Quantitative Methoden zur prozessbegleitenden Evaluation von Designvorschlägen

Patrick Fischer
SAP AG
Dietmar-Hopp-Allee 16
69190 Walldorf
patrick.fischer@sap.com

Martin Schrepp
SAP AG
Dietmar-Hopp-Allee 16
69190 Walldorf
martin.schrepp@sap.com

Theo Held
SAP AG
Dietmar-Hopp-Allee 16
69190 Walldorf
theo.held@sap.com

Bettina Laugwitz
SAP AG
Dietmar-Hopp-Allee 16
69190 Walldorf
bettina.laugwitz@sap.com

Abstract

Im Designprozess stehen oft viele Entscheidungen an, die nur schwer aufgrund von vorhandenem Wissen getroffen werden können. Aufwändige Userbefragungen sind in diesen Fällen meist nicht angebracht. Was hier benötigt wird, ist eine kostengünstige und schnelle Methode, zur Entscheidung zu kommen. In diesem Tutorial werden einige Methoden vorgestellt, die es ermöglichen schnell und mit wenig Aufwand an das nötige Feedback zu gelangen. Die vorgestellten Verfahren sind insbesondere auch zum Einsatz in Online-Studien geeignet.

Keywords

Quantitative Methoden, BTL-Skalierung, Magnitude Estimation, Conjoint Measurement

1.0 Einleitung

Während der Gestaltungsphase von Benutzungsoberflächen treten verschiedene Probleme auf, wie z.B. Entscheidung für ein konkretes visuelles Design oder die Art der Verteilung von Informationen. In vielen Fällen sind es Usability Experten, die durch ihr Wissen und ihre Erfahrung ad hoc solche Probleme lösen können.

Ist die Lösung eines Problems nicht offensichtlich, werden Endbenutzer in den Entwicklungsprozess einbezogen. Hierfür gibt es altbewährte Methoden, wie z.B. Usability Tests oder Usability Reviews. Bei diesen Methoden wird der direkte persönliche Kontakt bevorzugt. Endbenutzer bearbeiten im Labor Aufgaben, werden dabei beobachtet oder gefilmt und bekommen anschließend evtl. noch einen Fragebogen. Mitunter werden durch unstrukturierte Interviews zusätzliche Informationen gewonnen. Experten werten die Daten aus, beurteilen anschließend die Datenlage und kommen zu entsprechenden Vorschlägen, die in den Designprozess einfließen.

Ein Nachteil der Labor Tests bzw. anderer Methoden mit direktem Nutzerkontakt ist der Aufwand (finanziell und zeitlich). Für die im Designprozess auftretenden Probleme wird eine kostengünstige und schnelle Methode benötigt, um auf empirischer Basis Entscheidungen zu treffen.

Nun sind aus den Bereichen der empirischen Marktforschung und der experimentellen Psychologie Methoden bekannt, die prinzipiell in Frage kommen, Entscheidungshilfe zu leisten, z.B.: Magnitude Estimation, conjoint Measurement, vollständiger Paarvergleich sowie Fragebögen.

Insbesondere in Kombination mit der Nutzung des Inter-/Intranets, bieten sich solche quantitativen Methoden an, um zügig an Feedback zu gelangen. Nach unserer Erfahrung wird dies aber selten praktiziert. Gerade die Methoden, die helfen unter Zeitdruck Entscheidungen zu treffen, sind eine sinnvolle Ergänzung im Methodenrepertoire.

2.0 Skalierungsverfahren

Skalierungsverfahren sind etablierte Methoden der empirischen Wissenschaften. Sie ermöglichen es, die Qualität von Objekten in Bezug auf vorgegebene Kriterien zu messen. Die Messwerte werden dabei aus den Daten von Personen ermittelt, die die Alternativen in Bezug auf das Kriterium beurteilen. Für die Datenerhebung reichen in der Regel kleine Stichproben (10 – 20 Personen) aus.

Beim Einsatz im Bereich des User Interface Designs sind die Objekte in der Regel Designalternativen, z.B. Screens, einzelne Controls, oder Entwürfe für das visuelle Design. Als Kriterien für die Bewertung kommen z.B. die wahrgenommene Usability oder die Ästhetik des Designs in Frage.

