Fragebögen zur bestimmung der ergonomischen Qualität von Software

Kai-Christoph Hamborg
Universität Osnabrück
Seminarstr. 20
49069 Osnabrück
khamborg@uni-osnabrueck.de

Abstract
In diesem Beitrag wird zunächst geklärt, über welche Eigenschaften Fragebögen zur Evaluation von Software verfügen und für welche Zielsetzungen sie genutzt werden. Danach wird auf Qualitätskriterien für Fragebögen, auf die Durchführung von Fragebogenuntersuchungen und auf die Datenauswertung eingegangen. Es wird ein Überblick über ausgewählte Verfahren gegeben und schließlich Stärken und Schwächen von Fragebögen als Evaluationsinstrumente dargestellt und diskutiert.

Keywords
Fragebögen, Evaluation, Software-Ergonomie

1.0 Einleitung

Die Bewertung oder auch Evaluation von Software nach ergonomischen Gesichtspunkten wird mit unterschiedlicher Zielsetzung durchgeführt. Speziell bei dem Einsatz von Fragebögen für die Softwareevaluation geht es zumeist darum, die ergonomische Qualität einer Software numerisch zu bestimmen. Die Ergebnisdaten lassen sich dann da zu nutzen, um verschiedene Software miteinander zu vergleichen, einzelne Produkte in Bezug auf die Übereinstimmung mit bestimmten Gestaltungszielen oder -empfehlungen (z.B. auch aus internationalen Standards) zu überprüfen sowie gegen Normdaten zu testen. Der Vergleich von Software wird z.B. in Benchmarking-Untersuchungen und im Rahmen des Prototyping vorgenommen, im letzen Fall insbesondere mit dem Ziel, den Gestaltungsfortschritt eines Systems festzustellen. Darüber hinaus können Fragebögen jedoch auch dazu dienen, Schwachstellen und Probleme einer Software aus Nutzersicht zu identifizieren.

2.0 Fragebögen

Unter einem Fragebogen wird die schriftliche Zusammenstellung von Fragen oder Aussagen in Form sogenannter Items verstanden. Mit Hilfe dieser Items werden Urteile, zumeist der Nutzer, über ein Softwaresystem eingeholt. Die meisten gängigen Evaluations fragebogen umfassen mehrere Subskalen, die Items zur Bewertung einer Software auf speziellen Dimensionen wie z.B. Aufgaben-»angmessenheit« oder »Erlernbarkeit« enthalten. Fragebögen lassen sich nach dem Grad ihrer Standardisierung unterscheiden. Bei vollstandardisierten Fragebögen ist die Itemformulierung, die Darbietungsreihenfolge sowie das Antwortformat festgelegt. Die standardisierte Erfassung der Antworten erfolgt mit sogenannten Einschätzungs- oder auch Ratingskalen, die eine numerische Bewertung der Items erlauben. Bei teilstandardisierten Fragebögen sind die Itemformulierungen und die Darbietungsreihenfolge ebenfalls festgelegt. Die Antwortformate enthalten jedoch (auch) die Möglichkeit, Antworten frei zu formulieren. Hierzu wird den befragten Personen häufig am Ende einer Subskala eines Fragebogens oder jeweils in Verbindung mit einem Item die Gelegenheit gegeben, freitextliche Anmerkungen zu der bewerteten Software zu formulieren.

Die durch einen Fragebogen eingeholten Urteile können sich auf Fakten (»Die Software bietet mir eine Wiederhol-Funktion für wiederkehrende Arbeitsschritte.«), Beurteilungen, Bewertungen und Einstellungen (»Zur Erkundung des Systems durch Versuch und Irrtum wird ermutigt«) sowie Gefühle (»das System ist sehr unangenehm«) richten. Jede Art von Urteil basiert auf mehr oder weniger »selbstbezogenen« Auskünften, d.h. es drückt subjektiv gefärbte »Meinungen« der Befragten aus 5, die auf persönlichen Erfahrungen mit einer Software beruhen. Aus diesem Grund ist es wichtig, dass die befragten Personen bereits Erfahrungen mit der Software gewonnen haben, bevor sie

diese mit einem Fragebogen bewerten. Nur dann ist auch die notwendige Basis für eine Urteilsbildung gegeben. Ist eine Software noch nicht ausreichend bekannt, sollten die Untersuchungsteilnehmer zumindest vor der Befragung möglichst für den Anwendungskontext der Software repräsentative Aufgaben in Nutzungsszenarien bearbeitet haben.

