A. M. Heinecke, H. Paul (Hrsg.): Mensch & Computer 2006: Mensch und Computer im Struktur*Wandel*.
München, Oldenbourg Verlag, 2006, S. 115-124

# Qualitätssicherung im Usability-Testing – zur Reliabilität eines Klassifikationssystems für Nutzungsprobleme

Kai-Christoph Hamborg, Tom Hoemske, Frank Ollermann

Universität Osnabrück, Fachbereich Humanwissenschaften

**Zusammenfassung**

Der Beitrag beschäftigt sich mit Maßnahmen zur Qualitätssicherung im Usability-Testing. Die Bedeutung der Klassifikation von Nutzungsproblemen für die Qualitätssicherung wird aufgezeigt und ein Klassifikationssystem dargestellt, das im Folgenden empirisch in Bezug auf seine Reliabilität überprüft wird. Hierzu wurden zwei Usability-Tests durchgeführt und die erhobenen Nutzungs- bzw. Usability-Probleme anschließend klassifiziert. Insgesamt erweist sich das Klassifikationssystem als reliabel, es wird jedoch nicht von allen Klassifizierungsmöglichkeiten Gebrauch gemacht. Stärken des Klassifikationsansatzes sowie Konsequenzen für dessen Weiterentwicklung werden diskutiert.

## 1 Einleitung

Seit einiger Zeit wird die Qualität von Usability-Tests kritisch diskutiert. Im Mittelpunkt der Diskussion stehen die Reliabilität und Validität dieser Evaluationsmethodik sowohl in der praktischen Anwendung als auch im Forschungskontext.

Probleme bei der Anwendung von Usability-Tests in der Praxis berichten Molich und Mitarbeiter (Molich et al. 2004). In einer ersten Studie wurden vier, in einer zweiten Studie neun Usability-Labore beauftragt, ein Softwareprodukt mittels eines Usability-Tests zu untersuchen. Neben zahlreichen Unterschieden in Bezug auf die Durchführung der Tests, die erzielten Ergebnisse und die Berichtlegung, war der wohl hervorstechendste Befund, dass in der ersten Studie nur ein einziges Usability-Problem aus einer Menge von 141, in der zweiten Studie kein einziges Problem aus einer Menge von insgesamt 310 Problemen von allen Beratungsfirmen übereinstimmend erkannt wurde. Kessner et al. (2001) kommen zu ähnlichen Ergebnissen. Sechs Usability-Teams führten unabhängig voneinander einen Usability-

Test mit derselben Software durch. Von insgesamt 36 identifizierten Usability-Problemen wurde keines von allen Teams erkannt, zwei wurden durch fünf Teams, vier durch vier Teams, sieben durch drei und weitere sieben durch zwei Teams erkannt.

Die Befunde deuten darufhin, dass Usability-Tests in der Praxis (und auch in der Forschung; s. Gray & Salzman 1998) offensichtlich sehr uneinheitlich durchgeführt werden. Die geringe Übereinstimmung bei der Identifikation von Usability-Problemen zeigt die unzureichende Reliabilität von Usability-Tests und stellt die Vertrauenswürdigkeit der Befunde in Frage. Aus diesem Grund scheinen Maßnahmen zur Qualitätssicherung für den Bereich des Usability-Testing und allgemein für die gestaltungsunterstützende Evaluation von Software geboten (Molich et al. 2004).

Einen Ansatz zur Qualitätssicherung im Usability-Testing stellen Andre und Mitarbeiter (2001) mit dem User Action Framework (UAF) vor. Bei dem UAF handelt es sich um ein Rahmengebäude zur Unterstützung des gesamten Usability-Engineering-Prozesses und hier insbesondere die Tätigkeiten: Interaktionsdesign, formative Evaluation, Problemklassifikation und -dokumentation sowie Berichtlegung.

