UX Fragebögen: Verwenden wir die richtigen Methoden?

Martin Schrepp[1], Bernard Rummel[1]

User Experience, SAP SE[1]
martin.schrepp@sap.com, bernard.rummel@sap.com

Zusammenfassung

Fragebögen sind eine weit verbreitete Methode zur Messung von User Experience (UX). Die Methoden zur Konstruktion und Validierung von UX Fragebögen orientieren sich aktuell an den etablierten Methoden der psychologischen Testtheorie. Allerdings gibt es einige prinzipielle Unterschiede zwischen einem UX Fragebogen und einem klassischen psychologischen Test, z.B. einem Persönlichkeitstest. Wir beschreiben diese Unterschiede und diskutieren die Frage, ob man etablierte Konzepte der psychologischen Testtheorie wirklich unverändert auf die Konstruktion von UX Fragebögen übertragen kann.

1 Einleitung

Wie bei anderen Messmethoden auch, muss man bei einem UX Fragebogen sicherstellen, dass die gemessenen Werte gewisse Gütekriterien erfüllen. Schon bei der Konstruktion des Fragebogens wird versucht, durch geeignete Methoden zu erreichen, dass die entstehenden Skalen von den Befragten auch als unterschiedliche UX Aspekte wahrgenommen werden. Der neue Fragebogen wird dann in weiteren Studien validiert, wobei man sich an den klassischen Gütekriterien psychologischer Tests, also Objektivität, Reliabilität und Validität orientiert.

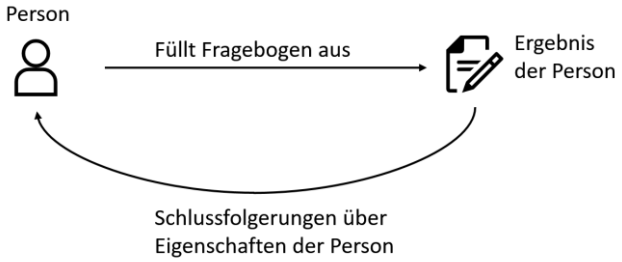
Diese Fokussierung auf die aus der Psychologie bekannten Methoden hat zwei Ursachen. Zum einen haben diese Methoden eine lange Tradition, sind gut ausgearbeitet und über eine Vielzahl von Lehrbüchern leicht zugänglich. Zum anderen haben die meisten Fragebogen-Entwickler im UX Bereich einen psychologischen Hintergrund.

Allerdings gibt es zwischen psychologischen Tests und UX Fragebögen einige strukturelle Unterschiede. Ziel dieses Beitrags ist es, diese zu beschreiben und zu hinterfragen, ob diese Unterschiede einige häufig verwendete Methoden nicht etwas fragwürdig machen.

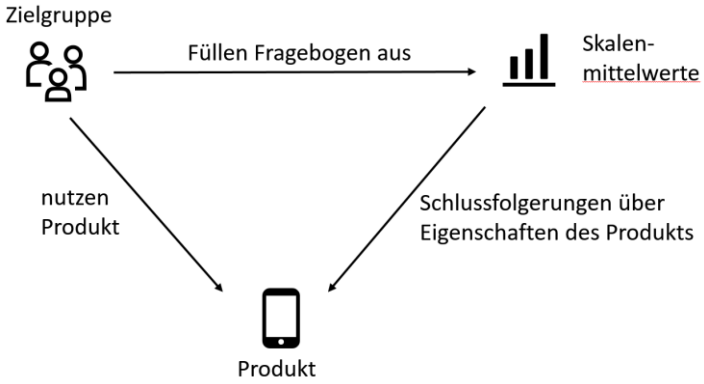
Veröffentlicht durch die Gesellschaft für Informatik e. V. 2018 in
R. Dachselt, G. Weber (Hrsg.):
Mensch und Computer 2018 – Workshopband, 02.–05. September 2018, Dresden.
 https://doi.org/10.18420/muc2018-ws16-0325

2 Psychologische Tests und UX Fragebögen

Bei psychologischen Test, z.B. einem Persönlichkeitstest oder einem Eignungstest, wird aus dem Testergebnis einer Person auf Eigenschaften dieser Person geschlossen, z.B. Introversion, Neurotizismus oder die Eignung für einen Beruf oder eine Ausbildung. Interpretiert wird also das Ergebnis auf Personenebene.



