

Gestaltungsunterstützende Evaluation von Software: Zur Effektivität und Effizienz des IsoMetrics^L Verfahrens

Kai-Christoph Hamborg

Universität Osnabrück

Abstract

Für die gestaltungsunterstützende Evaluation von Software stellen sogenannte Inspektionsverfahren eine ökonomische Alternative zu Testmethoden dar. Neben ökonomischen Aspekten ist aber für den Methodeneinsatz auch die Effektivität und Effizienz der Methoden in Bezug auf die Fähigkeit, gestaltungsrelevante Informationen zu erheben, entscheidend. In diesem Beitrag wird die Effektivität und Effizienz des IsoMetrics^L Verfahrens als fragebogengestützte Inspektionsmethode im Vergleich zu Testmethoden und der Methode der Heuristischen Evaluation untersucht. Die Ergebnisse zeigen, dass IsoMetrics^L bei geringeren Kosten mehr und auch schwerwiegendere Systemmängel als die Vergleichsmethoden erkennt.

1 Einleitung

Die formative bzw. gestaltungsunterstützende Evaluation ist ein wichtiger Bestandteil iterativer Softwareentwicklung (s. z.B. ISO 13407, 1999). Ihr Ziel ist die ergonomische Qualitätsverbesserung und –sicherung. Erhoben werden erstens quantitative Daten, um die Realisierung von Benutzbarkeitszielen im Entwicklungsprozess zu überprüfen und zweitens qualitative Daten über Schwächen eines Produktprototyps, aus denen Maßnahmen zur Verbesserung der Gebrauchstauglichkeit abgeleitet werden sollen (s. Wright & Monk 1989, Hix & Hartson 1993, Karat 1997).

Zu den am häufigsten eingesetzten Vorgehensweisen im Bereich der formativen Evaluation zählen Nutzertests (s. Rubin 1994, Dumas & Redish 1999). Nutzertests haben sich als wirkungsvolle Methodik etabliert, deren geringe Durchführungsökonomie jedoch bemängelt wird (s. Karat 1997, Virzi 1997). Dem Bedarf an einfach und schnell einzusetzenden Evaluationsmethoden (Oppermann & Reiterer 1994) wird damit nicht entsprochen. Diese Kritik ist besonders darum ernst zu nehmen, da die Verfahrensökonomie als ein wichtiges Bewertungskriterium für die Praxistauglichkeit“ von Evaluationsmethoden gilt (Hampe-Neteler 1994, S. 130).

Eine kostengünstige Alternative zu Testmethoden bieten Inspektionsmethoden (s. Nielsen & Mack 1994). Vergleichuntersuchungen zur Effektivität und Effizienz von Inspektions- und Testmethoden zeigen widersprüchliche Ergebnisse. Während Befunde von Jeffries et al. (1991) sowie Virzi et al.(1993) zugunsten von Inspektionsmethoden ausfallen, bewerten Karat et al. (1992) Inspektionsmethoden im Vergleich zu Testmethoden als weniger effektiv. Zu den Inspektionsmethoden zählt auch das fragebogengestützte IsoMetrics-Verfahren (Gediga, Hamborg & Düntsch 1999, Gediga & Hamborg 1999). Zur Klärung der Leistungsfähigkeit dieses Verfahrens wurden Untersuchungen durchgeführt, in denen es mit typischen Test- und Inspektionsmethoden verglichen wurde.

2 Untersuchungen

In mehreren Vergleichsuntersuchungen wurde bereits das „Kosten/Nutzenverhältnis“ von Evaluationsmethoden untersucht, indem der für die Problemgenerierung erforderliche Aufwand (Zeit und Kosten) ins Verhältnis zu der Anzahl und der Qualität erhobener Probleme gesetzt wurde (Jeffries et al. 1991, Karat et al. 1992, Smilowitz et al. 1993, Virzi et al. 1993). Dieser Vorgehensweise wurde in der vorliegenden Untersuchung gefolgt. Als Testmethoden wurden die Methode des lauten Denkens (Lewis 1982) und, als deren Weiterentwicklung, die Methode der Videokonfrontation (Hamborg & Greif 1999) berücksichtigt. Aus dem Bereich der Inspektionsmethoden diente die Heuristische Evaluation (Nielsen 1994) als Vergleichsmethode.

Die *Methode des lauten Denkens* ist eine Methode zur Erfassung bewusster handlungsbegleitender Kognitionen und Emotionen, die Versuchsteilnehmer bei der Nutzung einer Software äußern und von denen auf Problempunkte der Software geschlossen wird (Lewis 1982). Die Äußerungen der Probanden werden in einer einfachen Form mit Papier und Bleistift (s. Nielsen 1993) oder aufwändiger mit Tonband oder auch Videotechnik (Lewis 1982) aufgezeichnet. Videoaufnahmen haben den Vorteil, dass zusätzliche „Kontextinformationen“ registriert werden, die ggf. zum Verständnis der Äußerungen und deren Auswertung herangezogen werden können.

