

Qualitätssicherung in der Softwareevaluation

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

Tom Hoemske
Universität Osnabrück
Seminarstraße 20
49069 Osnabrück
Hoemske@web.de

Abstract

Der vorliegende Beitrag beschreibt Ansätze zur Qualitätssicherung für die Planung und Durchführung von software-ergonomischen Evaluationsuntersuchungen sowie für die Datenauswertung und Berichterlegung.

Keywords

Usability, Usability Tests, Qualitätssicherung, Lautes Denken, User Action Framework, Common Industry Format.

1.0 Einleitung

Die gestaltungsunterstützende, formative Evaluation von Prototypen spielt in der iterativen Entwicklung von Software eine wichtige Rolle. Hierfür steht eine breite Palette unterschiedlicher Methoden bereit, wie z.B. Usability Tests oder Inspektionemethoden 7,12, die zu unterschiedlichen Zeitpunkten im Entwicklungszyklus zum Einsatz kommen können. Das Ziel formativer Evaluationsuntersuchungen besteht darin, Nutzungs- bzw. Usabilityprobleme einer Software bereits im Entwicklungsprozess zu erkennen.

Trotz des hohen Stellenwerts formativer Evaluation in der Systementwicklung werden Praktiker bislang nur wenig, häufig aber auch gar nicht bei der Anwendung entsprechender Methoden unterstützt. Die meisten Methoden sind unzureichend definiert und standardisiert und die praktische Handhabung ist kaum oder gar nicht beschrieben. Es fehlen Hilfsmittel wie Handbücher, Vorgehensmodelle und Auswertungsprozeduren, die die Methodenanwendung von der Planung über die Durchführung bis zur Datenauswertung, Interpretation und

Kommunikation anleiten. Dieser Mangel hat zur Konsequenz, dass Praktiker die Anwendung der Methoden ständig neu erfinden müssen 1, S. 108.

Hierunter leidet sowohl die Praktikabilität der Methoden als auch die Objektivität bei der Untersuchungsdurchführung und Datenauswertung. Entsprechend ist die Zuverlässigkeit (Reliabilität) und Gültigkeit (Validität) der Ergebnisdaten meistens stark eingeschränkt. Hierin kann eine Ursache dafür liegen, dass der Einsatz derselben wie auch verschiedener Evaluationsmethoden bei ein und demselben Produkt mit nur geringer Übereinstimmung zur Entdeckung derselben Usability Problemen führt 14. Erschwert werden Bemühungen zur Qualitätssicherung in der Softwareevaluation durch die große Anzahl bestehender Methoden, die mit unterschiedlichen Zielen in verschiedenen Kontexten zur Anwendung kommen. Es fehlt Wissen darüber, zu welchem Zeitpunkt im Entwicklungskontext welche Methode am besten geeignet ist.

Vor diesem Hintergrund erscheinen Maßnahmen für die Qualitätssicherung äußerst wichtig, damit Praktiker bei der

Entscheidung über den Methodeneinsatz unterstützt werden, und mit vertretbarem Aufwand zuverlässige und gültige Ergebnisse erzielen und damit in der Usability Forschung aussagekräftigen Erkenntnisse über die Leistungsstärke unterschiedlicher Evaluationsmethoden gewonnen werden können. Im Folgenden werden vor dem Hintergrund eines allgemeinen Ablaufmodells für formative Evaluationsuntersuchungen Problempunkte aufgezeigt und Maßnahmen zur Verbesserung der Situation vorgeschlagen. Das Ablaufmodell umfasst die folgenden Punkte:

- Planung,
- Durchführung,
- Datenauswertung,
- Interpretation,
- Berichterlegung und Kommunikation der Befunde,
- Definition von Maßnahmen und Erstellen eines Veränderungsplans.