Bei UX Fragebögen messen wir den Eindruck, den eine repräsentative Gruppe von Nutzern von einem Produkt hat. Wir sind hier nicht an den Daten einzelner Personen interessiert. Aus diesen Daten schließen wir dann auf die UX Qualität des Produkts.



D.h. bei UX Skalen wird der Mittelwert über eine Zielgruppe interpretiert und dieser Mittelwert soll genau gemessen werden, nicht die Meinung eines einzelnen Nutzers.

3 Auswirkungen auf das Reliabilitätskonzept

Aufgrund des Ergebnisses eines Tests werden Diagnosen gestellt, Behandlungen empfohlen oder bei Leistungstests die Zugänge zu Studienplätzen vergeben. D.h. für die getestete Person kann das Ergebnis nicht unerhebliche Konsequenzen haben. Deshalb ist es wichtig, dass das Ergebnis auf Personenebene möglichst genau und reproduzierbar ist. Daraus leitet sich auch das Reliabilitätskonzept der klassischen Testtheorie ab (Lienert, 1989).

Reliabilität ist hier als Korrelation der Skalenwerte mit den Skalenwerten einer parallelen Skala (d.h. einer theoretisch angenommenen Skala, die den gleichen „wahren“ Score produziert und die gleiche Varianz der Testergebnisse aufweist) definiert. Die Grundidee ist also, dass zwei unabhängige Messungen der gleichen Person ähnliche Ergebnisse produzieren.

Die Reliabilität einer Skala kann man nicht direkt berechnen, sondern nur schätzen (eine untere Grenze angeben). Das für UX Fragebögen hierfür fast ausschließlich angewendete Verfahren ist die Berechnung des (standardisierten) Cronbach α Koeffizienten (Cronbach, 1951):

$$\alpha = \frac{n * \bar{r}}{(1 + (n-1) * \bar{r})}$$

wo $\bar{r}$ die mittlere Korrelation aller n Items der Skala ist.

Items eines UX Fragebogens werden immer im Kontext des evaluierten Produkts interpretiert. Die Korrelationen der Items einer Skala und damit α werden also zwischen zwei Produkten variieren. Damit ist die Reliabilität keine Eigenschaft des Fragebogens, sondern immer abhängig vom gemessenen Produkt!

In UX Fragebögen interpretieren wir eigentlich nie die Antworten einer Person, sondern immer Mittelwerte über eine repräsentative Nutzergruppe. Diese kann man auch bei geringer Reliabilität noch relativ genau schätzen. Wir zeigen das anhand einer kleinen Simulation.

Grundlage ist ein Datensatz zum UEQ mit 240 Befragten. Wir nutzen eine der Skalen (4 Items) als Basis für die Simulation. Der Skalenmittelwert war 5,64, die Standardabweichung 1,14 und der Wert für α ist 0.81 (also recht hoch). Jetzt ziehen wir aus dem Datensatz 200-mal zufällig eine Stichprobe der Größe 30 und berechnen jeweils α und den Mittelwert der Skala. Die Ergebnisse sind in Abbildung 1 zu sehen.

