Welche software-ergonomischen Evaluationsverfahren können *was* leisten ?

Bernd Holz auf der Heide
Technische Universität München

Zusammenfassung

Die Benutzungsfreundlichkeit kommerzieller Dialogsysteme kann mit Hilfe einer systematischen Evaluation gezielt verbessert werden. Welche Verfahren sind hierzu geeignet? Um diese Frage zu beantworten, wurden folgende Evaluationsansätze theoriegeleitet (weiter)entwickelt und in einer Längsschnittstudie empirisch miteinander verglichen: Fehler- und Zeitanalysen anhand von Rechnerprotokollen, Videoaufzeichnungen und Beobachtungsprotokollen; mündliche und schriftliche Benutzerbefragungen sowie Expertenurteile mittels software-ergonomischer Checklisten. Anhand der vorliegenden Ergebnisse werden die Vor- und Nachteile sowie die in der Praxis sinnvollen Einsatzbereiche und -bedingungen der einzelnen Evaluationsverfahren erläutert.

1 Mehr Benutzungsfreundlichkeit

Der Benutzungsfreundlichkeit bzw. Benutzbarkeit von Dialogsystemen kommt eine zentrale Bedeutung zu. Dennoch werden in der betrieblichen Praxis immer noch viele Benutzer mit Systemen konfrontiert, die ihren Ansprüchen, Schwierigkeiten und Bedürfnissen nicht gerecht werden. Um diese Situation zu verbessern und die Benutzbarkeit der Software zu erhöhen, werden von Forschern, Entwicklern und Anwendern mehrere unterschiedliche Wege eingeschlagen. Die folgenden Beispiele skizzieren drei dieser Wege und die dabei auftretenden Hindernisse.

... durch gezielten Einkauf

Immer häufiger werden Programme nicht mehr speziell entwickelt, sondern als "Standardsoftware" eingekauft und ggf. organisations- und anwendungsspezifisch modifiziert. Da sich die auf dem Markt angebotenen Systeme hinsichtlich ihrer technischen und betriebswirtschaftlichen Aspekte oft nur wenig unterscheiden, rücken bei der Kaufentscheidung software-ergonomische Qualitätsmerkmale und Fragen der Zufriedenheit und Akzeptanz der Endnutzer in den Vordergrund (Piepenburg & Rödiger [15]). Um sich nun bei einer Auswahlentscheidung nicht ausschließlich auf die Angaben der Hersteller und den persönlichen Eindruck verlassen zu müssen, benötigt man praxisgerechte Bewertungsverfahren, die einen direkten Vergleich zwischen alternativen Systemen ermöglichen.

... durch Normen und Richtlinien

Ein weiterer Weg, mehr Benutzungsfreundlichkeit zu erreichen, ist die Vorgabe software-ergonomischer Normen und Richtlinien, die bestimmte Eigenschaften eines Dialogsystems fordern. Hier gibt es mittlerweile eine ganze Reihe firmeninterner, firmenübergreifender, nationaler und internationaler Normen, Gestaltungsrichtlinien und -vorschriften (z.B. Apple [1], DIN 66234 Teil 8 [3], IBM [10], ISO 9241 Part 10 [11], Smith & Mosier [19]).

Um die Einhaltung dieser Forderungen zu kontrollieren und ihnen damit den entsprechenden Nachdruck zu verleihen, werden geeignete Verfahren zur sog. "Prüfung auf Normenkonformität" benötigt. Die hier bis heute vorherrschende Situation wird durch die Verfasser der DIN 66234 Teil 8, die fünf "Grundsätze ergonomischer Dialoggestaltung" beschreibt, treffend charakterisiert: "Es ist derzeit noch nicht möglich, die Erfüllung einzelner ... Leitsätze objektiv zu überprüfen, da geeignete Überprüfungsverfahren noch nicht bekannt sind" ([3], S. 1).

... durch neue Entwicklungsstrategien

Ein meines Erachtens sehr erfolgversprechender Ansatz zur Verbesserung der software-ergonomischen Qualität von Dialogsystemen bezieht sich auf den Entwicklungsprozeß selbst (Floyd [4]). Hier wird weniger versucht zu klären, *was* Benutzungsfreundlichkeit ist, sondern vielmehr, *wie* man benutzungsfreundliche Systeme entwickeln kann.

Neben den Prinzipien *Benutzerbeteiligung* und *Prototyping* ist die *systematische empirische Bewertung* der zu entwickelnden Systeme eine ganz zentrale Maßnahme (Gould & Lewis [6]). D.h. die Benutzungsfreundlichkeit der erstellten Software-Prototypen wird frühzeitig und gezielt durch Testläufe mit tatsächlichen oder potentiellen Benutzern anhand typischer Arbeitsaufgaben evaluiert. Hierzu werden geeignete Verfahren benötigt. Die Prototypen werden auf der Basis der Evaluationsergebnisse überarbeitet und erneut getestet. Dieser Zyklus aus (Re)Design und Evaluation wird solange durchlaufen, bis die Tests befriedigende Ergebnisse ergeben.