Die meisten Skalierungsmethoden sind in Bezug auf die Datenerhebung recht einfach. Sie eignen sich damit sehr gut dafür, während des Designprozesses schnell zu Entscheidungen zwischen alternativen Entwürfen zu kommen. Wegen ihrer einfachen Struktur sind Skalierungsverfahren auch sehr gut für Online-Studien geeignet.

Wir beschreiben drei Skalierungsverfahren, die wir teilweise schon in einigen Projekten erfolgreich eingesetzt haben.

2.1 BTL Skalierung (Vollständiger Paarvergleich)

Die Daten für eine BTL Skalierung (Bradley & Terry 1952; Luce 1959) werden über einen vollständigen Paarvergleich ermittelt. Dabei werden *n* Personen jeweils alle Paare von Alternativen (in zufälliger Reihenfolge) vorgelegt. Die Person entscheidet dann für jedes Paar, welche Alternative sie bevorzugt.

Die Alternativen sollten gleich häufig an den beiden Positionen auftauchen, um Reihenfolgeeffekte auszuschließen. Dies wird oft dadurch erreicht, dass jedes Paar zweimal mit vertauschten Positionen dargeboten wird. Aus den Daten der *n* Beurteiler kann dann die sogenannte Dominanzmatrix berechnet werden. Diese listet für jedes Paar (*a*, *b*) von Alternativen auf, wie häufig Alternative *a* gegenüber Alternative *b* bevorzugt wurde.

Aus der Dominanzmatrix werden dann die Skalenwerte *S*(*a*) der Alternativen ermittelt. Hier liegt die Annahme zugrunde, dass die Häufigkeit, mit der eine Alternative einer anderen vorgezogen wird, nur von den Skalenwerten der Alternativen abhängt. Konkret geht man von folgendem Zusammenhang[1] aus:

$$P(a,b) = \frac{S(a)}{S(a) + S(b)}$$

Die BTL Skalierung liefert eine Verhältnisskala. Angenommen wir skalieren mehrere visuelle Designs bzgl. ihrer Ästhetik und für zwei dieser Designs *a*, *b* gilt $S(a) = 2\ S(b)$. Dann kann man schlußfolgern, daß Design *a* als doppelt so ästhetisch wahrgenommen wird wie Design *b*.

BTL Skalierung bietet sich insbesondere an, wenn man wenige[2] Designalternativen in Bezug auf eine interessante Ausprägung (Ästhetik, wahrgenommene Usability, etc.) vergleichen will. Besonders geeignet ist das Verfahren bei einfachen Reizen, z.B. visuellen Designs, bei denen man im Paarvergleich jeweils zwei Alternativen parallel darbieten kann.

Eine ausführliche Darstellung der BTL Skalierung findet man z.B. in Gediga (1998). Anwendungen auf das Formulardesign sind z.B. in Schrepp et al. (2007) beschrieben.

2.2 Conjoint Measurement

Die Conjoint Analyse oder verbundene Messung wurde zur Messung multidimensionaler Objekteigenschaften entwickelt. Das heutige Hauptanwendungsgebiet der Methode ist die empirische Marktforschung.

Im Bereich des User Interface Designs eignet sich die Conjoint Analyse immer dann, wenn die alternativen Designentwürfe durch eine Anzahl von unabhängig voneinander variierbaren Attributen charakterisiert sind. In einer Anwendung auf das Formulardesign (Schrepp et al. 2007) wurde z.B. untersucht, wie sich die Anordnung der Feldbezeichner, die Balance des Layouts und die Visualisierung der Feldgruppierung auf die Ästhetik des Formulardesigns auswirken.

Bei der Conjoint Analyse ist jede Alternative durch die Ausprägung der Attribute beschreiben. Das Verfahren liefert nun pro Attribut Skalenwerte, die beschreiben, welchen Einfluss das jeweilige Attribut auf das vorgegebene Beurteilungskriterium hat. In der oben beschriebenen Untersuchung zum Formularlayout lieferte das Verfahren z.B. Skalenwerte für die drei Attribute *Anordnung der Feldbezeichner*, *Balance des Layouts* und *Visualisierung der Gruppen*. Die Skalenwerte beschreiben den Einfluss des jeweiligen Faktors auf die Ästhetik des Layouts.