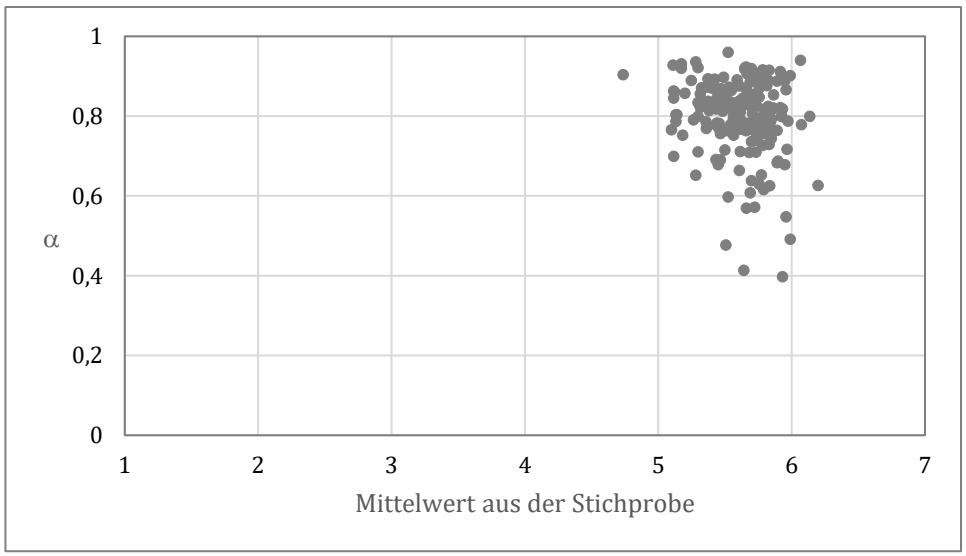


Abbildung 1: Skalenmittelwerte und α für 200 Stichproben der Größe 30.

Offenbar ist α recht anfällig gegen Sampling-Effekte. Der Skalenmittelwert ist deutlich stabiler. Müssen wir uns also Sorgen machen, wenn wir in einer Untersuchung mit begrenzter Teilnehmerzahl eine geringe Reliabilität für eine Skala messen? Wenn wir nur den Skalenmittelwert über die Zielgruppe interpretieren, offenbar nicht!

In einer Variation der Simulation wurden zusätzliche Zufallsfehler implementiert, d.h. die gezogenen Antworten wurden noch mit einem Zufallsmechanismus verrauscht, um eine geringe Reliabilität zu induzieren. Abbildung 2 zeigt den Zusammenhang zwischen der Abweichung des Skalenwerts der Stichprobe vom Skalenwert über alle 240 Daten (Y-Achse) und dem α Wert der Stichprobe (X-Achse). Offenbar ist hier nur ein schwacher Zusammenhang zu erkennen (Trendlinie).

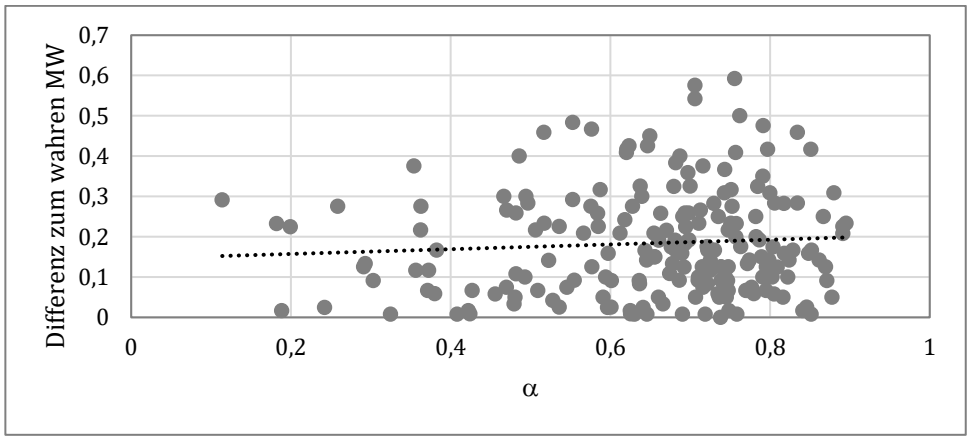


Abbildung 2: Zusammenhang zwischen α und der Genauigkeit der Schätzung des Skalenmittelwerts.

4 Testkonstruktion über Faktorenanalyse

UX Fragebögen werden meist mit Hilfe von Faktorenanalysen konstruiert. Dabei beginnt man mit einer großen Menge potentieller Items, die die UX im relevanten Produktbereich abbilden. Dann werden mehrere Produkte von einer größeren Zahl von Personen mit diesen Items beurteilt. Über eine Faktorenanalyse werden dann Faktoren (unbenannte hypothetische UX Aspekte) ermittelt, die die Skalen des Fragebogens bilden. Pro Skala (Faktor) wird eine kleine Menge von Items ausgewählt, die den zugrundeliegenden UX Aspekt repräsentieren (typischerweise die mit der höchsten Ladung auf dem Faktor). Der Name der Skala wird dann so gewählt, dass er die Gemeinsamkeit der Items der Skala gut beschreibt. So sind z.B. der AttrakDiff2 (Hassenzahl et al., 2003), der UEQ (Laugwitz et al., 2006) und der VISAWI (Mooshagen & Thielsch, 2010) entstanden.

