Woran erkenne ich einen guten User Experience Fragebogen?

Meinald T. Thielsch[1], Gerrit Hirschfeld[2]

Institut für Psychologie, Westfälische Wilhelms-Universität Münster [1]
Quantitative Methoden, Hochschule Osnabrück[2]

thielsch@uni-muenster.de, g.hirschfeld@hs-osnabrueck.de

Zusammenfassung

Der vorliegende Beitrag stellt die Merkmale guter User Experience (UX) Fragebögen anhand von zentralen Kriterien heraus. Neben den klassischen Gütekriterien, insbesondere Reliabilität und Validität, werden hierbei die Grundlagen eines Verfahrens sowie die praktische Anwendung diskutiert. Eine beispielhafte Illustration erfolgt anhand etablierter UX Fragebögen im Spiegel der Erkenntnisse aus Evaluationsforschung und Psychometrie.

1 Einleitung

Die Auswahl von geeigneten Fragebögen im Rahmen einer User Experience (UX) Evaluation ist oftmals nicht einfach. Zudem kommen meist verschiedene qualitative und quantitative Verfahren aus der empirischen Sozialforschung kombiniert zum Einsatz. Der vorliegende Beitrag fokussiert auf Fragebogenverfahren: Woran können AnwenderInnen die Qualität eines Instruments festmachen? Antworten darauf werden im Folgenden in Form eines Überblicks gegeben der über die klassischen Hauptgütekriterien wie Objektivität, Reliabilität und Validität (siehe bspw. Bühner, 2010; Moosbrugger & Kelava, 2012) hinausgeht ohne diese zu vernachlässigen. Die dargestellten Aspekte können damit gleichzeitig eine Orientierung geben, worauf FragebogenautorInnen besonderen Wert legen sollten.

2 Solide Grundlagen: Klare Konstruktdefinition und klarer Anwendungsbereich

Es ist der zentrale erste Schritt: Die AutorInnen eines Fragebogens müssen klar definieren wozu ihr Instrument dienen soll. Welches Konstrukt wird erfasst? Was ist der

Veröffentlicht durch die Gesellschaft für Informatik e. V. 2018 in
R. Dachselt, G. Weber (Hrsg.):
Mensch und Computer 2018 – Workshopband, 02.–05. September 2018, Dresden.
 https://doi.org/10.18420/muc2018-ws16-0489

Anwendungsbereich? Beispiel: Bei der „System Usability Scale“ (SUS; Brooke, 1996) haben die AnwenderInnen direkt ein Gefühl, wozu die Skala geeignet sein könnte – zur Bewertung von Usability von Systemen. Bei der „Usability Metric for User Experience – Lite“ (UMUX-Lite, Lewis et al., 2013) ist das schon ein wenig unklarer. Zumindest ist nicht direkt ersichtlich, dass es sich hierbei um eine (durchaus gut gemachte) Kurzversion der SUS handelt. Beim SUS gibt der Autor eine klare Definition und Bezugspunkte was Usability umfasst (Brooke, 1996), dies fehlt wiederum in der Darstellung der UMUX-Lite. Natürlich kann mit dem Wissen, dass die UMUX-Lite eine SUS-Kurzversion ist, im Originaltext nachgelesen werden; das ist aber umständlich und erschwert die einfache, direkte Anwendung.
Aus der Konstruktdefinition sollten sich zudem die Entwicklung und Zusammenstellung der Befragungsitems direkt ableiten – dies stellt die Grundlage einer hohen inhaltlichen Validität eines Fragebogens dar.

Des Weiteren sollten die AutorInnen klar festlegen, für welche Befragten und welche Inhalte ihr Instrument geeignet ist, damit wird auch klar wo ein Einsatz wahrscheinlich nicht sinnvoll ist. Zum Beispiel: Der Fragebogen zur Erfassung von „Website- Clarity, Likeability, Informativenass, and Credibility“ (Web-CLIC, Thielsch und Hirschfeld, in press) kann laut Autoren für *Jugendliche ab 14 Jahren und Erwachsene* zur Beurteilung von *Website-Inhalten* eingesetzt werden. Es gibt zwar erste Anwendungserfahrungen bei Informationsmedien jenseits von klassischen Websites (bspw. Online-Geschäftsberichte, siehe Thielsch & Wirth, 2017) – aber dennoch ist unmissverständlich, dass der Web-CLIC vorrangig für die Website-Evaluation erstellt und validiert wurde. Für eine generelle Bewertung von interaktiven Produkten, wie beispielsweise einer Smartwatch, ist dieser Fragebogen nicht erstellt worden. Ebenso gilt für einen Einsatz des Web-CLICs in einer Altersgruppe unter 14 Jahren: Es müsste zunächst kritisch geprüft werden, ob der Fragebogen hier überhaupt anwendbar ist.

