 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).

[29] Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

[30] Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R.

[31] Matell, M. S., & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology*, *56*(6), 506.

[32] Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and psychological measurement*, *31*(3), 657-674.

[33] Minge, M., & Riedel, L. (2013). meCUE-Ein modularer fragebogen zur erfassung des nutzungserlebens. *Mensch & Computer 2013: Interaktive Vielfalt*.

[34] Moosbrugger, H., & Kelava, A. (2012). Testtheorie und Fragebogenkonstruktion.

[35] Müller, K., & Schrepp, M. (2013, August). Visuelle Komplexität, Ästhetik und Usability von Benutzerschnittstellen. In *Mensch & Computer* (pp. 211-220).

[36] Ortner, T. M. (2005). *Möglichkeiten und Grenzen adaptiver Persönlichkeitsfragebogen*. Pabst Science Publishers.

[37] Redaelli, C., & Riva, G. (2011). Flow for Presence Questionnaire. In *Digital Factory for Human-oriented Production Systems* (pp. 3-22). Springer, London.

[38] Reichheld, F. F. (2003). The one number you need to grow. *Harvard business review*, *81*(12), 46-55.

[39] Sauro, J., & Lewis, J. R. (2009, April). Correlations among prototypical usability metrics: evidence for the construct of usability. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1609-1618).

[40] Schrepp, M., & Thomaschewski, J. (2019). Eine modulare Erweiterung des User Experience Questionnaire. *Mensch und Computer 2019-Usability Professionals*.

[41] Takatalo, J., Häkkinen, J., Kaistinen, J., & Nyman, G. (2010). Presence, involvement, and flow in digital games. In *Evaluating user experience in games* (pp. 23-46). Springer, London.

[42] Thielsch, M. T., & Moshagen, M. (2011). Erfassung visueller Ästhetik mit dem VisAWI. *Tagungsband UP11*.

[43] Van Westendorp, P. H. (1976, September). NSS Price Sensitivity Meter (PSM)–A new approach to study consumer perception of prices. In *Proceedings of the 29th ESOMAR Congress* (Vol. 139167).