Was ist das Problem mit dieser gängigen Vorgehensweise? Wie schon erwähnt, werden Items von UX Fragebögen immer im Kontext des evaluierten Produkts interpretiert. Die Bedeutung eines Items kann sich hier also abhängig vom Produkt leicht ändern und damit auch seine Korrelation zu anderen Items.

Ein Item *Spart mir Zeit / Kostet mich Zeit* wird bei der Evaluation einer betriebswirtschaftlichen Software sicher als *Das Produkt spart mir Arbeitszeit* interpretiert, d.h. eine hohe Korrelation mit anderen Effizienz-Items aufweisen. Wird ein soziales Netzwerk untersucht, werden einige Teilnehmern dieses Item als *Ich verbringe zu viel Zeit in diesem Netzwerk* interpretieren. Die Korrelation mit anderen Effizienz-Items wird hier geringer sein. Das ist natürlich ein drastisches Beispiel. Aber auch kleinere Bedeutungsunterschiede können die Korrelationen zwischen Items beeinflussen und führen damit zu einer anderen Faktorenstruktur.

Aus diesem Grund ist es ratsam, immer eine größere Menge von Produkten in einer solchen Studie zu untersuchen. Aber natürlich gibt es hier praktische Grenzen, d.h. letztlich wird man sich auf eine überschaubare Menge von Produkten beschränken müssen. Daraus ergeben sich verschiedene Schwierigkeiten. Erstens ist es relativ schwierig die Skalenstruktur empirisch zu replizieren. Das kann eigentlich nur dann funktionieren, wenn man die Replikationsstudie mit den gleichen oder einer zumindest sehr ähnlichen Menge von Produkten durchführt.

Zweitens sind die aus dieser Art der Konstruktion entstehenden Skalen oft nicht besonders sauber semantisch abgegrenzt und in vielen Fällen enthalten zwei Fragebögen ähnliche oder gleich benannte Skalen, die aber auf der Ebene der Items betrachtet nicht identisch sind (Schrepp, 2018). Betrachten wir auch dazu ein konkretes Beispiel. Der AttrakDiff2 (Hassenzahl et al., 2003) und der UEQ (Laugwitz et al., 2008) enthalten beide eine Skala *Stimulation*:

- **UEQ Stimulation:** uninteressant/interessant, langweilig/spannend, aktivierend/einschläfernd, wertvoll/minderwertig
- **AttrakDiff2 Stimulation:** phantasielos/kreativ, originell/konventionell, innovativ/konservativ, neuartig/herkömmlich, mutig/vorsichtig, harmlos/herausfordernd, lahm/fesselnd

Beide Konzepte von Stimulation sind ähnlich, aber nicht identisch. Die ersten 4 Items der Skala aus dem AttrakDiff2 beschreiben den Aspekt der kreativen Gestaltung, die restlichen drei Items eher den Aspekt der spannenden Interaktion. Der Aspekt der kreativen Gestaltung wird im UEQ in einer eigenen Skala *Originalität* beschrieben, d.h. die Skala *Stimulation* im AttrakDiff2 ist eigentlich eine Kombination der Skalen *Stimulation* und *Originalität* im UEQ.

Natürlich sind beide Arten mit dem Konzept *Stimulation* umzugehen akzeptabel. Eine originelle Gestaltung weckt das Interesse des Nutzers und macht das Produkt interessanter, d.h. man kann *Originalität* durchaus als Teil von Stimulation auffassen. Man kann aber auch beide Konzepte getrennt erfassen. Diese unterschiedlichen Operationalisierungen wurden aber hier nicht bewusst gewählt, sondern resultieren aus unterschiedlichen Arten von Produkten in den Studien zur Fragebogenkonstruktion.

5 Schlussfolgerungen

Zwischen psychologischen Tests und UX Fragebögen bestehen strukturelle Unterschiede. Diese machen die Verwendung einiger Methoden der psychologischen Testtheorie bei Konstruktion und Validierung von UX Fragebögen zumindest fragwürdig. Dies wurde am Beispiel

des Reliabilitätskonzepts und der Verwendung der Faktorenanalyse zur Konstruktion von UX Fragebögen herausgearbeitet.