Die *Technik der Videokonfrontation* (Greif et al. 1986, Moll 1987, Hamborg & Greif 1999) sieht vor, dass zunächst eine Arbeitssequenz mit dem zu evaluierenden System per Videotechnik aufgezeichnet und daraufhin in einem Interview mit dem Untersuchungsteilnehmer analysiert wird. Hierbei kommen standardisierte und halbstandardisierte Frageformate zum Einsatz, die sich auf Probleme bei der Interaktion mit der evaluierten Software richten. Bei der Videoaufzeichnung ist in der einfachsten Variante eine Kamera auf den Monitor, an dem gearbeitet wird, gerichtet. Es hat sich jedoch als vorteilhaft erwiesen, zusätzlich die Tastatur und das Gesicht der Versuchspersonen aufzunehmen und auf einem Band zusammenzuschneiden (Hamborg & Greif 1999).

Die *Technik der Heuristischen Evaluation* entspricht im Prinzip einem strukturierten Expertenurteil (Nielsen 1994). Die Evaluation wird durch sog. Heuristiken gelenkt, die Eigenschaften benutzbarer Systeme beschreiben. Nielsen (Nielsen & Molich 1990, Nielsen 1994) schlägt neun Heuristiken vor (Einfacher und natürlicher Dialog, Spreche die Sprache des Benutzers, Minimiere die Gedächtnisbelastung des Benutzers, Sei konsistent, Gebe Rückmeldung, Sorge für schnelle Abbruchpunkte, Unterstütze Abkürzungen, Gute präzise Fehlermeldungen, Verhüte Fehler). Die Software wird jeweils einzeln von einem Evaluator in Bezug auf Abweichungen von den in den Heuristiken formulierten Qualitätsaspekten begutachtet.

IsoMetrics^L ist ein benutzer-orientiertes Evaluationsverfahren, das die 7 Gestaltungsgrundsätze der ISO 9241 Teil 10 durch 75 Items operationalisiert. Es gibt zwei Versionen des Verfahrens, die auf dem selben Itemsatz basieren: *IsoMetrics^S* (short) dient der summativen, *IsoMetrics^L* (long) der formativen Evaluation. In der Kurzversion wird jedes Item auf einer 5-stufigen Ratingskala beantwortet. Die Langversion besitzt zusätzlich ein auf die jeweilige Itemaussage bezogenes „Wichtigkeitsrating“ sowie Freiraum für die Darstellung konkreter auf das Item bezogener Schwächen des Systems. Folgende Informationen werden durch *IsoMetrics^L* erhoben:

- Maßzahlen auf den Benutzbarkeitsdimensionen, die als Anhaltspunkte für den Entwicklungsfortschritt benutzt werden können.
- Konkrete Hinweise auf Fehlfunktionen und Schwachstellen aus Benutzersicht.
- Gewichte von Problemklassen, die empirisch aus der Benutzersicht gewonnen werden.

Grundlage für den Methodenvergleich war die Evaluation eines elektronischen Bibliothekssystems, das zum Untersuchungszeitpunkt über Telnet zugänglich war und entsprechend über eine zeichenorientierte Benutzungsschnittstelle verfügte. Als Indikatoren für die Effektivität der Methoden wurden die Menge und die Qualität der erhobenen Anmerkungen erhoben.

Bei der Evaluation wurde Empfehlungen bezüglich des Einsatzes der Methoden, insbesondere des Untersuchungsablaufs, der Anzahl der Evaluatoren und, im Falle der Heuristischen Evaluation auch deren Qualifikation, so weit wie möglich entsprochen. Für Usability Tests werden unterschiedliche Angaben über die erforderliche Stichprobengröße gemacht. Nielsen (1993, S. 224) schlägt 3-5 Personen für die Methode des lauten Denkens vor, Rubin (1994) ca. 10 Personen. In der vorliegenden Untersuchung nahmen jeweils 8 Personen an der Evaluation mit der Methode des lauten Denkens und der Videokonfrontation teil. Für IsoMetrics^L wurde eine im mittleren Empfehlungsbereich (8 - 20 Evaluatoren, Gediga & Hamborg 1999, S. 230) liegende Stichprobengröße von 15 Personen gewählt. Die Heuristische Evaluation sollte laut Nielsen (1993) mit 3-5 Evaluatoren, die über Expertise im Bereich der Mensch-Computer Interaktion resp. Software-Ergonomie verfügen, durchgeführt werden. In der hier dargestellten Studie nahmen fünf Evaluatoren teil. Sie hatten zuvor Kurse zum Thema „Software-Ergonomie“ besucht oder entsprechende Konzepte bei der Entwicklung eigener Applikationen angewendet. Damit entspricht ihre Qualifikation wenigstens der Qualifikation derjenigen, mit deren Hilfe der Nachweis der Effektivität der Methode erbracht wurde (Informatikstudierende, die einen Kurs zum Thema „User Interface Design“ besucht hatten, und Leser einer Computerzeitschrift, s. Nielsen & Molich 1990).