Der im UAF vorgesehene Ansatz zur Problemklassifikation soll eine reliable Kategorisierung und Beschreibung von Usability-Problemen ermöglichen (Andre et al. 2001). Dies ist ein zentraler Punkt für die Qualitätssicherung von Usability-Tests, da die Reliabilität der Klassifikation von Usability-Problemen die Güte der Ergebnisse von Usability-Tests insgesamt beeinflusst. Bei der Durchführung von Usability-Tests fallen zum großen Teil qualitative Daten an, wie z.B. Verbalisierungen von Nutzern, wenn mit der Methode des Lauten Denkens gearbeitet wird. Die Auswertung dieser Daten erfordert ein inhaltsanalytisches Vorgehen, das zumeist, nach Aufbereitung der Rohdaten, die Kategorisierung der aufbereiteten Daten beinhaltet. Die Reliabilität der Kategorisierung wird durch die Übereinstimmung von wenigstens zwei Personen (Rater), die die Zuordnung der Daten zu Kategorien vornehmen, statistisch bestimmt. Das Ergebnis der Klassifizierung ermöglicht im Folgenden die Interpretation der erkannten Usability-Probleme und unterstützt damit die Kommunikation der Ergebnisse sowie die Ableitung von Maßnahmen.

Die im UAF verwendete Taxonomie zur Klassifizierung von Usability-Problemen basiert im Kern auf einem einfachen Handlungsmodell (Norman 1986). In dem Modell werden sieben Handlungsphasen unterschieden, die bei der Interaktion mit dem Computer durchlaufen werden: 1. Festlegen des *Handlungsziels* und 2. Bildung einer das Ziel konkretisierenden *Handlungsabsicht* sowie eines Handlungsplans, 3. *Spezifikation der Handlung*: Bestimmung einer Handlungsabfolge zur Umsetzung der Absicht bzw. des Plans, 4. *Ausführung der Handlung* in Form von Systemeingaben, 5. *Wahrnehmung des Systemzustands* nach Verarbeitung der Eingaben durch das System, 6. *Interpretation des Systemzustands*, 7. *Bewertung des Systemzustands* in Bezug auf das verfolgte Ziel und den Handlungsplan.

Für das UAF wurde dieses Handlungsmodell adaptiert und in den so genannten Interaktionszyklus (Interaction Cycle) mit den Phasen: *Planung*, *physische Handlungsausführung* und *Bewertung* übernommen (siehe Abbildung 1).

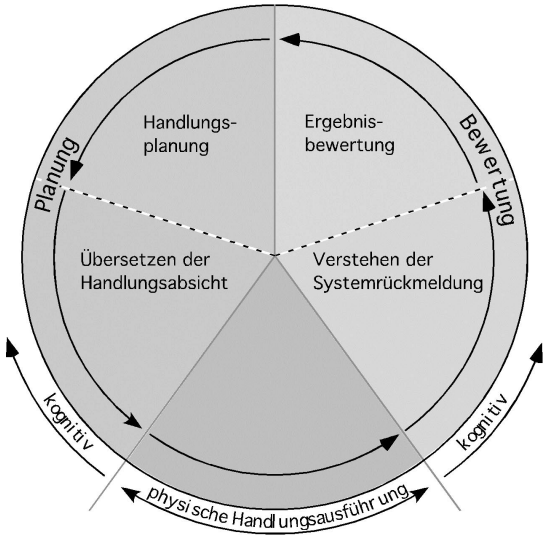


*Abbildung 1: UAF-Interaktionszyklus*

Die von Norman unterschiedenen Handlungsphasen lassen sich den im Interaktionszyklus unterschiedenen Phasen entsprechend Tabelle 1 zuordnen.