3.0 Gütekriterien

Bei der Auswahl eines Fragebogens müssen bestimmte Gütekriterien Beachtung finden. Hierzu zählen die Objektivität, Reliabilität und Validität als Hauptgütekriterien sowie die Ökonomie und Nützlichkeit (s. 1).
Nur wenn die Hauptgütekriterien gegeben sind, kann wirklich davon ausgegangen werden, dass der Fragebogen zuverlässig funktioniert und den erhobenen Daten vertraut werden kann. Aus diesem Grund sollte man auch vorsichtig mit »selbstgestrickten« und nicht systematisch beforschten Fragebögen sein. Im Zweifelsfall ist es immer sinnvoller auf bewährte Fragebögen zurückzugreifen. Auch sollten einzelne Skalen nicht verkürzt oder sonst wie verändert dargeboten werden. Eine wichtige Ausnahme gibt es allerdings: Es kann aus bestimmten Gründen sinnvoll sein, bei der Evaluation auf komplette Subskalen, die sich auf noch nicht in einem Prototypen implementierte Funktionalitäten beziehen, zu verzichten.

4.0 Durchführung von Fragebogenuntersuchungen

Zu Beginn einer Fragebogenuntersuchung ist es wichtig, die zu befragenden Nutzer zu identifizieren und ggf. zu bearbeitende Nutzungsszenarien auszuwählen (s. 3). Fragebögen werden zumeist als »Papier-und-Bleistift« Version vorgelegt und handschriftlich bearbeitet. Eine Alternative besteht in der computergestützten Darbietung von Fragebögen, ggf. auch per Intra- oder Internet.
Die Ergebnisdaten aus Ratingskalen liegen in numerischer Form als quantitative Daten vor. Einige Fragebögen erfassen aber zusätzlich auch qualitative Daten z.B. Informationen über Schwachstellen und unterstützen damit auch die Definition von Gestaltungsbedarf für eine Software.

4.1 Auswertung quantitativer Daten

Ausgangspunkt für die numerische Bewertung einer Software sind die Antworten der Untersuchungspersonen auf den Ratingskalen des Fragebogens. Nach der Bearbeitung durch die Untersuchungsteilnehmer werden die Itemwerte aller befragten Personen – meist in Form von Mittelwerten – pro Subskala numerisch zusammengefasst. Wird untersucht, ob sich verschiedene (Entwicklungs-) Versionen einer Software oder bereits fertig gestellte Software voneinander unterscheiden, eignet sich für die Datenauswertung die Profilanalyse. Bei der Profilanalyse werden Unterschiede zwischen den verglichenen Softwaresystemen für die Mittelwerte der Skalen eines Fragebogens statistisch geprüft (z.B. mit t-Tests oder Varianzanalyse). Auch kann auf Basis der Ergebnisprofile eine Software mit Normwerten für definierte Produkte und Nutzungskontexte verglichen werden.

Eine weitere Möglichkeit, Befragungsdaten zu analysieren bietet die Datenauswertung an Hand von Prozentwerten. Hierdurch lässt sich überprüfen, in welchem Ausmaß vorgegebene kritische Ausprägungen auf den Skalen erreicht werden (Cutoff-Analyse). Hierzu wird die Struktur des Antwortschemas daraufhin untersucht, ob die Software in einer Dimension einem festgelegten Bewertungskriterium entspricht, also z.B. als »mittelmäßig« oder besser eingestuft wird. Pro Person wird bei diesem Auswertungsmodell der prozentuale Anteil der Antworten mit positiver Beurteilung (also »besser als neutra«) bestimmt. Aus den Einzelbeurteilungen lässt sich eine mittlere Cutoff-Beurteilung bilden, die aussagt, ob das Bewertungskriterium durch die Software erreicht wurde (s. 3, S. 233ff).

4.2 Auswertung qualitativer Daten

Den Ausgangspunkt für die Analyse qualitativer Daten bilden offene, frei formulierte Anmerkungen zu einer Software, aus denen sich konkrete Informationen über Schwachstellen aber auch positive Eindrücke extrahieren lassen.