2.0 Ansätze zur Qualitätssicherung

2.1 Planung

Auch wenn an Evaluationsuntersuchungen nicht unbedingt die gleichen Anforderungen wie an experimentelle Studien gestellt werden, so gibt es doch einige Punkte, die bei der Durchführung von Evaluationsuntersuchungen Beachtung finden sollten. Insbesondere sollten bei der Planung die folgenden Aspekte berücksichtigt werden:

- Definition der Zielsetzung der Evaluationsuntersuchung,
- Ableitung von Evaluationskriterien aus Anforderungsanalysen mit Bezugnahme auf Konstrukte der Gebrauchstauglichkeit,
- Bestimmung der zu untersuchenden Merkmale oder Funktionen der Software, auf die die Kriterien angewendet werden sollen,
- Ableitung von Untersuchungsszenarien aus Anforderungsanalysen,
- Auswahl der geeigneten Methode oder Methodenkombination,
- Festlegung des Untersuchungsdesigns und Entscheidung über die adäquate Teilnehmerzahl.

2.2 Durchführung und Datenerhebung

Die Anwendung der Methode(n) sollte – soweit vorhanden – den Durchführungsempfehlungen und Verfahrensrichtlinien folgen. Nur hierdurch kann sichergestellt werden, dass ein Verfahren auch von unterschiedlichen Personen in vergleichbarer Weise ange-

wendet wird. Das Vorgehen sollte während der Versuchsdurchführung genau dokumentiert werden.

Insbesondere bei verhaltensorientierten Evaluationsmethoden ist es ratsam, das Verhalten der Probanden bei der Nutzung des evaluierten Systems aufzuzeichnen, da die Evaluatoren durch gleichzeitige Beobachtung und Registrierung meist überfordert sind. Auch sollte klar spezifiziert werden, welche Verhaltensweisen registriert werden sollen (siehe Hassenzahl und Seewald in diesem Band). Durch unsachgemäße Handhabung von Methoden gehen häufig Informationen verloren 1, S. 108.

2.3 Datenauswertung

Für die wenigsten Methoden zur formativen Evaluation von Software werden Angaben darüber gemacht, wie die erhobenen Daten zu behandeln sind. Dabei ist die für die Qualitätssicherung zentrale Frage bei der Datenauswertung, wie sich die erhobenen Daten so auswerten lassen, dass in stringenter Form Usability-Probleme erkannt werden aus denen sich Gestaltungsvorschläge ableiten lassen.

Bei den in einer formativen Evaluationsuntersuchung erhobenen Daten handelt es sich zumeist um qualitative Daten, die Datenauswertung ist daher in der Regel inhaltsanalytisch ausgerichtet. Die folgenden Schritte lassen sich hierfür unterscheiden.

2.3.1 Identifikation relevanter Informationen/Vorkategorisierung

Die Auswertung des Datenmaterials beinhaltet zumeist in einem ersten Schritt die Identifikation der Informationen, die sich inhaltlich auf Schwächen und Mängel des untersuchten Systems bezie-

hen. Hierzu kann es sinnvoll sein, zunächst die Informationen zu identifizieren, die sich offensichtlich nicht auf das evaluierte System und dessen Bedienung beziehen. Dieser Schritt ist besonders bei dem Einsatz der Methode des Lauten Denkens notwendig, da von den Probanden meist viele Verbalisierungen vorgenommen werden, die nicht problem-relevant sind, wie z.B.

»Strategieaussagen«, »Aussagen über die eigene Person« etc. s. dazu 3.

2.3.2 Explikation

Das Ziel der Explikation besteht darin, ggf. zusätzliches Material heranzutragen, das die verbliebenen Informationen erläutert, erklärt, ausdeutet und deren Verständnis erweitert 13. Dies kann mit Hilfe einer Kontextanalyse geschehen, durch die zusätzliche Informationen aus dem zur Verfügung stehenden Quellen herangezogen werden, um die registrierten Daten interpretieren zu können. Explikationen sind häufig für die weitere Verarbeitung von Daten aus Protokollen des Lauten Denkens notwendig, da die Verbalisierungen selten ohne Informationen über den Handlungszusammenhang und den Referenzbereich der Software zu verstehen sind.