*Tabelle 1: Zuordnung der Phasen – UAF-Interaktionszyklus und Normans Handlungsmodell (Andre et al. 2001)*

| **Normans Handlungsmodell** | **Phase im Interaktionszyklus** | **Bewertungsaspekte** |
|---|---|---|
| Festlegung des Handlungsziels | Planung – hohe Ebene | können Nutzer die allgemeinen Anforderungen bestimmen, um mit einer Handlung beginnen zu können? |
| Bildung der Handlungsabsicht | Planung – hohe Ebene | können Nutzer festlegen was zu tun ist, um eine Handlung ausführen zu können? |
| Handlungsspezifikation | Planung – Übersetzung der Handlungsabsicht | können Nutzer festlegen, wie sie ihre Handlungsabsicht in Handlungen umsetzen? |
| Handlungsausführung | Physische Handlungsausführung | können Nutzer Handlungen einfach ausführen? |
| Wahrnehmung des Systemszustandes | Bewertung – Systemrückmeldung verstehen | können Nutzer die Systemrückmeldung wahrnehmen? |
| Interpretation des Systemzustandes | Bewertung – Systemrückmeldung verstehen | können Nutzer die Systemrückmeldung verstehen? |
| Evaluation des Systemszustandes mit Bezug auf Ziele und Intentionen | Bewertung – Bewertung des Ergebnisses | können Nutzer den Erfolg des Ergebnisses feststellen? |

Die Phasen: Planung, Handlungsausführung und Bewertung bilden Kategorien, denen die in einem Usability-Test erhobenen Probleme in einem ersten Schritt zugeordnet werden. Jeder der drei Kategorien sind über maximal drei Hierarchieebenen weitere Kategorien unterge-

ordnet. Die Unterkategorien konkretisieren die übergeordneten Kategorien schrittweise in Bezug auf handlungs- und gestaltungsbezogene Ursachen der Nutzungsprobleme. Die Unterkategorien variieren inhaltlich für die übergeordneten Kategorien. Das folgende Beispiel soll das Vorgehen erläutern. Ein Usability-Problem wurde mit Bezug auf den Interaktionszyklus der *Bewertungsphase* (Der Nutzer hat die Ergebnisse seiner Interaktion mit dem System nicht erkannt) zugeordnet (Ebene 1). Auf der zweiten Ebene stehen drei Kategorisierungsalternativen, von denen das Problem der Kategorie 1.1 *Die Rückmeldung des Systems über den Systemstatus ist nicht ausreichend* zugewiesen wird. Auf der dritten Ebene sind wiederum vier Kategorisierungsalternativen vorhanden, von denen das Problem der zweiten Kategorie (*Die Darstellung des Systemfeedbacks ist unzureichend*) zugeordnet wird. Auf der folgenden vierten Ebene werden wiederum acht weitere Kategorisierungsalternativen angeboten. Das Problem wird wie folgt kategorisiert: *Sensorische Wahrnehmbarkeit der Rückmeldungsdarstellung ist beeinträchtigt* da der Text zu klein dargestellt wurde.

Diese schrittweise konkretisierende Kategorisierung soll zu einer möglichst eindeutigen Eingrenzung der Ursachen von Nutzungsproblemen führen und damit nachfolgende Designentscheidungen unterstützten. Die Klassifikationspfade sind im wesentlichen hierarchisch organisiert, an einigen Stellen ist jedoch ein Wechsel in den Klassifikationspfaden möglich; unterschiedliche Pfade können so u.U. zur gleichen Endkategorie führen. Die Klassifikation wird durch ein Softwarewerkzeug, den UAF-Viewer unterstützt, durch den die Kategorien als Hyperdokument dargeboten und erläutert werden (siehe Abbildung 3).