Die Daten für eine Conjoint Analyse werden in der Regel über ein Ranking-Verfahren ermittelt. Bei einem Ranking-Verfahren ordnet ein Beurteiler alle Alternativen entsprechend seiner Präferenz direkt an. D.h. jeder Beurteiler produziert hierbei eine Rangreihe der Alternativen. Dieses Verfahren hat den Vorteil, dass es bei einer überschaubaren Anzahl von zu beurteilenden Objekten sehr schnell durchzuführen ist.

Die Skalenwerte der Attribute werden aus den beobachteten Rangreihen über eine Regressionsanalyse bestimmt.

Die Conjoint Analyse eignet sich immer dann, wenn man den Einfluß bestimmter Faktoren auf den Eindruck untersuchen will, den ein Design beim Benutzer hervorruft.

Eine übersichtliche Darstellung über die wesentlichen Aspekte der Conjoint Analyse gibt Klein (2002).

2.3 Magnitude Estimation

Anders als bei den beiden oben vorgestellten *indirekten* Skalierungsverfahren handelt es sich bei der Magnitude Estimation um einen Ansatz zur *Direktskalierung*. Das Verfahren wurde von Stevens (1946) vorgeschlagen. In einem typischen Magnitude Estimation Experiment ist es die Aufgabe der Versuchsperson, Reizen numerische Werte in der Weise zuzuordnen, dass die zugeordneten Werte proportional zur Empfindungsgröße des wahrgenommenen Reizes sind. Im Allgemeinen wird zunächst einem Referenz-Reiz ein gegebener Wert zugeordnet (z.B. 100) und die Versuchsperson ist angehalten alle folgen-

1 Dabei ist *P*(*a*, *b*) die Wahrscheinlichkeit, dass *a* gegenüber *b* bevorzugt wird und *S*(*a*) bzw. *S*(*b*) sind die Skalenwerte der Alternativen *a* und *b*.

2 Der zeitliche Aufwand für einen vollständigen Paarvergleich steigt mit der Zahl der Alternativen stark an.

den Wertezuordnungen auf diesen Referenzwert zu beziehen. Wenn die Versuchsperson also einen Reiz als dreimal so intensiv (z.B. „laut“, „hell“, „angenehm“) wie den Referenzreiz empfindet, sollte sie diesem Reiz eine dreimal so hohe Zahl zuordnen (z.B. 300 bei einer Referenz von 100). Ohne auf die formalen Details dieses Skalierungsverfahrens einzugehen, sei hier nur erwähnt, dass nach Stevens auf Basis der Daten eines Magnitude Estimation Experiments eine *Verhältnisskala* konstruiert werden kann.

Magnitude Estimation ist eine weitverbreitete Methode in der empirischen Marktforschung und nach wie vor auch in der experimentellen Psychologie. Interessant für Usability Professionals ist, dass eine Anwendung des Verfahrens für die Skalierung von Interaktionssequenzen bezüglich ihrer wahrgenommenen Usability vorgeschlagen wurde. McGee (2003) und Rich & McGee (2004) präsentieren mit der „Usability Magnitude Estimation“ einen Ansatz, der für sich beansprucht, effizient und verlässlich Verhältnisskalen zur wahrgenommenen Usability von Softwareprodukten zu generieren. Die Rolle der Reize spielen bei diesem Verfahren einzelne Interaktionssequenzen, die die Versuchsperson bei der Bedienung eines Softwareprodukts zu durchlaufen hat. Wie bei traditionellen Magnitude Estimation-Experimenten wird der Versuchsperson mit Hilfe neutraler Interaktionen eine Bewertungsreferenz vorgegeben. Die Interaktionssequenzen des zu beurteilenden Softwareprodukts werden dann in Bezug auf die Referenz mit Zahlenwerten beurteilt.

Da die Versuchsperson bei allen Zuordnungen numerischer Werte auch die vorher erfolgten Zuordnungen berücksichtigen muss, ist die Durchführung eines solchen Verfahrens potentiell mit einer *hohen kognitiven Belastung* verbunden.