2.1 Untersuchungsdurchführung und Datenauswertung

Alle Untersuchungen wurden von geschulten studentischen Hilfskräften geleitet. Der Untersuchungsverlauf mit den Testmethoden und IsoMetrics^L gestaltete sich wie folgt: Zunächst wurde mittels Fragebogen die computerbezogene Vorerfahrung der Probanden erhoben. Daraufhin erhielten sie 15 Minuten Zeit, das System zu explorieren und bearbeiteten anschließend vier typische Aufgaben (unterschiedlich komplexe Datenbankabfragen). Die Probanden wurden für die Untersuchung nicht speziell in das System eingeführt, erhielten jedoch ein kurzes Bedienungshandbuch, in dem die grundlegenden Befehle des Systems erläutert waren. Nach den Aufgabebearbeitungen wurde die Software mit den jeweiligen Methoden evaluiert.

Die Evaluatoren, die mit der Heuristischen Evaluation arbeiteten, wurden instruiert, die Nutzungsmöglichkeiten des Programms umfassend zu erkunden. Bei der anschließenden Bewertung sollten Abweichungen des Programms von den Heuristiken so genau wie möglich und unter Angabe der Systemeigenschaften, auf die sie sich beziehen, angemerkt werden.

Die Auswertung des Datenmaterials umfasste die Identifikation der Informationen, die sich auf Schwächen und Mängel des untersuchten Systems bezogen. Dazu wurden zunächst aus der Gesamtmenge aller erhobenen Informationen (s. Tabelle 1, Spalte: Informationen insgesamt) diejenigen eliminiert, die sich *nicht* auf das evaluierte System und dessen Bedienung bezogen. Gegebenenfalls wurden die extrahierten Anmerkungen ergänzt (Explikation), so dass sie durch Dritte ohne zusätzliche Informationen verständlich waren. Bei der Methode des Lauten Denkens und der Videokonfrontation geschah dies auf Basis der Videoaufzeichnungen und der registrierten Tastatureingaben. Da weder bei der Heuristischen Evaluation noch bei IsoMetrics^L die Aufgabebearbeitungen aufgezeichnet wurden, basierte die Explikation im ersten Fall nur auf den schriftlich verfügbaren Informationen, bei IsoMetrics^L auf den Items des Fragebogens.

3 Ergebnisse

3.1 Anzahl der durch die Methoden erhobenen Anmerkungen

Bei der Mengenbestimmung der erhobenen Anmerkungen wurden redundante, d.h. mehrfach genannte Anmerkungen nicht berücksichtigt, da sie keine neuen Informationen beinhalten.

Die Methoden unterscheiden sich in Bezug auf die Menge der erhobenen Anmerkungen (s. Tabelle 1, Spalte: Anm. ohne Redundanzen) signifikant ($\chi^2 = 239,15$, $df = 3$, $p = ,000$). Die Inspektionsmethoden (Heuristische Evaluation, IsoMetrics^L) verfügen über ein deutlich besseres Verhältnis von erhobenen Informationen (Tabelle 1, Spalte: Anmerkungen insgesamt) zu verwertbaren Anmerkungen (Tabelle 1, Spalte: Anm. mit und ohne Redundanzen) als die Testmethoden (Lautes Denken, Videokonfrontation).

Methoden	N	Informationen insgesamt	Anm. mit Redundanzen	Anm. ohne Redundanzen	Anm./Evaluator (ohne Redundanzen)
Lautes Denken	8	453	61	48	6,00 (SD = 5,1)
Video-konfrontation	8	131	69	58	7,25 (SD = 6,1)
Heuristische Evaluation	5	88	88	79	15,80 (SD = 3,2)
IsoMetrics ^L	15	345	310	245	16,33 (SD = 10,1)

Tabelle 1: Anmerkungshäufigkeiten, alle Methoden

Im Einzelvergleich zeigt sich, dass das Mengenverhältnis der erhobenen redundanzfreien Anmerkungen von Heuristischer Evaluation und IsoMetrics bei ca. 3 zu 1 zugunsten von IsoMetrics^L liegt. Wird die Menge der Anmerkungen an der Anzahl der Evaluatoren relativiert, ist der personenbezogene Output bei Heuristischer Evaluation und IsoMetrics etwa gleich groß (s. Tabelle 1), wobei zu bemerken ist, dass es sich bei den Evaluatoren, die mit der Heuristischen Evaluation arbeiteten, um Experten handelte, bei IsoMetrics^L dagegen um in Bezug auf software-ergonomisches Wissen „naive“ Nutzer.