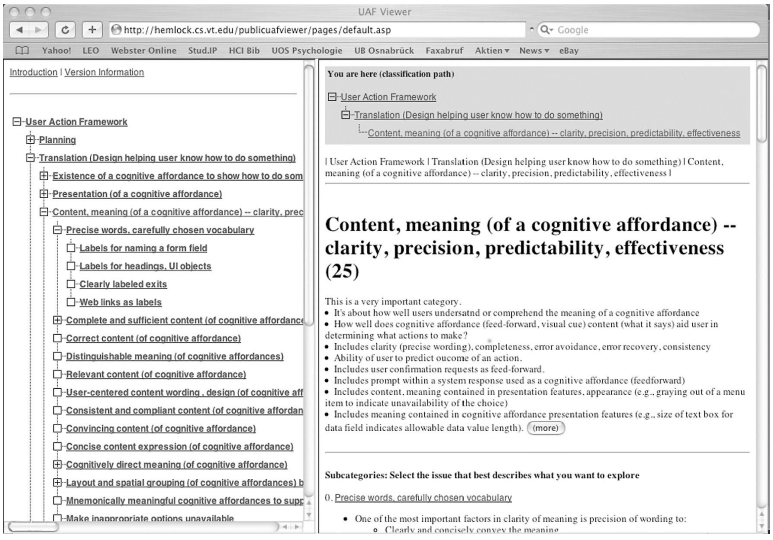


*Abbildung 2: UAF-Viewer (http://hemlock.cs.vt.edu/uaf/)*

Die Anwendung dieses Kategoriensystems verspricht eine deutliche Verbesserung der Reliabilität der Klassifikation von Usability-Problemen. Der empirische Nachweis erfolgte jedoch bisher lediglich auf der Basis einer selektiven Stichprobe eindeutig klassifizierter und häufig aufgetretener Nutzungsprobleme (Andre et al. 2001).

Gegenstand des vorliegenden Beitrags ist die Überprüfung der Reliabilität der Taxonomie an Hand empirisch erhobener Usability-Probleme.

# 2 Untersuchung

Für die Untersuchung wurde ein Online-Kursmanagementsystem in zwei Usability-Tests mit verschiedenen Test-Methoden einer formativen Evaluation unterzogen. Die identifizierten Usability-Probleme wurden mit dem UAF-Kategorisierungssystem klassifiziert, um anschließend dessen Reliabilität zu bestimmen.

## 2.1 Stichprobe

Insgesamt nahmen an der Untersuchung 28 Studierende teil. Studierende zählen zu den Hauptnutzern des untersuchten Systems. Der Altersdurchschnitt der Teilnehmer betrug 24,93 Jahre (SD 6,84). Von den Probanden waren 26 weiblich und zwei männlich. Für die Untersuchung wurden nur Probanden mit geringer bis mittlerer Expertise bezüglich des untersuchten Systems ausgewählt, da erstens die Nutzung des Programms für diese Nutzergruppe erleichtert werden sollte und zweitens, da die im Usability-Test verwendete Methode des Lauten Denkens bei Nutzern mit hoher Expertise auf Grund der eingeschränkten Verbalisierbarkeit der hochautomatisierten Handlungsabläufen zugrundeliegenden psychischen Prozesse nur bedingt funktionsfähig ist.

## 2.2 Methode

### 2.2.1 Software-System

Gegenstand der Usability-Tests war ein Kursmanagementsystem, das Lehrende und Studierende bei der Organisation von Lehrveranstaltungen sowie die Kommunikation zwischen Studierenden untereinander und zwischen Lehrenden und Studierenden unterstützt. Für die Evaluation des Systems wurde ein Nutzerkonto auf einem Testserver eingerichtet, damit die Untersuchung nicht durch Änderungen des Produktivsystems beeinträchtigt wurde.

### 2.2.2 Methoden

Die Usability-Tests wurden mit zwei unterschiedlichen Methoden durchgeführt, der Methode des Lauten Denkens und der Videokonfrontation.

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. In der vorliegenden Untersuchung orientierte sich die Anwendung der Methode des Lauten Denkens an den Empfehlungen von Ericsson und Simon (1980) sowie Boren und Ramey (2000). Angeleitet durch eine kurze schriftliche Instruktion und nach einer Übungsphase mit der

Methode äußerten die Probanden (Pbn) während der Bearbeitung von Testaufgaben kontinuierlich ihre Bewusstseinsinhalte. Der Versuchsleiter saß dabei in ca. 1,5 Metern Abstand im äußeren Sichtfeld der Pbn um ggf. nach Zielerreichung bzw. Zeitablauf zur nächsten Aufgabe überleiten zu können, Fragen zur Durchführung zu beantworten und oder auf die Instruktion zum Lauten Denken hinzuweisen. Die Interaktion der Probanden mit dem System wurde mittels Videotechnik aufgenommen.