Wenn keine expliziten Angaben zum Anwendungsbereich eines UX Fragebogens gemacht werden, sollte man kritisch prüfen, inwiefern bisherige Teststichproben und Beurteilungsobjekte der AutorInnen zur eigenen geplanten Untersuchung passen.

3 Erfüllung zentraler Gütekriterien

Ein zentraler Unterschied zwischen klassischen psychometrischen Fragebögen und UX Instrumenten liegt im Beurteilungsgegenstand: Bei einem klassischen psychometrischen Tests, wie beispielweise Intelligenztests, werden verschiedene Personen hinsichtlich eines Merkmals – der Intelligenz – untersucht. Bei einem UX Test werden verschiedene Personen hinsichtlich ihrer Einschätzung eines Merkmals eines interaktiven Produktes befragt – der Fragebogen soll aber für verschiedene Produkte gelten (vgl. dazu auch Gediga et al., 1999). Kann daraus geschlossen werden, dass klassische Gütekriterien nicht im UX Bereich anwendbar sind?

Nein, im Gegenteil: Die Anforderungen an die Validierung eines Fragebogens sind sogar erhöht. Es bedarf nicht nur einer Prüfung der Stichproben, für die der Fragebogen angewandt werden kann, sondern auch für welche Objekte sich der Fragebogen eignet. Wurde ein UX Messinstrument beispielsweise nur an einem einzigen Produkt mit einer rein studentischen

Stichprobe getestet, so sind nur valide Aussagen für Studenten und diese Produktklasse möglich. Soll ein UX Fragebogen einen breiten Einsatzbereich haben, so ist auch eine entsprechend breite Prüfung der Validierung des Instrumentes notwendig.

Für die AnwenderInnen gilt: Ein Fragebogen sollte nur dann zum Einsatz kommen, wenn Hinweise darauf vorliegen, dass er für das Testszenario und die Zielgruppe geeignet ist. Hierzu werden insbesondere die Aussagen der AutorInnen zu Reliabilität und Validität herangezogen.

3.1 Reliabilität

Das Gütekriterium der Reliabilität bezieht sich auf Messgenauigkeit der Merkmalserfassung. In Fragebögen prüft man die Reliabilität typischerweise auf zwei Arten. Zum einen in Hinblick auf den Zusammenhang (Korrelation) aller Items untereinander (die sogenannte interne Konsistenz), dazu wird meist Cronbachs α als Maß angegeben. Werte kleiner .8 gelten hierbei als niedrig, Werte von .8 bis .9 als ausreichend, Werte größer .9 als gut (vgl. Bühner, 2010). Zum anderen kann man die Messwerte aus zwei verschiedenen Messzeitpunkten miteinander korrelieren und so die Stabilität bestimmen (Retest-Reliabilität). Die Korrelationen als Zusammenhangsmaß können zwischen 0 und 1 schwanken, 0 bedeutet keinen Zusammenhang, 1 wäre eine perfekte Übereinstimmung.

Eine alleinige Betrachtung der internen Konsistenz, meist über Cronbachs α, bringt verschiedene Nachteile mit sich. Insbesondere ist Cronbachs α von der Anzahl der Items abhängig, zudem muss eine valide Skala nicht zwingend homogen sein. Weiterhin sollten einen sehr hohe α-Werte bedenklich stimmen: Gibt es bei einem α von beispielsweise .98 überhaupt noch Unterschiede zwischen den Items oder wird dort x-mal das gleiche in minimal veränderten Formulierungen abgefragt? Somit wird klar, Cronbachs α ist kein optimales Maß für Reliabilität (siehe Cronbach, 2004). Fazit: In der Überprüfung von UX Instrumenten müssen weitere Verfahren, wie beispielsweise die Berechnung der Retest-Reliabilität, zum Einsatz kommen.

3.2 Validität

Die Validität zeigt an, in welchem Ausmaß ein Verfahren ausschließlich das Merkmal erfasst, dass es erfassen soll. Rein theoretisch sollte ein valider UX Fragebogen nicht von der Lesegeschwindigkeit oder der allgemeinen Lebenszufriedenheit eines Befragten beeinflusst sein[1].