Für die Verwertung offener Antworten empfiehlt es sich, die enthaltenen Anmerkungen zu kategorisieren und zu priorisieren. Hierzu sind verschiedene Auswertungsschritte erforderlich.
Zunächst müssen die Anmerkungen, die für die Gestaltung der Software relevant sind, als solche erkannt und von anderen nicht aussagekräftigen Inhalten (z.B. allgemeinen Unmutsäußerungen wie: »mir gefällt das Programm grundsätzlich nicht«) getrennt werden. Der »Gehalt« einer Anmerkung kann z.B. dadurch definiert werden, ob Aspekte angesprochen werden, die den Nutzer bei der Arbeit mit der Software schwerwiegend behindern oder die spezifische Vorteile darstellen. Für die Gestaltung der Software sollten entsprechende Inhalte besonders dann berücksichtigt werden, wenn sie von mehreren Personen häufig oder von einzelnen Person wiederholt genannt werden.

5.0 Fragebögen zur Evaluation von Software

Für die Evaluation von Software wurden in den letzten Jahren mehrere Fragebogen entwickelt. Zum großen Teil

fand diese Arbeit an Universitäten oder anderen Forschungseinrichtungen statt. Bedauerlicherweise werden daher die wenigsten von ihnen durch Verlage vertrieben. Auch gibt es für nur wenige der Fragebögen Handbücher, die deren Nutzung sowie die Auswertung und Interpretation der Daten anleiten.
Die folgenden Bewertungsansätze von Fragebogen zur Evaluation von Software lassen sich unterscheiden:

1 Erfassung der Zufriedenheit oder anderer Gefallensaspekte

2 Erfassung der ergonomischen Qualität insbesondere der Gebrauchstauglichkeit (usability) oder der Übere instimmung mit einzelnen Gestaltungsgrundsätzen aus Normen und Standards, die im Zusammenhang mit dem Konzept der Gebrauchstauglichkeit stehen.

Sowohl die Zufriedenheit als auch die Gebrauchstauglichkeit einer Software werden durch die folgenden Fragebogen angesprochen:

- Questionnaire for User Interface Satisfaction (QUIS, 15),
- Post Study System Usability Questionnaire (PSSUQ, 13),
- Software Usability Measurement Inventory (SUMI, 10),
- Attrakdiff 7.

Speziell für die Bewertung der Gebrauchstauglichkeit von Software entsprechend des internationalen Standards zur Dialoggestaltung ISO 9241-10 9 aus Nutzersicht wurden die folgenden Fragebögen konstruiert.:

- ISONorm Fragebogen [14]
- IsoMetrics Fragebogen [4].

Für den PSSUQ und den QUIS Fragebogen sind keine autorisierten deutschsprachigen Versionen bekannt. Umfangreiche Dokumentation und Manuale, die die Anwendung, Datenauswertung und Berichtlegung unterstützen, gibt es für den SUMI 12 und IsoMetrics Fragebogen [2].
Alle genannten Verfahren wurden –in unterschiedlicher Intensität – auf Haupt- und teilweise auch Nebengütekriterien überprüft. Für eine detaillierte Darstellung sei hier auf gesonderte Veröffentlichungen [5], [6] oder die Originalliteratur verwiesen.

6.0 Stärken und Schwächen

Es gibt wohl kaum eine Methode zur Evaluation von Software, die nicht über spezifische Stärken und Schwächen verfügt. Ein grundsätzlicher Schwachpunkt von Fragebögen besteht darin, dass die Beantwortung bzw. Beurteilung der Items stark von dem Erinnerungsvermögen, der Selbstwahrnehmung und der Aufmerksamkeit der Probanden abhängt und sowohl für unwillkürliche Fehler und Verzerrungen als auch für absichtliche Verfälschungen anfälliger ist als verhaltensbasierte Benutzbarkeitstests (Usability-Tests). Fehlersituationen oder ähnliche kritische Ereignisse, die während der Interaktion mit einer Software auftreten können, fließen in die Beantwortung eines Fragebogens retrospektiv aus der Erinnerung der Nutzer ein und können daher Verfälschungstendenzen unterliegen [8].
Auch absichtliche Verfälschungen können auftreten, z.B. wenn die Bewertung einer von den Nutzern wenig akzeptierten Software schlechter verzerrt wird, um Veränderungen zu bewirken.