Der entsprechende Vergleich der Methode des Lauten Denkens und der Videokonfrontation mit IsoMetrics^L zeigt, dass das Mengenverhältnis verwertbarer Anmerkungen bei ca. 5 bzw. 4 zu 1 zugunsten von IsoMetrics^L liegt. Auch bei Relativierung der erhobenen Anmerkungen an der Zahl der Untersuchungsteilnehmer zeigt sich, dass durch IsoMetrics^L mehr als doppelt so viele Anmerkungen wie durch die Methode des Lauten Denkens und die Videokonfrontation erhoben werden.

3.2 Bedeutsamkeit der Probleme

Die formative Evaluation von Software führt in der Regel zu einer Liste von Anmerkungen die Schwachpunkte, Probleme oder Mängel des Produkts beschreiben. Für die weitere Verwertung der Anmerkungen ist es notwendig, diese nach ihrer Bedeutsamkeit bzw. ihrem Problemgehalt zu

priorisieren (s. Nielsen 1993), da sie häufig von unterschiedlicher Qualität sind. Hierzu wurden die erhobenen Anmerkungen mit einer von Nielsen (1993) formulierten Skala bewertet, die zuvor in die deutsche Sprache übersetzt und um einen Punkt („kann ich nicht beurteilen“) ergänzt wurde (s. Tabelle 2).

Prioritätsstufe	Erläuterung
Benutzbarkeitskatastrophe (1)	Das Problem muss nach software-ergonomischen Gesichtspunkten auf jeden Fall zum nächstmöglichen Zeitpunkt behoben werden.
Schwerwiegendes Benutzbarkeitsproblem (2)	Es ist aus software-ergonomischer Sicht sehr wichtig, dass das Problem behoben wird. Es sollte eine hohe Priorität erhalten.
Kleineres Benutzbarkeitsproblem (3)	Die Behebung des Problems sollte nach software-ergonomischen Gesichtspunkten eine geringere Priorität haben.
Kosmetisches Problem (4)	Nach software-ergonomischen Gesichtspunkten muss das Problem nicht unbedingt behoben werden, es sei denn, im Projekt ist noch extra Zeit dazu vorhanden.
Kein Benutzbarkeitsproblem (5)	Nach meiner Auffassung handelt es sich hierbei nicht um eine Benutzbarkeitsproblem (z.B. positive Bewertungen)
Kann ich nicht beurteilen (6)	Der Anmerkungstext ist unverständlich oder aus anderen Gründen kann eine Problem diagnose weder getroffen noch abgelehnt werden.

Tabelle 2: Skala zur Bestimmung der Bedeutsamkeit der Anmerkungen (nach Nielsen 1993)

Nielsen (1994) berichtet aus eigenen Untersuchungen als Übereinstimmungsmaß für die Bedeutsamkeitsbestimmung mit seiner Skala bewerteter Anmerkungen eine signifikante mittlere Korrelation (Kendall's Konkordanz Koeffizient, $W = .31$, $Chi^2 = 123,3$, $df = 39$, $p < .01$, s. Nielsen 1994, S. 49) bei elf Beurteilern und schließt daraus, dass die Übereinstimmung der Urteiler nicht zufällig ist. Jedoch merkt er an, dass die Interrater-Reliabilität vergleichsweise niedrig ist. Als Auswertungsprozedur empfiehlt er daher, den Mittelwert aus mehreren Ratings zu bilden, um die Reliabilität der Beurteilung zu erhöhen (Nielsen 1994, S. 50).

Das Bedeutsamkeitsrating wurde in den vorliegenden Untersuchungen von drei Ratern durchgeführt. Bei der Auswertung wurden nur Anmerkungen berücksichtigt, die entsprechend der angegebenen Skalierung von wenigstens 2 der 3 Rater einen Wert zwischen 1 und 5 erhielten und bei denen Minimum und Maximum nicht mehr als zwei Skalenpunkte auseinander lag. Gaben zwei der drei Rater an, die Anmerkung nicht priorisieren zu können, wurde die Anmerkung ebenfalls nicht berücksichtigt. Nach Anwendung dieser Ausschlusskriterien konnten 81,6 % (IsoMetrics^L), 91,1 % (Heuristische Evaluation), 97,9% (Methode des Lauten Denkens) und 98,3% (Videokonfrontation) der redundanzfreien Anmerkungen ausgewertet werden.

Wie bei Nielsen (1994) ist die Interrater-Reliabilität recht gering. Der „Intra-Class“-Korrelationskoeffizient beträgt $r_1 = 0,38$ für einen und $r_k = 0,64$ für alle Rater (s. Bortz & Döring 1995, S. 252f). Entsprechend des Vorschlags von Nielsen wurden daher die Mittelwerte aus den Ratings gebildet. Demnach wurden die durch die Methode der Videokonfrontation erhobenen Anmerkungen am höchsten priorisiert ($M = 2,75$), gefolgt von IsoMetrics^L ($M = 2,97$) der Methode des Lauten Denkens ($M = 3,00$) und der Heuristischen Evaluation ($M = 3,18$). Die mittlere Bedeutsamkeit der durch die Methoden erhobenen Anmerkungen unterscheidet sich jedoch nicht statistisch bedeutsam (KW-Test; $Chi^2 = 5,258$, $df = 3$, $p = ,154$).