Die *Technik der Videokonfrontation* (Hamborg & Greif 1999; Moll 1987) sieht vor, dass zunächst eine Arbeitssequenz mit dem zu evaluierenden System per Videotechnik aufgezeichnet und daraufhin in einem Interview mit dem Pbn analysiert wird. Hierbei kommen standardisierte und halbstandardisierte Frageformate zum Einsatz. In der vorliegenden Untersuchung wurde ein Interviewleitfaden zur Identifikation kritischer Nutzungsereignisse eingesetzt, der sich am UAF-Interaktionszyklus orientierte. Als kritische Ereignisse wurden hierbei alle Unterbrechungen einer Handlungssequenz sowie längere Planungs- und Explorationsphasen gewertet. Zur Identifikation der kritischen Ereignisse wurden die Pbn zu jeder Testaufgabe nach ihrem Bearbeitungsziel gefragt. Im anschließenden Schritt wurden mit den Pbn die Aufgabenschritte erhoben, die zur Erreichung des jeweiligen Aufgabenziels durchgeführt worden waren. Passagen, in denen kritische Ereignisse auftraten, wurden bei Bedarf mehrfach und ggf. verlangsamt betrachtet, um die Erinnerungsprozesse der Pbn zu unterstützen. Für jeden Aufgabenschritt wurden Fragen zu den einzelnen Phasen des Interaktionszyklus und damit verbundenen Probleme gestellt (z.B. „Waren die Rückmeldungen aus Ihrer Sicht ausreichend und verständlich“). Auf diese Weise wurden die Ziele, Arbeitsschritte und dabei auftretenden Nutzungsprobleme in allen Arbeitsaufgaben identifiziert.

### 2.2.3 Testaufgaben und Versuchsablauf

Die Untersuchungsteilnehmer bearbeiteten drei unterschiedliche Aufgaben: Herunterladen einer Datei, Nutzung des Forums einer Veranstaltung sowie Heraufladen einer Präsentation in den Dateiordner einer Veranstaltung. Die Testaufgaben wurden den Pbn in randomisierter Reihenfolge dargeboten und jeweils schriftlich instruiert. Die Aufgaben waren innerhalb eines Szenarios miteinander verbunden, das einer typischen Nutzung des Systems entsprach.

Zum Untersuchungsbeginn wurden die Pbn gebeten, sich die schriftlichen Instruktionen genau durchzulesen und ggf. Fragen zu den Aufgaben und dem Szenario zu stellen. Weiterhin wurden sie instruiert, die Aufgaben vor der Bearbeitung genau durchzulesen und jeweils eine Aufgabe komplett vor Beginn einer neuen Aufgabe zu bearbeiten.

# 3 Ergebnisse

Während in der Literatur als Resultat formativer Evaluation von identifizierten Problempunkten („problems“, Nielsen 1993, 156) gesprochen wird, ist die genaue Bewertung, was ein Usability-Problem genau darstellt, nicht so einfach, wie häufig suggeriert. Wie bereits einleitend angesprochen, wurden bei den in dieser Untersuchung eingesetzten Methoden zunächst Verbalisierungen zu dem evaluierten System erhoben. Bei der Auswertung dieser Verbalisierungen handelt es sich um ein *inhaltsanalytisches* Problem. Verbalisierungen mit

Problemgehalt müssen als solche erkannt und von Anmerkungen mit anderem Inhalt getrennt werden, um daraufhin der Kategorisierung durch wenigstens zwei Rater unterzogen werden zu können. Die Reliabilität des Klassifikationssystem lässt sich mit Hilfe von Übereinstimmungskoeffizienten bestimmen.