2.3.3 Klassifikation, Kategorisierung und Priorisierung von Problemen

Zu bemerken ist, dass Probanden nur in seltenen Fällen Usability Probleme von sich aus benennen. Vielmehr wird in der Regel durch einen Evaluator von den Beobachtungs- oder Befragungsdaten auf Usability Probleme geschlossen. Voraussetzend hierfür muss definiert worden sein, was unter einem Usability Problem verstanden werden soll. Dies kann sowohl mit Bezug auf bestimm-

te Gestaltungsziele sowie Ereignisse, die die effektive und effiziente Aufgabenbearbeitung behindern 10, geschehen als auch mit Bezug auf Auslösebedingungen für das Auftreten von Fehlhandlungen.

Um aus den erhobenen Daten Aussagen über Problempunkte einer evaluierten Software ableiten zu können, werden in der Literatur verschiedene Kategorisierungsansätze vorgeschlagen. Usability Probleme lassen sich demnach z.B. nach inhaltlichen Aspekten wie z.B. Art der begangenen Handlungsfehler, Phasen im Handlungsprozess 9, der Problemspezifität 5 oder ihrer Bedeutsamkeit 10; 15 kategorisieren. Ein Problempunkt bei der Kategorisierung, insbesondere bei der Priorisierung von Usability-Problemen ist die oft niedrige Reliabilität der Kategoriensysteme 9; 10, die zumeist durch die Übereinstimmung zweier oder mehrerer Rater bei der Zuordnung von Problemen zu Problemklassen überprüft wird.

Einen Ansatz zur Qualitätssicherung für die Datenkategorisierung stellen Hartson et al. 9 vor: Im Rahmen ihres Konzepts des User Action Framework (UAF) formulieren sie ein Klassifikationssystem, dass auf dem Konzept des Interaction Cycle – eines Modells der Mensch-Maschine-Interaktion 16 – basiert. Das UAF sieht vor, die in Usability Tests erhaltenen Nutzungsprobleme den unterschiedlichen Phasen im Interaction Cycle (siehe Abbildung 1) zuzuordnen und dann in vier weiteren Klassifikationsschritten detaillierteren Unterkategorien zuzuordnen. Die Anwendung dieses Kategoriensystems verspricht eine deutliche Verbesserung der Gütekriterien, wie bei der Klassifikation von Verbesserungsvorschlägen und Nutzungsproblemen empirisch belegt wurde 1.

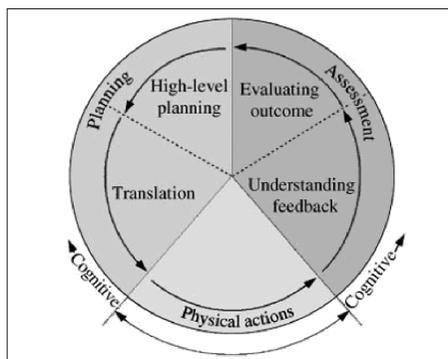


Abbildung 1: UAF Ratingmodell

2.4 Berichtlegung

Berichte über die Ergebnisse einer Evaluationsuntersuchung, so genannte »Usability Reports« s. 4; 17 haben unterschiedliche Funktionen. Zu denen zählt, die Befunde der Evaluationsuntersuchung zu kommunizieren, veränderungsbezogenes Handeln anzuleiten, Veränderungen des evaluierten Systems zu unterstützen und natürlich auch, die getane Arbeit zu belegen 17. Da sich die Berichte häufig an Personen richten, die keine Experten aus dem Bereich der Ergonomie sind, sondern in der Regel Systementwickler 10, sollten die Berichte auch für diese Zielgruppe verständlich geschrieben sein. In einem »Usability Report« sollten nach einem »executive summary« die Zielsetzung und der genaue Ablauf der Untersuchung dargelegt werden. Es wird vorgeschlagen, zu Beginn des Ergebnisteils eine Zusammenfassung der Untersuchungsbefunde zu präsentieren 4. Die detaillierte Ergebnisdarstellung sollte die identifizierten Problempunkte und ggf. die durch die Evaluatoren formulierten Veränderungsvorschläge - geordnet nach inhaltlichen Kategorien und nach Prioritäten - beinhalten. Problembeschreibungen und Problemlösungen sollten getrennt dargelegt werden (s. 10).

Derzeit gibt es Standardisierungsbemühungen für die Berichtlegung von Usability Tests durch das Common Industry Format (CIF) 18.