Es gibt für die Validität eines Verfahrens nicht einen einzelnen nummerischen Wert – Validität wird stets in Bezug auf eine Vielzahl von Untersuchungswegen argumentiert. Im Idealfall kommen mehrere Verfahren zur Anwendung: Man vergleicht beispielsweise einen neuen Fragebogen mit vorhandenen Verfahren, die das gleiche messen (erwartet werden hierbei hohe Korrelationen = konvergente Validität) oder die etwas ganz anderes messen (erwartet werden hierbei niedrige Korrelationen = divergente Validität). Zudem können Expertenurteile und andere vergleichende Kriterien herangezogen (= konkurrente Validität) oder Experimente

[1] Praktisch zeigen sich Einflüsse differentieller Faktoren auf die User Experience, die Effektgrößen sind in bisherigen ersten Studien dabei aber eher klein (vgl. Thielsch & Thielsch, 2018).

durchgeführt werden (= experimentelle Validität). Außerdem kann geprüft werden, ob das Instrument zwischen verschiedenen Zielobjekten unterscheiden kann (= diskriminative Validität) und ob die angenommenen Faktoren mittels einer konfirmatorischen Faktorenanalyse reproduziert werden können (= faktorielle Validität). Die Prüfung der Validität an *mindestens* zwei verschiedenen Stichproben ist ein übliches und sehr sinnvolles Vorgehen, um Stichprobenfehler zu vermeiden.

3.3 Interpretation der Gütekriterien

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, dass ein hoch reliables Verfahren auch objektiv ist. Gezielte Prüfungen der Objektivität finden bei UX Instrumenten oft nicht statt – solange ausreichende Prüfungen der Reliabilität und Validität vorliegen ist dies akzeptabel. Nicht akzeptabel ist, wenn keine Prüfung der Validität erfolgt ist. Liegen keine Informationen zur Validität eines Fragebogens vor, dann sollte das Verfahren nicht zum Einsatz kommen. Hinweise zur Interpretation der Reliabilität wurden oben bereits kurz gegeben, bei der Validität gilt unter anderem:

- Der Fragebogen muss in einem solchen Maß valide sein, dass ein Einsatz besser ist als eine Unterlassung.
- Die Validität sollte bei neuen Fragebögen höher als bei älteren sein oder die Durchführungsökonomie sollte sich bei gleicher Validität verbessern.
- Die Höhe der Validität sollte proportional zur Wichtigkeit der Entscheidung sein.

Die Gütekriterien eines Verfahren sollten zudem an ausreichend großen Stichproben aus der Zielpopulation bestimmt worden sein. Die meisten Gütekriterien, die auf Korrelationskoeffizienten beruhen, können ab einer Stichprobengröße von ca. 250 Probanden hinreichend exakt geschätzt werden (vgl. Schönbrodt & Perugini, 2013). Zur Identifikation stabiler Strukturen in Faktorenanalysen sind durchaus größere Stichproben notwendig (vgl. Costello & Osborne, 2005; Hirschfeld et al., 2014), eine Daumenregel kann hier sein, mindesten 500 Befragte als Berechnungsgrundlage für explorative Faktorenanalysen zu wählen.

Insgesamt zeigt sich damit: Die Interpretation der Güte von UX Fragebögen verlangt Fachwissen – sowohl im Anwendungsbereich UX als auch hinsichtlich der zentralen psychometrischen Kriterien. Am Erwerb dieser Expertise führt kein Weg vorbei. Dies ist insbesondere wichtig, da leider nicht immer die Gütekriterien von UX Fragebögen vollständig bestimmt und eindeutig dargestellt werden.

4 Auswertungs- und Interpretationshinweise

In der Forschung wird in der Regel auf die Hauptgütekriterien Reliabilität und Validität geschaut. Dies ist prinzipiell gut und richtig – aber für die praktische Anwendung sind eine Reihe von Nebengütekriterien und weitere grundsätzliche Informationen relevant:

- **Auswertungshinweise und Fragebogenvorlagen**: Liegen entsprechende Informationen zu einem Instrument vor, so ist die Anwendung deutlich leichter als wenn erst mühselig Items, Skalen- und Antwortformate aus einer Fachpublikation zusammengesucht werden müssen.
- **Benchmarks/Interpretationshilfen**: Für AnwenderInnen ist es enorm hilfreich, wenn die eigenen Testwerte eingeordnet werden können. Dies können Umrechnungshilfen für Fragebogenwerte in Schulnoten sein (wie für die SUS, siehe Tabelle 2 in Lewis et al., 2015), optimale Schwellenwerte (wie für den VisAWI, siehe Hirschfeld & Thielsch, 2015), oder Benchmarks (wie für den Web-CLIC, siehe Thielsch & Hirschfeld, in press).
- **Ökonomie**: Der Nutzen eines Verfahrens sollte auf jeden Fall höher sein als seine Kosten. Dabei stehen im Bereich UX derzeit eine Reihe von Instrumenten aus der Forschung zur freien Verfügung – einziger Lohn für die AutorInnen ist die angemessene Zitation beim Einsatz der Verfahren. In manchen Situationen ist damit die kostenpflichtige Lizensierung kommerzieller Verfahren nur dann gerechtfertigt, wenn diese a) eine höhere Güte haben, b) bei vergleichbarer psychometrischer Güte differenziertere Ergebnisse anbieten, oder c) mit entsprechender hochwertiger Beratungsleistung einhergehen.
- **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. Es gibt psychometrisch sehr sauber konstruierte Fragebogenverfahren, die aber leider sehr viele Items umfassen. Nicht in jeder Situation sind diese anwendbar. Daraus folgt: AutorInnen langer Verfahren können die Anwendung durch entsprechende Kurzversionen erleichtern.