Befunde der modernen psychologischen Messtheorie zeigen zudem, dass Magnitude Estimation nicht notwendigerweise zu einer Verhältnisskalierung führt. Eine Theorie von Narens (1996) bietet eine Axiomatisierung der zugrunde liegende Messstruktur und zeigt die notwendigen Bedingungen für die Identifikation des realen Skalentyps auf. Zimmer et al. (2004) stellen einen Vergleich von BTL-Skalierung und Magnitude Estimation an und betonen ebenso das Problem des undeterminierten Skalentyps bei der Magnitude Estimation.

3.0 Fragebogen

Eine schnelle und einfache Möglichkeit, subjektive Einschätzungen von Endbenutzern zu erheben, sind Fragebögen. Zur Evaluation von Software stehen verschiedene solcher Verfahren zur Verfügung. Diese können sich unterscheiden hinsichtlich ihrer Zielsetzung, der Art der erzeugten Daten (quantitativ, ggf. zusätzlich qualitativ) oder auch des Erhebungsaufwands.

Manche Fragebögen sind in erster Linie darauf ausgelegt, eine grobe Einschätzung von Produktmerkmalen zu liefern. Ein Beispiel hierzu ist die System Usability Scale SUS (Brooke 1996), der es erlaubt, aufgrund von 10 Items einen Gesamt-Usability-Wert zu ermitteln, der zwischen 0 und 100 rangiert.

Ein differenzierteres Bild hinsichtlich der Produkteigenschaften kann beispielsweise der IsoMetricsS (Gediga & Hamborg 1999) liefern, der sich an den Kriterien für Gebrauchstauglichkeit gemäß ISO 9241 Teil 10 orientiert. Der Benutzer beurteilt die Ausprägung der entsprechenden Eigenschaften, wie etwa der Aufgabenangemessenheit oder der Fehlerrobustheit.

Der UEQ (Laugwitz et al. 2006) wurde hingegen mit dem Ziel konstruiert, eine schnelle, möglichst unmittelbare Einschätzung des Produkts bzw. der Zufriedenstellung des Benutzers gemäß ISO 9241 Teil 11 zu unterstützen. Die Auswertung liefert dann intervallskalierte Werte auf fünf Skalen, darunter „Originalität“ oder „Vorhersagbarkeit“.

Die genannten Fragebögen liefern quantitative Daten und sind dadurch in der Auswertung sehr effizient und besonders für eine computergestützte Erhebung geeignet. Andere Verfahren erlauben ein teil-standardisiertes Vorgehen, in dem nicht nur ein Wert auf einer Ratingskala ausgewählt werden kann, sondern zusätzlich Raum für freitextliche Äußerungen gegeben ist (z. B. IsoMetricsL, Gediga & Hamborg 1999; ErgoNorm-Benutzerfragebogen, Dzida et al. 2000). Die Ergebnisse der qualitativen Anteile können Erkenntnisse über Stärken und Schwächen des Produkts mit einem Detaillierungsgrad liefern, wie es den rein quantitativen Verfahren nicht möglich ist. Preis hierfür sind ein erhöhter Erhebungs- und Auswertungsaufwand.

Auch wenn es dem Interaktionsdesigner jederzeit freisteht, eigene Fragen zu formulieren oder Fragensätze zu erstellen, sollte nach Möglichkeit den veröffentlichten Verfahren den Vorzug geben, da hier relevante Gütekriterien wie die Objektivität, Reliabilität und Validität dokumentiert und nachvollziehbar sind. Von der Abänderung vorhandener Fragebögen ist grundsätzlich abzuraten. Für die Einschätzung oder den Vergleich von Produkten hinsichtlich einzelner Dimensionen wie Ästhetik oder wahrgenommene Usability durch Endnutzer sind außerdem die oben beschriebenen Skalierungsverfahren eher geeignet.

Einen Überblick über den Einsatz von Fragebögen im Bereich Softwareergo-

nomie bieten z. B. Dzida et al. (2000) oder auch Hamborg (2004).