Die Daten aus beiden Usability-Tests wurden in vier aufeinander aufbauenden Schritten analysiert. In einem ersten Schritt wurden die Verbalisierungen aus beiden Untersuchungen transkribiert. Nach der Transkription wurde die Daten segmentiert und in Bezug auf ihre Verwertbarkeit vorkategorisiert. Als nicht verwertbar wurden solche Aussagen der Probanden klassifiziert, die sich auf Personen oder Situationen, nicht aber auf das Programm bezogen. Nach der Vorauswahl verwertbarer Aussagen wurden diese expliziert. Das Ziel der Explikation besteht darin, Verbalisierungen der Pbn ggf. so zu ergänzen, dass sie durch Dritte ohne zusätzliche Informationen verständlich sind. In einem letzten Schritt wurden innerhalb der Menge aller Aussagen diejenigen identifiziert, die Usability-Probleme adressierten. Ein Usability-Problem wurden als Störung des Planungs- oder Handlungsflusses definiert. Die Operationalisierung wurde wie folgt vorgenommen: Ein Usability-Problem liegt dann vor, wenn der Nutzer oder die Nutzerin:

- Schwierigkeiten hat festzulegen, was zu tun ist um mit der Aufgabe beginnen zu können
- Schwierigkeiten hat, über eine geeignete Strategie bzw. einzusetzende Operationen/ Handlungsschritte für das weitere Vorgehen zu entscheiden
- nicht genau festlegen kann, durch welche Handlungsschritte Ziele umgesetzt werden können
- Schwierigkeiten hat, Handlungen zur Zielerreichung auszuführen
- das Feedback auf Handlungen nicht wahrnehmen oder verstehen kann
- den Erfolg des eigenen Handelns nicht in vollem Umfang feststellen kann.

Die Identifikation der Usability-Probleme und deren Zuordnung zu dem Kategoriensystem wurde von drei unabhängigen Ratern durchgeführt. Redundante Probleme wurden zusammengefasst.

Die Kategorisierung erfolgte mit Hilfe einer in die deutsche Sprache übersetzten und um Beispiele ergänzten Fassung des UAF-Viewers (Version 3.3, November 2004). Die Zuordnung der Usability-Probleme erfolgte zunächst zu einer der Handlungsphasen und setzte sich dann schrittweise über die hierarchisch untergeordneten Ebenen des Kategoriensystems fort.

Andre et al. (2001) schlagen vor, die Übereinstimmung der Ratings immer jeweils für eine Ebene zu berechnen, weil nicht in allen Kategorien eine Zuordnung bis zur untersten Ebene des UAFs möglich ist. Diesem Vorgehen wurde in der vorliegenden Studie gefolgt. Nach der Zuweisung eines Usability-Problems zu einer Kategorie auf der ersten Ebene des Kategoriensystems (z.B. „Wie beginne ich mit der Aufgabe?“ ⇨ Kategorie Planung), musste der Rater entscheiden, welcher Kategorie auf der zweiten Ebene das Problem zuzuordnen war (z.B. Probleme bei der Bildung von Teilzielen für die Bearbeitung einer Aufgabe ⇨ Kategorie Zielzerlegung) usw.. In der Untersuchung wurden die Rater dazu angehalten, bei der Kategorisierung möglichst die gesamte Hierarchie zu durchlaufen, sie wurden jedoch nicht forciert, die Nutzungsprobleme jeweils der untersten Ebene des Kategoriensystems zuzu-

weisen, da in einem vorangegangenen Ratertraining deutlich wurde, dass ein Teil der vorliegenden Problembeschreibungen auf Grund fehlender Details der Problembeschreibung dies nicht erlaubten. Von den 128 Usability-Problemen, die der ersten Kategorienebene zugewiesen wurden, ließen sich 96% auch den Kategorien der zweiten Ebene, 40% auf der dritten Kategorienebene aber nur noch 3% auf der vierten Ebene zu ordnen.