Da durch die Testmethoden und die Heuristische Evaluation die Kategorie „Benutzbarkeitskatastrophe“ kaum abgedeckt wurde, wurden die Anmerkungen auf Basis der gemittelten Ratings den größeren Problemkategorien: 1. Katastrophe/schwerwiegendes Problem, 2. kleineres/kosmetisches Problem und 3. kein Problem zugewiesen.

Ratingintervall	Katastrophe/schwerwiegendes Problem (1-2,4)	kleineres/kosmetisches Problem (2,41-4,4)	kein Problem Rating (4,41 – 5)
Heuristische Evaluation	19	42	11
Lautes Denken	17	22	7
Videokonfrontation	16	40	1
IsoMetrics ^L	65	108	27

Tabelle 3: Bedeutsamkeit der Anmerkungen

Die Verteilung der Anmerkungen (s. Tabelle 3) zeigt, dass durch die untersuchten Methoden im *relativen* Vergleich nicht unterschiedlich viele Anmerkungen innerhalb der Bedeutsamkeitskategorien erhoben wurden ($\chi^2 = 10,31$, $df = 6$, $p = ,112$). Im absoluten Vergleich wurden mit IsoMetrics^L jedoch sowohl deutlich mehr Benutzbarkeitskatastrophen als auch kleinere und kosmetische Probleme, aber auch mehr Anmerkungen ohne Problemgehalt als durch die Vergleichsmethoden erhoben. Die Unterschiede sind statistisch signifikant (Katastrophe/schwerwiegendes Problem: $\chi^2 = 58,42$, $df = 3$, $p < ,000$, kleineres/kosmetisches Problem: $\chi^2 = 80,69$, $df = 3$, $p < ,000$, kein Problem: $\chi^2 = 32,26$, $df = 3$, $p < ,000$). Auffällig ist die vergleichsweise geringe Anzahl von Anmerkungen ohne Problemgehalt bei der Methode der Videokonfrontation.

3.3 Effizienz der Methoden

Zur Bestimmung der Effizienz der Methoden wurden vorbereitende Tätigkeiten (Entwicklung der Testaufgaben, technische Vorbereitungen, wie die Vorbereitung der Laborräume, Überprüfung der Rechner, Einrichtung der Videoanlage und der Audioaufnahmegeräte), die Untersuchungsdurchführung und die Ergebnisaufbereitung und -analyse (nachbereitende Tätigkeiten) berücksichtigt. Der Zeitaufwand und die daraus folgenden Kosten für diese Tätigkeiten wurden auf Grund von Befragungen der Versuchsleiterinnen und eines Ergonomieexperten geschätzt¹.

Jede der hier berücksichtigten Methoden kann durch geschulte Hilfskräfte/Assistenten durchgeführt werden (10 €/Std). Der Zeitaufwand ist für die Videokonfrontation auf Grund des videogestützten Interviews am höchsten (3,25 Std), für die Methode des Lauten Denkens ist er um eine Stunde geringer (2,25 Std). Für die Heuristische Evaluation ist der zeitliche Aufwand recht gering, da vier Evaluatoren durch eine Untersuchungsleiterin parallel betreut werden können. Bei einem Zeitaufwand von 4 Stunden für eine Heuristische Evaluationsuntersuchung ergibt sich pro Evaluator ein durchschnittlicher Zeitbedarf von einer Stunde, für 5 Evaluatoren also ein Zeitbedarf von 5 Stunden. Bei der Evaluation mit IsoMetrics^L können ebenfalls 4 Untersuchungsteilnehmer gleichzeitig betreut werden. Für eine Untersuchung mit 15 Personen ergibt sich damit ein Zeitaufwand von ca. 10 Stunden bei einer Dauer von 2,5 Stunden pro Untersuchung.

¹ Alle Kosten wurden ursprünglich in DM bestimmt und erst nachträglich in Euro umgerechnet (gerundet)