4.0 Anwendungsbeispiel

Aus dem Tagesgeschäft der Autoren soll im Folgenden die Vorgehensweise für eine BTL-Skalierung kurz umrissen werden.

Es standen sechs Designvorschläge von Widgets für eine Web-Applikation zur Verfügung. Ziel war es, so zügig wie möglich, eine Entscheidung für eine der sechs Designalternativen zu treffen. Die Widgets wurden in eine vorhandene Screenstruktur integriert und davon Screenshots angefertigt. Mithilfe eines Softwaretools auf Basis von PXLab (Irtel 2007) wurde ein Online-Paarvergleich generiert und im Intranet von SAP zugänglich gemacht. Alle Produktmanager und User Interface Designer aus dem CRM Bereich wurden per E-Mail aufgefordert den Paarvergleich durchzuführen.

Insgesamt wurden aus den sechs Designalternativen 15 Paare generiert. Die Versuchspersonen mussten mithilfe der Pfeiltasten ihre Entscheidung bezüglich der Ästhetik für jedes Paar treffen (siehe Abbildung 1). Für alle Entscheidungen benötigten die Versuchspersonen im Durchschnitt 5 Minuten.

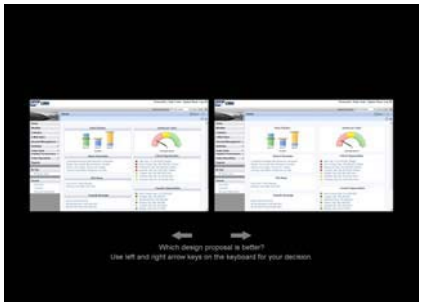

Abb. 1: Screenshot zu einer Entscheidung für einen der beiden Designvorschläge.

Das Softwaretool übernahm automatisch die Datenauswertung und lieferte die Skalenwerte für jede Designalternative in numerischer und grafischer Form. Die Daten wurden nach der BTL-Prozedur (Bradley & Terry 1952; Luce 1959) ausgewertet. Zu diesem Zweck wurde auf das „eba“-Modul (Wickelmaier 2008) für die Statistiksoftware R zurückgegriffen.

Der zeitliche Aufwand zur Abwicklung der gesamten Studie, belief sich lediglich auf wenige Stunden. An einem Tag konnte der Paarvergleich erstellt, die Aufforderung zur Teilnahme per Email versandt und die Ergebnisse schon einige Stunden später gesichtet werden. Dies ermöglichte eine zügige Entscheidung für einen der Designvorschläge auf Basis von empirischen Daten.

5.0 Fazit

Skalierungsverfahren erlauben es den Einfluss bestimmter Designeigenschaften auf den durch das Design hervorgerufenen Eindruck sehr genau zu bestimmen. Dies kann insbesondere bei Fragen relevant sein, bei denen eine Abwägung zwischen verschiedenen Faktoren bedeutsam ist. Zum Beispiel stellt sich bei der Gestaltung von Web-Anwendungen oft die Frage, ob der durch die Verwendung eines komplexeren Style-Sheets verbesserte visuelle Eindruck die dadurch evtl. hervorgerufene Verschlechterung bei der zum Seitenaufbau notwendigen Zeit aufwiegt. Solche Fragen können mit den beschriebenen Skalierungsverfahren gut untersucht werden.

Der Einsatz von Fragebögen ist dann besonders sinnvoll, wenn ein Maß für bestimmte komplexe Produkteigenschaften (eingeschätzte Usability, Benutzerzufriedenheit) benötigt wird. Fragebogendaten können zu verschiedenen Zeitpunkten mit unterschiedlichen Befragten zu verschiedenen Produkten erhoben werden und erlauben es dennoch bis zu einem gewissen Grad, Vergleiche zwischen verschiedenen Produkten oder Produktversionen anzustellen.

Im Designalltag können quantitative Methoden, wie die oben dar gestellten, bei Designentscheidungen effiziente Unterstützung bieten. Dies gilt insbesondere, wenn die Datenerhebung und die Auswertung web- bzw. computergestützt erfolgt. Die im Tutorial vorgestellte Software ermöglicht das Durchführen von Online-Studien mit BTL-Skalierung, Magnitude Estimation sowie Fragebogen.