Die Bestimmung der Raterübereinstimung wurde mit Hilfe des Kappa-Koeffizienten (κ, Cohen 1960) vorgenommen. Sie erfolgte getrennt für die Ratings der unterschiedenen Kategorienebenen. Hierbei wurden nur die Probleme einbezogen, für die von allen drei Ratern eine Zuordnung vorlag. Die berechnete Raterübereinstimmung erweist sich auf den Ebenen 1-3 als gut (siehe Tabelle 2).

*Tabelle 2: Raterübereinstimmung für die verschiedenen UAF Ebenen (Kappa)*

| | **Raterübereinstimmung (κ)** | **Anzahl der Probleme** |
|---|---|---|
| Problemrating | 0.77 | |
| UAF Ebene 1 | 0.80 | 128 (100,00%) |
| UAF Ebene 2 | 0.74 | 123 (96,09%) |
| UAF Ebene 3 | 0.72 | 52 (40,63%) |
| UAF Ebene 4 | - | 4 (3,12%) |

# 4 Diskussion

Nach den Befunden der vorliegenden Untersuchung erlaubt der im UAF vorgesehene Kategorisierungsansatz eine reliable und theoriegeleitete Bestimmung und Unterscheidung von Nutzungsproblemen. Es hat sich jedoch gezeigt, dass nur knapp die Hälfte aller Usability-Probleme auf der dritten und ein marginale Anzahl auf der nochmals konkreteren, vierten Kategorienebene verortbar war. Die Ursache hierfür lag hauptsächlich in der nicht ausreichend konkreten Beschreibung der Usability-Probleme. Dass sich Usability-Probleme in Bezug auf ihre Konkretheit unterscheiden, ist ein bekanntes Phänomen, wobei die Bewertung genereller und spezifischer Problembeschreibungen in der Literatur unterschiedlich ausfällt (s. Dumas & Redish 1999; Gediga & Hamborg 1997). Im Rahmen des UAF bedeutet fehlende Konkretheit jedoch die eingeschränkte Rückführbarkeit der Usability-Probleme auf ihre Ursachen. Ein Ansatzpunkt, um mit diesem Problem umzugehen, bietet die weitere Auseinandersetzung mit Methoden des Usability-Testings und die Qualität der durch sie erfassten Usability-Probleme. Nach unseren Erfahrungen werden z.B. durch die Methode der Videokonfrontation konkrete Problemnennungen erhoben als durch die Methode des Lauten Denkens. Eine weitere Möglichkeit, den Anteil konkreter Problembeschreibungen zu erhöhen, besteht in einer sorgfältigeren Explikation der erhobenen Daten. Werden im Rahmen der inhaltsanalytischen Auswertung die Quellen zur Beschreibung der Probleme im Detail berücksichtigt, sollte dies zu einem größeren Anteil konkreterer Problembeschreibungen führen. Beide Fragen sollten in folgenden Untersuchungen geklärt werden.

Gerade bei einer großen Anzahl erkannter Nutzungsprobleme kann zusätzlich zu der Unterscheidung von Problemqualitäten die Bedeutsamkeit (severity) der erkannten Nutzungsprobleme eine wichtige Information enthalten, um Gestaltungsentscheidungen zu unterstützen.

Eine entsprechende Entscheidungsgrundlage wird durch das UAF bisher nicht geboten. Auch in dieser Hinsicht gibt es noch Entwicklungsbedarf der in diesem Beitrag vorgestellten Taxonomie. Als Kriterien zur Priorisierung sind die Auftretenshäufigkeit einzelner Usability-Probleme, die subjektive Einschätzungen des Schweregrads durch die Nutzer oder die mit den einzelnen Usability-Problemen verbundenen ökonomischen oder psychischen Kosten denkbar (Hassenzahl 2000; Hassenzahl et al. 1997).