Aktivitäten	Lautes Denken	Videokonfrontation	Heuristische Evaluation	IsoMetrics
Techn. Vorbereitung	2 Std	2,5 Std	1 Std	1 Std
Durchführungszeit	2,25 *8 = 18 Std.	3,25 *8 = 26 Std.	5 Std.	10 Std.
Analysezeit	6 *8 = 48 Std	4,5 *8 = 36 Std	1,5 *5 = 7,5 Std	1,5*15 = 22,5 Std
Kosten				
Evaluatoren/-Probanden	616,00 € (8 Pbn)	616,00 € (8 Pbn)	2429,00 € (5 Evaluatoren)	1.155,00 € (15 Pbn)
Techn. Vorbereitung	20,00 €	25,00 €	10,00 €	10,00 €
Durchführung	180,00 €	260,00 €	50,00 €	100,00 €
Auswertung/Analyse	4.800,00 €	3.600,00 €	750,00 €	2.250,00 €
Summe	5.616,00 €	4.501,00 €	3.239,00 €	3.515,00 €
Katastrophe/schwerwiegende Probleme	17	16	19	65
Kosten pro Katastrophe/schwerwiegendem Problem	330,35 €	281,31 €	170,47 €	54,08 €

Tabelle 4: Kostenberechnung für Evaluationsmethoden

Die technischen Vorbereitungen sind für die Testmethoden (Lautes Denken, Videokonfrontation) deutlich höher als für die Inspektionsmethoden. Dies ist in erster Linie auf die Einrichtung der erforderlichen Ton- bzw. Videotechnik zurückzuführen. Der Aufwand für die technischen Vorbereitungen wurde nur einmal in die Kostenkalkulation aufgenommen, da sie in der Regel nur zu Beginn einer Untersuchungsserie, danach aber nicht mehr im vollem Umfang erforderlich sind.

Die Datenanalyse und Auswertung erfordert geschulte Fachkräften (100 €/Std). Wie der Durchführungsaufwand ist auch der Aufwand für die Analyse der durch die Testmethoden erhobenen Informationen höher als für die Inspektionsmethoden. Dies ist u.a. darauf zurückzuführen, dass der Kontext der durch die Probanden vorgenommenen Äußerungen häufig nachträglich auf Basis der Videoaufzeichnungen erschlossen werden musste. Insbesondere muss auf die erforderliche aufwändige inhaltsanalytische Analyse der mit der Methode des Lauten Denkens erhobenen Äußerungen hingewiesen werden: sie waren zum großen Teil unverständlich, wenig geordnet und bestanden oft nur aus Teilsätzen, meistens ohne expliziten Bezug zu dem evaluierten Programm. Aufgrund des halbstandardisierten Vorgehens bei dem videogestützten Interview sind die Daten der Videokonfrontation stärker vorstrukturiert und erfordern daher einen geringeren Auswertungsaufwand. Bei der Heuristischen Evaluation wird der Kontext in den meisten Fällen durch die Evaluatoren selbst hergestellt, dies gilt ebenfalls für die Befragung mit IsoMetrics^L.

Die Kosten für die Untersuchungsteilnehmer der Untersuchungen mit den Testmethoden und mit IsoMetrics^L wurden mit 77 € pro Person berechnet. Für die Evaluatoren (i.d.R. Spezialisten aus dem Bereich der Software-Ergonomie), die mit der Heuristischen Evaluation arbeiteten, wurde ein halber Arbeitstag mit 486 € (also ein Tagessatz von 972 €) berechnet.

Die Kostenanalyse zeigt, dass die Heuristische Evaluation die günstigste der untersuchten Methoden ist, gefolgt von IsoMetrics^L, der Videokonfrontation und dem Lauten Denken. Werden die

Kosten mit der Anzahl der hoch bedeutsamen Probleme (Benutzbarkeitskatastrophen, schwerwiegende Probleme) in Beziehung gesetzt, muss jedoch IsoMetrics^L als die kosteneffizienteste Methode bewertet werden (s. Tabelle 4). Die Kosten für eine Benutzbarkeitskatastrophe oder ein schwerwiegendes Problem, erhoben mit IsoMetrics^L, betragen ca. 54 €, mit der Heuristischen Evaluation schon ca. das Drei- und mit der Videokonfrontation oder der Methode des Lauten Denkens mehr als das Fünf- bzw. das Sechsfache. Hierbei ist zu beachten, dass die Kosten für die erforderliche technische Ausrüstung nicht mit in die Analyse einbezogen wurden. Sie liegen für die Methode der Videokonfrontation deutlich am höchsten, da die Datenaufzeichnung mehrere Kameraeinstellungen vorsieht. Im Falle der Methode des Lauten Denkens reicht grundsätzlich eine Kameraeinstellung.

4 Zusammenfassung und Diskussion

Das Ziel der hier dargestellten Untersuchungen bestand darin, Aussagen über die Effektivität und Effizienz des IsoMetrics^L Verfahrens im Vergleich zu anderen etablierten Evaluationsmethoden vorzunehmen. Die Ergebnisse lassen sich folgendermaßen zusammenfassen:

Durch IsoMetrics^L werden im *absoluten* Vergleich 3 – 5 mal mehr Anmerkungen als durch die Vergleichsmethoden erhoben. Selbst wenn die Anmerkungshäufigkeiten an der Anzahl der Evaluatoren *relativiert* werden, erfasst IsoMetrics^L gut doppelt so viele Anmerkungen wie die Methode des Lauten Denkens und die Videokonfrontation und etwa gleich viele Anmerkungen wie die Heuristische Evaluation. IsoMetrics^L kann damit als die effektivste der untersuchten Methoden bezeichnet werden.