6.0 Literatur

Bradley, R. A.; Terry, M. E. (1952): Rank Analysis of Incomplete Block Designs: I. The method of Paired Comparisons. Biometrika, Vol. 39, S. 324-345.

Brooke, J. (1996): SUS: a "quick and dirty" usability scale. In: Jordan, P. W.; Thomas, B.; Weerdmeester, B. A.; McClelland, A. L. (Hrsg.): Usability Evaluation in Industry. London: Taylor and Francis.

DIN EN ISO 9241-10 (1996): Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten. Teil 10: Grundsätze der Dialoggestaltung. Berlin: Beuth Verlag.

DIN EN ISO 9241-11 (1999): Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten. Teil 11: Anforderungen an die Gebrauchstauglichkeit - Leitsätze. Berlin: Beuth Verlag.

Dzida, W.; Hofmann, B.; Freitag, R.; Redtenbacher, W.; Baggen, R.; Geis, T.; Beimel, J.; Zurheiden, C.; Hampe-Neteler, W.; Hartwig, R.; Peters, H. (2000): Gebrauchstauglichkeit von Software: ErgoNorm: Ein Verfahren zur Konformitätsprüfung von Software auf der Grundlage von DIN EN ISO 9241 Teile 10 und 11. Schriftenreihe der Bundesanstalt für Arbeitschutz und Arbeitsmedizin. Dortmund: Bundesanstalt für Arbeitschutz und Arbeitsmedizin.

Gediga, G. (1998): Skalierung. Münster: Lit Verlag.

Gediga, G.; Hamborg, K.-C. (1999): IsoMetrics: Ein Verfahren zur Evaluation von Software nach ISO 9241-10. In: Holling, H.; Gediga, G. (Hrsg.): Evaluationsforschung. Göttingen: Hogrefe, S. 195 - 234.

Hamborg, K.-C. (2004): Fragebögen zur Bestimmung der ergonomischen Qualität von Software. In: Hassenzahl, M.; Peissner, M. (Hrsg.): Usability Professionals 2004. Berichtband des zweiten Workshops des German Chapter der Usability Professionals Association e. V.. Stuttgart: German Chapter der Usability Professionals Association e. V., S. 92-95.

Irtel, H. (2007). *PXLab: The Psychological Experiments Laboratory* (Vers. 2.1.11). Online im Internet: URL: http://www.pxlab.de (Stand 19. 6. 2007).

Laugwitz, B.. Schrepp, M. & Held, T.(2006): Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten. In: A.M. Heinecke & H. Paul (Hrsg.): Mensch & Computer 2006, 125–134. München: Oldenbourg Verlag.

Luce, R. D. (1959): Individual choice behavior: A theoretical analysis. New York: Wiley.

Klein, M. (2002): Die Conjoint-Analyse. Eine Einführung in das Verfahren mit einem Ausblick auf mögliche sozialwissenschaftliche Anwendungen. In: Zentralarchiv für empirische Sozialforschung (Hrsg.), ZA-Information 50, 7–45. Köln: Universität zu Köln.

McGee, M. (2003): Usability magnitude estimation. Proceedings HFES, 47th Annual Meeting, 691-695.

Narens, L. (1996): A theory of magnitude estimation. Journal of Mathematical Psychology, 40, 109-129.

Rich, A. & McGee, M. (2004): Expected usability estimation. Proceedings HFES, 48th Annual Meeting, 912-916.

Schrepp, M., Held, T. & Fischer, P. (2007): Untersuchung von Designpräferenzen mit Hilfe von Skalierungsmethoden. MMI-Interaktiv, Nr. 13, S. 72- 82.

Stevens, S. S. (1946): On the theory of scales of measurement, Science, 103, 677-680.

Wickelmaier, F. (2008, März): The eba package. URL: http://ftp5.gwdg.de/pub/misc/cran/web/packages/eba/eba.pdf (Stand: 26.05.2008).

Zimmer, K., Ellermeier, W. & Schmid, C. (2004): Using probabilistic choice models to investigate auditory unpleasantness. Acta Acustica united with Acustica, 90, 1019-1028.