Eine Stärke der UAF-Klassifikation im Vergleich zu anderen theoriegeleiteten Ansätzen der Problemkategorisierung, wie etwa Fehlertaxonomien (s. z.B. Frese & Zapf 1992), besteht darin, dass auch Nutzungsprobleme berücksichtigt werden, die nicht zu Fehlern führen. Für die Interaktion spielen diese Probleme aber durchaus eine Rolle, da sie zu Zusatzaufwand bei der Systembedienung führen können. Entsprechend sollte die hier vorgestellte Taxonomie ein breiteres Problemspektrum erkennen helfen.

Schließlich muss aber auch darauf hingewiesen werden, dass die Güte der Klassifizierung von Usability-Problemen nur einen, wenn auch recht bedeutenden, Beitrag zur Qualitätssicherung im Bereich des Usability-Testing leisten kann. Das in diesem Beitrag eingesetzte Verfahren scheint hierzu beizutragen. Weitere Maßnahmen, die sich auf die Planung und Durchführung von Usability-Tests sowie auf Berichtlegung und die Kommunikation der Befunde richten, sind jedoch darüber hinaus notwendig, um die Qualität von Usability-Tests zu verbessern (Molich et al. 2004).

## Literaturverzeichnis

Andre, T. S.; Hartson, H. R.; Belz, S. M.; McCreary, F. A. (2001): The user action framework: a reliable foundation for usability engineering support tools. International Journal of Human-Computer Studies, 54, S. 107-136.

Boren, T.; Ramey, J. (2000): Thinking Aloud. Reconciling Theory and Practice. IEEE Transactions on professional communication, 43(3), S. 261-278.

Cohen, J. (1960): A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, S. 37-46.

Dumas, J. S.; Redish, J. C. (1999): A Practical Guide to Usability Testing (Revised Edition). Exter: Intellect Books.

Ericsson, K. A.; Simon, H. A. (1980): Verbal reports as data. Psychological Review, 87, S. 215-251.

Frese, M.; Zapf, D. (Hrsg.). (1992): Fehler bei der Arbeit mit dem Computer – Ergebnisse von Beobachtungen und Befragungen im Bürobereich. Bern: Huber.

Gediga, G.; Hamborg, K.-C. (1997): 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. Stuttgart: Teubner. S. 145-155.

Gray, W. D.; Salzman, M. C. (1998): Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. Human-Computer Interaction, 13, S. 203-261.

Hamborg, K.-C.; Greif, S. (1999): Heterarchische Aufgabenanalyse. In: H. Dunckel (Hrsg.), Handbuch psychologischer Arbeitsanalyseverfahren. Zürich: vdf. S. 147-177.

Hassenzahl, M. (2000): Prioritizing usability problems: data-driven and judgement-driven severity estimates. Behaviour & Information Technology, 19(1), S. 29-42.

Hassenzahl, M.; Prümper, J.; Sailer, U. (1997): Die Priorisierung von Problemhinweisen in der software-ergonomischen Qualitätssicherung. In: R. Liskowsky; B. M. Velichkovsky; W. Wünschmann (Hrsg.), Software-Ergonomie '97. Usability Engineering: Integration von Mensch-Computer-Interaktion und Software-Entwicklung. Stuttgart: Teubner. S. 191-201.

Kessner, M.; Wood, J.; Dillon, R. F.; West, R. L. (2001). On the reliability of usability testing, CHI 2001 Extended Abstracts. New York, NY: ACM. S. 97-98.

Molich, R.; Ede, M.; Kaasgaards, K.; Karyukin, B. (2004): Comparative usability evaluation. Behaviour & Information Technology, 23(1), S. 65-74.

Moll, T. (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.

Norman, D. A. (1986): Cognitive Engineering. In: D. A. Norman; S. W. Draper (Hrsg.), User Centered System Design. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers. S. 31-61.

**Kontaktinformationen**

Universität Osnabrück
Fachbereich Humanwissenschaften
Lehreinheit Psychologie
Arbeits- und Organisationspsychologie
Kai-Christoph Hamborg, Tom Hoemske, Frank Ollermannn

Seminarstr.20
D-49069 Osnabrück Germany
khamborg@uni-osnabrueck.de