Ziel des CIF ist die Steigerung der Transparenz von Usability Tests. Durch CIF wird nicht der Gegenstand des Usability Tests spezifiziert, sondern welche Inhalte in der Berichtlegung über summative Evaluationsuntersuchungen zu berücksichtigen sind. Hierzu gehören die Evaluationsaufgaben der Testpersonen, die genutzten Geräte und Software, die Evaluationsprotokolle, die Ergebnisse der Evaluation, die Spezifizierung der Evaluationsmethoden sowie die Datenerhebungsmechanismen. Eine Erweiterung von CIF auf Internet Usability Tests, Hardware Usability Berichtlegung und die Kommunikation von Usability-Untersuchungen ist geplant

2.5 Vermittlung und Kommunikation

Zur Vermittlung der Befunde aus Evaluationsuntersuchungen eignen sich Usability Reviews 6. Das Ziel eines Usability-Reviews besteht darin, ausgehend von den priorisierten und kategorisierten Problembeschreibungen Optimierungsvorschläge zu erarbeiten und daraus Entwicklungsbedarf abzuleiten.

Teilnehmer an einem Usability Review sollten ein Moderator, Entwickler, ein Ergonom und wenigstens ein Nutzer des Systems oder dessen Vertreter sein. Zu Beginn eines Reviews berichtet der Ergonom über die Ergebnisse der Evaluation. Die Ergebnispräsentation einer formativen Evaluation stützt sich im Wesentlichen auf die Kategorisierung und Priorisierung der erkannten Problempunkte.

Nach der Ergebnispräsentation werden Vorschläge für die Behandlung der Problempunkte vom Ergonom unterbreitet, in der Gruppe diskutiert, ggf. Gegenvorschläge gesammelt und über die Behandlung entschieden. Zum Abschluss des Reviews werden die Ergebnisse

zusammengefasst und als ein nachbereitender Schritt ein Aktionsplan für die Veränderung der Software festgelegt.

3.0 Diskussion

In diesem Beitrag wurde die Bedeutung von Maßnahmen der Qualitätssicherung für Evaluationsuntersuchungen, wie sie im Zusammenhang mit der ergonomischen Gestaltung von Software praktiziert werden, dargelegt. Es stellt sich die Frage, warum diese sowohl in der Praxis als auch in der Forschung selten berücksichtigt werden. Eine mögliche Antwort liegt im Spannungsfeld von zeitlichem und organisatorischem Aufwand für die Implementierung von qualitätssichernden Elementen und den »Anforderungen der Praxis«: Zeitdruck und begrenzte Mittel hindern oft daran, für Vorbereitung, Durchführung und Dokumentation von Untersuchungen die notwendigen Ressourcen bereitzustellen. Wenn dann noch die Ansicht hinzukommt, mit weniger Mitteln vergleichbare Ergebnisse erzielen zu können, dann wundert es kaum, dass aus pragmatischen Gründen Qualitätsaspekte vernachlässigt werden.

Wie kann dieser Tendenz begegnet werden? Zunächst fehlen Rahmenmodelle, die den gesamten Evaluationsprozess konsistent unterstützen. Das UAF könnte ein solches Rahmenmodell bieten, wenn das Modell weiter ausformuliert und erprobt wird.

Darüber hinaus fehlen für den Großteil der bestehenden Evaluationsmethoden genaue Verfahrens- bzw. Evaluationsmodelle, Manuale und Hilfsmittel wie z.B. für die Datenaus- und -verwertung. Weiterhin fehlt es an reliablen und validen Erkenntnissen, die die Effizienz und Effektivität bestehender Methoden aufzeigen und Empfehlungen über den Anwendungszeitpunkt im Entwicklungsprozess oder über effektive

Methodenkombinationen ermöglichen. Gleichzeitig bedarf es auch Schritte zur Vermittlung der Erkenntnis, dass Evaluationsuntersuchungen nur dann von hohem Nutzen sind, wenn ihre Ergebnisse zuverlässig und gültig sind. Voraussetzung hierfür ist die Anwendung der Methoden nach den Regeln der Kunst.