Es ist fraglich, ob das klassische Reliabilitätskonzept für UX Fragebögen geeignet ist. Erstens ist die Reliabilität hier keine Eigenschaft des Fragebogens, sondern ist abhängig vom gemessenen Produkt. Zweitens sind wir eigentlich immer an den Skalenmittelwerten über eine hinreichend große Nutzergruppe interessiert und diese lassen sich auch noch mit geringer Reliabilität einer Skala gut schätzen. Andererseits brauchen wir, wie für andere Messmethoden auch, ein geeignetes Konzept zur Messgenauigkeit und damit Qualität von UX Skalen. Wie könnte ein geeignetes Reliabilitätskonzept für UX Fragebögen aussehen?

Die Konstruktion von UX Fragebögen über Faktorenanalysen führt zu nicht immer semantisch sauber abgegrenzten Skalen. Das macht es für den Anwender schwierig wirklich zu verstehen, was die Skalen eines Fragebogens messen. Auch für wissenschaftliche Arbeiten, die auf Ergebnissen von Fragebögen beruhen, hat das negative Konsequenzen. Nehmen wir mal an, wir finden in einer solchen Arbeit einen interessanten Zusammenhang zwischen Stimulation und einer anderen relevanten Produkteigenschaft. Was dieser Zusammenhang inhaltlich bedeutet, hängt aber davon ab, ob wir mit dem AttrakDiff2 oder dem UEQ gemessen haben. Ist das den Lesern und Autoren solcher Publikationen immer klar? Wie können Autoren von Fragebögen die Anwender besser unterstützen, um die Skalen vernünftig zu interpretieren?

Macht es, zumindest aus Sicht von Designern, nicht mehr Sinn, klar zu definieren, welche UX Qualitäten wir messen wollen und dann ausgehend davon passende Items zu konstruieren? D.h. sollten wir nicht eher im Vorfeld klären, was wir messen wollen und die für uns relevanten UX Faktoren dann sauber beschreiben, statt empirisch vorzugehen und dann die über eine Faktorenanalyse gefundenen Skalen zu nutzen?

Literatur

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, pp. 297-334.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität In: Ziegler, J. & Szwillus, G. (Hrsg.), Mensch & Computer 2003. Interaktion in Bewegung, S. 187-196, Stuttgart, Leipzig: B.G. Teubner.

Lienert, Gustav A. (1989). Testaufbau und Testanalyse. München: Psychologie Verlags Union.

Laugwitz, B.; Schrepp, M. & Held, T. (2006). Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten. A.M. Heinecke & H. Paul (Eds.): Mensch & Computer 2006 - Mensch und Computer im Strukturwandel. Oldenbourg Verlag, S. 125 – 134

Moshagen, M. & Thielsch, M. T. (2010). Facets of visual aesthetics. International Journal of Human-Computer Studies, Vol. 68, S. 689-709.

Schrepp, M. (2018). User Experience mit Fragebögen messen. ISBN: 9781986843768. Printed by CreateSpace (Amazon).

Autoren

Schrepp, Martin

Dr. Martin Schrepp studierte Mathematik und Psychologie an der Universität Heidelberg. 1990 Abschluss als Diplom-Mathematiker. 1990 – 1993 Promotion in Psychologie. Seit 1994 bei der SAP AG tätig. Er ist einer der Autoren des User Experience Questionnaire (UEQ) und hat Erfahrungen in der praktischen Anwendung zahlreicher anderer UX Fragebögen. Er ist auch Autor zahlreicher Beiträge zu methodischen Fragen im UX Bereich.

Rummel, Bernard

Bernard Rummel studierte Psychologie an der Universität Kiel (Diplom 1990). Nach neun Jahren am Schiffahrtmedizinischen Institut der Marine kam er 2000 zu SAP. Er arbeitet weiterhin im DIN-Normenausschuss Ergonomie – Benutzungsschnittstellen an Normenreihen wie der ISO 9241, sowie als Fachprüfer des UXQB für Usability Testing und Evaluation (CPUX-UT). Seit 2011 beschäftigt er sich bei SAP mit der Quantifizierung von Gebrauchstauglichkeit und Usability Benchmarking.