Die durch die untersuchten Evaluationsmethoden erhobenen Anmerkungen sind im Mittel nicht unterschiedlich bedeutsam. Auch kann im *relativen* Vergleich keine methodenspezifische Konzentration auf eine der unterschiedlichen Bedeutsamkeitskategorien festgestellt werden. Die Analyse der *absoluten* Anmerkungshäufigkeiten zeigt aber, dass durch IsoMetrics^L insgesamt die meisten Benutzbarkeitskatastrophen und schwerwiegenden sowie kleinen und kosmetischen Probleme erhoben werden aber auch die meisten Anmerkungen ohne Problemgehalt. Die für die Verwertung von Evaluationsbefunden wichtige Bestimmung des Problemgehalts der erhobenen Anmerkungen ist insgesamt nicht unproblematisch. Die Zuordnung zu Problemklassen fällt (s. a. Nielsen 1994) wenig reliabel aus. Daher sollten die Befunde vorsichtig interpretiert werden. Da Problempriorisierungen für die Erforschung der Effektivität von Evaluationsmethoden als auch für deren Anwendung eine wichtige Rolle spielen, ist an dieser Stelle weiterer Forschungsbedarf angezeigt.

Die Befunde zur Effizienz der Methoden zeigen Stärken der Heuristischen Evaluation und des IsoMetrics^L-Verfahrens im Vergleich zu den Testmethoden. Letztere sind deutlich teurer, obwohl die Anzahl der erforderlichen Untersuchungsteilnehmer im Verhältnis zu IsoMetrics^L geringer ist. Wird die Anzahl der erhobenen Benutzbarkeitskatastrophen und schwerwiegenden Probleme ins Verhältnis zu den Einsatzkosten gesetzt, ist IsoMetrics^L deutlich die effizienteste Methode. Pro Benutzbarkeitskatastrophe oder schwerwiegendem Problem fallen für IsoMetrics^L nur ein Drittel und weniger der für die Vergleichsmethoden erforderlichen Kosten an. Die Untersuchungsergebnisse bestätigen damit, dass Testmethoden aufwändiger als Inspektionsmethoden sind (s.a. Jeffries et al. 1991, Karat et al. 1992, Virzi et al. 1993, Smilowitz et al. 1993) aber auch, dass Inspektionsmethoden eine effiziente Alternative zu Testmethoden darstellen.

Über die bisherige Darstellung hinaus bestehen die Stärken von IsoMetrics darin, dass es sich problemlos am Einsatzort der Software unter Bezugnahme auf den Nutzungskontext einsetzen lässt und direkte sowie aktive Formen der Partizipation unterstützt. Es erfordert keine aufwändige Ausstattung oder Ergonomie-Experten als Evaluatoren. Weiterhin basiert das Verfahren auf der internationalen Norm ISO 9241 Teil 10 (1996), auf die auch bei der rechtlichen Umsetzung der EU Bildschirmrichtlinie (EWG 90/270/EWG) durch die Bildschirmarbeitsverordnung Bezug genommen wird und die damit ein relevantes Konstrukt für die Gestaltung von Software darstellt. Schwächen des Verfahrens bestehen im Vergleich zu Testmethoden darin, dass kein unmittelbarer Verhaltensbezug bei der Systemevaluation besteht.

Grundsätzlich sprechen die Befunde der Vergleichsuntersuchung für einen kombinierten Methodeneinsatz, durch den die unterschiedlichen Stärken der Methoden genutzt werden können (und auch darum, weil die Anzahl der durch die Methoden gemeinsam erkannten Probleme recht gering ist, s. Hamborg 2001). Sinnvoll erscheint insbesondere eine Kombination von IsoMetrics^L mit der Methode der Videokonfrontation, durch die viele qualitativ hochwertige und wenig irrelevante Anmerkungen erhoben werden.