5 Fazit

Die dargestellten Aspekte zu den Grundlagen eines Fragebogenverfahrens im Sinne von Konstruktdefinition, Anwendungsbereich, zu den zentralen Gütekriterien und deren Interpretation sowie zu Rahmenbedingungen der Anwendung in der Praxis, geben einen Rahmen, sowohl zur Beurteilung, als auch zur Konstruktion von UX Instrumenten. Der vorliegende Beitrag kann dabei nur Akzente setzen, welche Aspekte besonderer Betrachtung bedürfen. Eine allgemeine Darstellung zur Bewertung von Test- und Fragebogenverfahren liefert Kersting (2006); eine spezifische Übersicht im UX Bereich der Website-Evaluation Thielsch (2017).

In der Praxis sollen UX Befragungen Entscheidungsprozesse unterstützen – und hierfür möglichst verlässliche Ergebnisse liefern. Fragebögen sollen helfen Ist-Zustände zu beschreiben und zu bewerten. Aber nur wenn dieser gegenwärtige Ist-Zustand treffend und valide beschrieben ist, wird klar ob und welche Maßnahmen notwendig sind. Ist die Zustandsbeschreibung falsch, ist das Risiko erhöht Entscheidungen auf Basis von falschen Annahmen zu treffen. In manchen Projekten kann dies weitreichende negative Konsequenzen mit sich bringen. Ist die Bewertungsgrundlage durch eine Evaluation hingegen hinreichend valide, kann eine optimale UX für die NutzerInnen erreicht werden.

Literaturverzeichnis

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.

Costello, A. B., & Osborne, J. W. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Education*, *10*(7), 1–9. http://doi.org/10.1.1.110.9154

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391–418.

Gediga, G., Hamborg, K.-C., & Duntsch, I. (1999). The IsoMetrics usability inventory: an operationalization of ISO9241-10 supporting summative and formative evaluation of software systems. *Behaviour & Information Technology*, *18*(3), 151–164.

Hirschfeld, G., von Brachel, R. & Thielsch, M. T. (2014). Selecting items for Big Five questionnaires: At what sample size do factor loadings stabilize? *Journal of Research in Personality, 53,* 54-63. http://dx.doi.org/10.1016/j.jrp.2014.08.003

Hirschfeld, G. & Thielsch, M. T. (2015). Establishing meaningful cut points for online user ratings. *Ergonomics, 58*(2), 310-320. http://dx.doi.org/10.1080/00140139.2014.965228

Kersting, M. (2006). Zur Beurteilung der Qualität von Tests: Resümee und Neubeginn. *Psychologische Rundschau*, 57, 243-253

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.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring Perceived Usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction*, *31*(8), 496–505. http://doi.org/10.1080/10447318.2015.1064654

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

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

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. & 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

Thielsch, M. T. & Wirth, M. (2017). Web-based annual reports at first contact: corporate image and aesthetics. *Technical Communication, 64* (4), 282-296.

Autoren

Thielsch, Meinald T.

PD Dr. Meinald T. Thielsch (Dipl.-Psych.) studierte an der Westfälischen Wilhelms-Universität Münster und ist dort seit 2004 am Institut für Psychologie tätig. Seit 2014 ist er Akademischer Rat in der Organisations- und Wirtschaftspsychologie im Bereich „Beratung und Fortbildung für Organisationen“. Als Lehrbeauftragter war er an den Universitäten Bonn und Fribourg (Schweiz) sowie der Fachhochschule Münster aktiv. Seine Arbeits- und Forschungsschwerpunkte sind User Experience, Wirtschaftspsychologie, Forschungs-Praxis-Transfer, Evaluation und Online-Forschung. Weitere Informationen finden sich unter www.meinald.de.

Hirschfeld, Gerrit

Prof. Dr. Gerrit Hirschfeld (Dipl.-Psych.) studierte an der Westfälischen Wilhelms-Universität Münster Psychologie und promovierte anschließend in Biologie. Nach seiner Promotion hat er mehrere Jahre am Deutschen Kinderschmerzzentrum gearbeitete und angewandte Studien zur Diagnostik und Intervention bei chronischen Schmerzen durchgeführt. Seit 2014 ist er Professor für Quantitative Methoden an der Hochschule Osnabrück. Im Rahmen von drittmittel-geförderten Projekten entwickelt er Methoden weiter, um optimale Grenzwerte für diagnostische Instrumente zu bestimmen. Weitere Informationen finden sich unter www.gerrithirschfeld.de.