Diesen Einschränkungen steht jedoch gegenüber, dass das durch einen Fragebogen erhobene generelle Urteil – aggregiert über viele Erfahrungssituationen der Nutzer–, einen größeren »Wirklichkeitsausschnitt« abbildet als dies z.B. bei einem Benutzbarkeitstest der Fall sein kann. Dieser Vorteil ist jedoch eng mit dem Nachteil der höheren Subjektivität der Urteilsbildung durch die Befragten verknüpft.
Während die Konstruktion von Fragebögen sehr aufwändig ist, ist die Anwendung im Vergleich zu den meisten Evaluationsmethoden sehr ökonomisch. Das betrifft sowohl die Datenerhebung als auch –auswertung. Damit verbunden ist auch, dass sich Fragebögen, anders als Benutzbarkeitstests mit geringem finanziellen und organisatorischen Aufwand außerhalb eines Labors an den Arbeitsplätzen der Nutzer einer Software einsetzen lassen. Durch computergestützte Versionen, die die Anwendung auf Arbeitsplatzrechnern oder im Intra-/Internet erlauben, lässt sich der Rücklauf, die Datenadministration sowie die Datenaus wertung automatisieren und nochmals ökonomischer gestalten. Hierdurch ist die Realisierung deutlich größerer Untersuchungsstichproben als bei anderen aufwändigeren Evaluationsmethoden möglich. Die ökonomische Nutzung von Fragebögen prädestiniert dieses Evaluationsinstrument auch für breitflächige Evaluationsuntersuchungen z.B. in großen Unternehmen, um aus ergonomischer Sicht problematische Software in bestimmten

Nutzungskontexten zu identifizieren und diese ggf. durch Folgeuntersuchungen mit aufwändigeren Analyseformen zu spezifizieren.

Eine weitere Stärke von Fragebögen besteht darin, dass die Benutzer einer Software selber die Bewertung der Software vornehmen können und nicht indirekt durch Verhaltensbeobachtungen oder andere Verfahren auf die Geeignetheit einer Software für bestimmte Nutzerpopulationen geschlossen werden muss. Dieser Ansatz entspricht aktuellen Entwicklungsmodellen und ihrer Forderung nach direkter Beteiligung der Nutzer einer Software im Entwicklungsprozess (s. 10).

7.0 References

1 Bortz, J. & Döring, N., Forschungsmethoden und Evaluation für Sozialwissenschaftler (2. Auflage). Berlin: Springer (1995).

2 Gediga, G., Hamborg, K.-C. & Willumeit, H., Das IsoMetrics Handbuch (Version 1.15a). Osnabrücker Schriftenreihe Software-Ergonomie, OSSE – 2. Osnabrück: Universität Osnabrück, Fachbereich Humanwissenschaft, Fachgebiet Arbeits- und Organisationspsychologie (1998).

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

4 Gediga, G., Hamborg K.-C. & Düntsch, I. The IsoMetrics Usability Inventory: An operationali sation of ISO 9241-10, Behaviour and Information Technology, 18, 151-164 (1999).

5 Gediga G. & Hamborg, K.-C., Ergonomische Evaluation von Software: Methoden und Modelle im Software-Entwicklungsprozess. Zeitschrift für Psychologie, 210 (1), 40 - 5 (2002).

6 Hamborg, K.-C., Gediga, G., Hassenzahl, M., Fragebogen zur Evaluation. In S. Heinsen & P. Vogt (Hrsg.), Usability praktisch umsetzen, (S. 172-187). München: Hanser (2003).

7 Hassenzahl, M., The effect of perceived hedo nic quality on product appealingness. International Journal of Human-Computer Interaction, 13, 479-497 (2002).

8 Hassenzahl, M. & Sandweg, N., From Mental Effort to Perceived Usability: Transforming Experiences into Summary Assessments. In Proceedings of the CHI 04 Conference on Human Factors in Computing Systems. Extended abstracts, 1283-1286 (2004).

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

10 ISO 13407, Human-centred design processes f or interactive systems. Genf: ISO (1999).

11 Kirakowski, J., The Software Usability Measurement Inventory: background and usage. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & I. L. McClelland (Eds.). Usability Evaluation in Industry (169-177). London: Taylor & Francis (1996).

12 Kirakowski, J., SUMI User Handbook. University College Cork: Human Factors Research Group (1998).

13 Lewis, J. R., IBM computer usability satisfac tion questionnaire: psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction, 7 (1), 57-78 (1995).

14 Prümper, J. & Anft, M., Die Evaluation von Software auf Grundlage des Entwurfs zur inter nationalen Ergonomie-Norm ISO 9241 Teil 10 als Beitrag zur partizipativen Systemgestaltung – ein Fallbeispiel. In: K.-H. Rödiger (Hrsg.), Software-Ergonomie ´93. Von der Benutzungsoberfläche zur Arbeitsgestaltung (S. 145-156). Stuttgart: Teubner (1993).

15 Shneiderman, B., Designing the User Interface. Strategies for Effective Human-Computer Interaction (3rd Edition). Reading, Massachusetts: Addison-Wesley (1998).