5 Literatur

- Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation für Sozialwissenschaftler* (2. Auflage). Berlin: Springer.
- Dumas, J.S. & Redish, J.C. (1999). *A Practical Guide to Usability Testing*. Revised Edition. Exter: Intellect Books.
- DIN EN ISO 9241-10 (1996). *Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten*. Teil 10: Grundsätze der Dialoggestaltung. Berlin: Beuth.
- Gediga, G. & Hamborg, K.-C. (1999). IsoMetrics: Ein Verfahren zur Evaluation von Software nach ISO 9241-10. In: H. Holling & G. Gediga (Hrsg.): *Evaluationsforschung*. Göttingen: Hogrefe, S. 195 – 234.
- Gediga, G., Hamborg, K.-C. & Düntsch, I. (1999). The IsoMetrics Usability Inventory: An operationalisation of ISO 9241-10, *Behaviour and Information Technology*, 18, S. 151-164.
- Greif, S. Tolksdorf, M. & Monecke, U. (1986). *Heterarchische Aufgabenanalyse. Eine Methode zur Untersuchung der Interaktionsprozesse am Computer*. Referat zum 35. Kongreß der Deutschen Gesellschaft für Psychologie in Heidelberg.
- Hamborg, K.-C. (2001). *Gestaltungsorientierte Evaluation von Software*. Habilitationsschrift. Osnabrück: Fachbereich 8, Humanwissenschaften der Universität Osnabrück.
- Hamborg, K.-C. & Greif, S. (1999). Heterarchische Aufgabenanalyse. In: H. Dunckel (Hrsg.). *Handbuch psychologischer Arbeitsanalyseverfahren*. Zürich: vdf, S. 147-177.
- Hampe-Neteler, W. (1994). *Software-ergonomische Bewertung zwischen Arbeitsgestaltung und Softwareentwicklung*. Frankfurt a.M.: Lang.
- Hix, D. & Hartson, H.R. (1993). *Developing User Interfaces*. New York: John Wiley & Sons.
- ISO 13407 (1999). *Human-centred design processes for interactive systems*. Genf: ISO.
- Jeffries, R., Miller, J.R., Wharton, C. & Uyeda, K.M. (1991). User Interface Evaluation in the Real World: A Comparison of Four Techniques. *Proceedings ACM CHI '91 Conference on Human Factors in Computing Systems*. New York: ACM, S. 119 – 124.

- Karat, J. (1997). User-Centered Software Evaluation Methodologies. In: M. Helander, T.K. Landauer & P. Prabhu (eds.): *Handbook of Human-Computer Interaction*. 2nd Completely Revised Edition. Amsterdam: Elsevier Science B.V., S. 689-704.
- Karat, C.-M., Campbell, R.L., Fliegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. *Proceedings ACM CHI 92 Conference. Monterey CA*. New York: ACM, S. 397 – 404.
- Lewis, C. (1982). *Using the "thinking aloud" method in cognitive interface design*. IBM Research Reports RC 9265 (40713), IBM Thomas J. Watson Research Center. Yorktown Heights: New York.
- Moll, Th. (1987). Über Methoden zur Analyse und Evaluation interaktiver Computersysteme. In K.-P. Fähnrich (Hrsg.): *Software-Ergonomie*. München: Oldenbourg, S. 179-190.
- Nielsen, J. (1993). *Usability Engineering*. Boston: AP Professional.
- Nielsen, J. (1994). Heuristic Evaluation. In: J. Nielsen & R.L. Mack (eds.): *Usability Inspection Methods*. New York: J. Wiley & Sons, S. 25-62.
- Nielsen, J. & Molich, R. (1990). Heuristic Evaluation of User Interfaces. In: J.C. Chew & J. Whiteside (eds.): *Chi'90 Conference Proceedings, Empowering People*. ACM Press, S. 249 – 256.
- Nielsen, J. & Mack, R. L. (eds.). (1994). *Usability Inspection Methods*. New York: John Wiley & Sons, Inc.
- Oppermann, R., & Reiterer, H. (1994). Software-ergonomische Evaluation. In: E. Eberle, H. Oberquelle & R. Oppermann (Hrsg.): *Einführung in die Software-Ergonomie*. Berlin: de Gruyter, S. 335-371.
- Rubin, J. (1994). *Handbook of Usability Testing*. New York: John Wiley & Sons, Inc.
- Smilowitz, E. D., Darnell, M.J. & Bensson, A.E. (1993). Are we overlooking some usability testing methods? A comparison of lab, beta, and forum tests. *Behaviour & Information Technology*, 13, NOS. 1 and 2, 183-190.
- Virzi, R. A. (1997). Usability Inspection Methods. In: M. Helander, T.K. Landauer & P. Prabhu (eds.): *Handbook of Human-Computer-Interaction*. 2nd completely revised edition. Amsterdam: Elsevier Science B.V., S. 705 – 715.
- Virzi, R.A., Sorce, J.F. & Herbert, L.B. (1993). A Comparison of Three Usability Evaluation Methods: Heuristic, Think Aloud, and Performance Testing. *Designing for diversity: Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting 1993*. Human Factors and Ergonomics Society: Santa Monica, S. 309-313.
- Wright, P. & Monk, A.F. (1989). Evaluation for Design. In: A. Scutcliffe & L. Macaulay (Hrsg.): *Proceedings of the HCI'89, Conference on People and Computers V*. Cambridge: University Press, S. 345-358.

Kontaktinformation

Kai-Christoph Hamborg
 Universität Osnabrück
 Fachbereich Humanwissenschaften, LE Psychologie, Arbeits- und Organisationspsychologie
 Seminarstr. 20
 49069 Osnabrück
 Email: khamborg@uos.de