4.0 References

- 1 Andre, T.S., Hartson, H.R., Belz, S.M. & McCreary, F.A., The user action framework: a reliable foundation for usability engineering support tools. *Int. J. Human-Computer Studies*, 54, 107-136 (2001).
- 2 Boren, T. & Ramey, J., Thinking Aloud. Reconciling Theory and Practice. *IEEE Transactions on professional communication*, 43, 3, 261-278 (2000).
- 3 Deffner, G., Heydemann, M. & Borstel, G. v., Ein Kategoriensystem und interaktives Ratingverfahren für die Vorverarbeitung von Protokollen des Lauten Denkens. *Archiv der Psychologie*, 136, S. 147-162 (1984).
- 4 Dumas, J.S. & Redish, J.C., *A Practical Guide to Usability Testing (Revised Edition)*. Exter: Intellect Books (1999).
- 5 Gediga, G. & Hamborg, K.-C., Heuristische Evaluation und IsoMetrics: Ein Vergleich. In: R. Liskowsky, B.M. Velichkovsky & W. Wünschmann (Hrsg.), *Software Ergonomie '97, Usability Engineering: Integration von Mensch-Computer-Interaktion und Software-Entwicklung* (S. 145-155). Stuttgart: Teubner (1997).
- 6 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).
- 7 Gediga, G. & Hamborg, K.-C., Ergonomische Evaluation von Software: Methoden und Modelle im Software-Entwicklungsprozess. *Zeitschrift für Psychologie*, 210 (1), 40-57 (2002).
- 8 Gray, W.D. & Salzman, M.C., Damaged Merchandise? A Review of Experiments that Compare Usability Evaluation Methods. *Human-Computer Interaction*, 13, 203-261 (1998).
- 9 Hartson, H.R. Andre, T.S., Williges, R.W. und van Rens, L., The user action framework: A theory-based foundation for inspection and classification of usability problems, *Proceedings of the 8th International Conference on Human-Computer Interaction: Ergonomics and User Interfaces, Vol.1 Mahway: Lawrence Erlbaum Associates, 1058-1062* (1999).
- 10 Jeffries, R., Usability Problem Reports: Helping Evaluators Communicate Effectively with Developers. In: J. Nielsen & R.L. Mack. (eds.), *Usability Inspection Methods* (S. 273-294). New York: J. Wiley & Sons (1994).
- 11 Karat, C.-M., A Comparison of User Interface Evaluation Methods. In: J. Nielsen & R.L. Mack (eds.), *Usability Inspection Methods* (S. 203-233). New York: J. Wiley & Sons (1994).
- 12 Karat, J., User-Centered Software Evaluation Methodologies. In M. Helander, T. K. Landauer & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction* (2nd ed.), (pp. 689-704.). Amsterdam: Elsevier (1997).
- 13 Mayring, P., *Qualitative Inhaltsanalyse* (6., durchgesehene Auflage). Weinheim: Beltz Deutscher Studienverlag (1997).
Molich, R. et al.: Comparative usability evaluation. In: *Behaviour & Information Technology*, 23, 1, 65-74 (2004).
- 14 Nielsen, J., *Usability Engineering*. Boston: AP Professional (1993).
- 15 Norman, D.A. *Cognitive Engineering*. In D.A. Norman & S.W. Draper, Eds. *User Centered System Design: New perspectives on Human-Computer Interaction*, 31-61. Hillsdale: Lawrence Erlbaum Associates (1986).
- 16 Rubin, J., *Handbook of Usability Testing*. New York: John Wiley & Sons, Inc (1994).
- 17 Scholtz, J. & Morse, E., A New Usability Standard and What It Means to You. *SIGCHI Bulletin*. May/June 2002, 11 (2002).

»Es ist erlaubt digitale Kopien in Papierform des ganzen Papers oder Teilen davon für den persönlichen Gebrauch oder zur Verwendung in Lehrveranstaltungen zu erstellen. Der Verkauf oder gewerbliche Vertrieb ist untersagt. Rückfragen sind zu stellen an den Vorstand des GC-UPA e.V. (Postfach 80 06 46, 70506 Stuttgart).
Proceedings of the 2nd annual GC-UPA Track Paderborn, September 2004
© 2004 German Chapter of the UPA e.V.«