Wie aus diesen drei Beispielen deutlich wird, besteht ein großer Bedarf an unterschiedlichen Verfahren zur Bewertung bzw. Evaluation der software-ergonomischen Qualität von Prototypen und fertigen Dialogsystemen. Dementsprechend wurden und werden für unterschiedliche Evaluationsanlässe zum Teil sehr spezifische Verfahren konzipiert. So finden sich in der Literatur etliche Zusammenstellungen von Vorgehensweisen und Verfahren, die zur Evaluation von Dialogsystemen genutzt werden können (z.B. Hampe-Neteler & Rödiger [8], Rauterberg [17]). Bisher wurde allerdings kein Versuch unternommen, die in der software-ergonomischen Forschung und Praxis häufig verwendeten Evaluationsverfahren *empirisch* hinsicht-

lich ihrer Möglichkeiten und Grenzen miteinander zu vergleichen und auf dieser Grundlage die jeweils sinnvollen Einsatzbereiche und -bedingungen aufzuzeigen.

Dieser Versuch wurde im Kontext des Forschungsprojekts PROTOS (Methoden zur Entwicklung und Bewertung von Prototypen für Dialogsysteme - BMFT 01 HK 088-6) unternommen: Zunächst wurden zentrale, in der software-ergonomischen Forschung und Praxis häufig eingesetzte Evaluationsansätze zusammengestellt und hinsichtlich ihrer zugrundeliegenden Evaluationsziele, -kriterien und -mittel systematisiert. Die praxisrelevanten Ansätze wurden im Rahmen von *Querschnittstudien* mit Systemen unterschiedlicher Funktionalität und Komplexität erprobt und zu anwendungsreifen Verfahren weiterentwickelt. Anschließend wurden diese Verfahren in einer *Längsschnittstudie* - in deren Verlauf eine Büroanwendung iterativ entwickelt wurde - parallel eingesetzt, so daß ein direkter Vergleich möglich war.

Welche Bewertungsverfahren können *was* leisten? Dieser Frage wird im folgenden nachgegangen, indem zunächst Anforderungen an erfolgversprechende Evaluationsverfahren formuliert werden. Daran anknüpfend werden die in der Längsschnittstudie eingesetzten Verfahren vorgestellt und die dabei gewonnenen Erfahrungen und Ergebnisse diskutiert.

2 Anforderungen an geeignete Evaluationsverfahren

Alle software-ergonomische Evaluationsverfahren sollten zunächst den üblichen testtheoretischen Anforderungen genügen, d.h. sie sollten möglichst objektiv, reliabel und valide sein. Um wirklich praxisrelevant zu sein, sollte die Erhebung der Daten und deren Auswertung nur einen relativ geringen finanziellen, zeitlichen und technischen Aufwand erfordern. Zudem sollten die Verfahren universell einsetzbar sein, um für Dialogsysteme unterschiedlicher Funktionalität und Komplexität anwendbar zu sein. Soweit zu den allgemeinen Anforderungen, die prinzipiell für alle Evaluationsverfahren gelten.

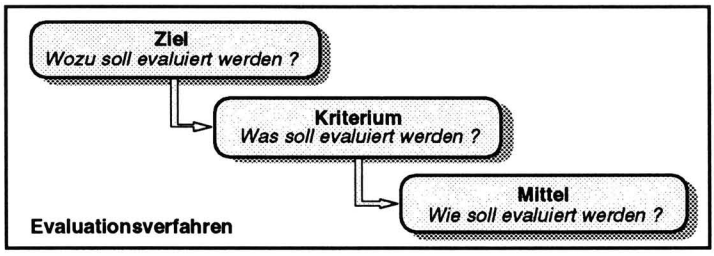

Abb. 1: Die drei Elemente eines Evaluationsverfahrens

Um die spezifischen Anforderungen zu bestimmen, muß der Vorgang "Evaluation eines Dialogsystems" etwas genauer betrachtet werden. Hierzu bietet sich eine Differenzierung nach den *Zielen*, *Kriterien* und *Mitteln* einer Evaluation an (Piepenburg & Rödiger [15]). Jedes Evaluationsverfahren läßt sich durch diese drei Elemente charakterisieren (Abb. 1).

Das Ziel der Evaluation bestimmt dabei das weitere Vorgehen, d.h. die Wahl angemessener Evaluationskriterien und deren Erhebung durch geeignete Evaluationsmittel. Die spezifischen Anforderungen an das einzusetzende Verfahren ergeben sich somit aus dem jeweiligen Evaluationsziel. Unter praxisorientierten Gesichtspunkten lassen sich hier in Anlehnung an Piepenburg und Rödiger [15] drei zentrale Evaluationsziele bzw. -anforderungen unterscheiden, die das Spektrum von der *formativen* Evaluation (bei der es um die Bewertung und Gestaltung einzelner Systemaspekte geht) bis zur *summativen* Evaluation (bei der das Dialogsystem in seiner Gesamtwirkung im Vordergrund steht) abdecken. Sie können durch drei Fragen charakterisiert werden, die eine entsprechende Evaluation beantworten soll:

- Die Evaluation soll die Frage "Which is better?" beantworten, d.h. einen direkten, globalen Vergleich zwischen alternativen Dialogsystemen ermöglichen; beispielsweise um für einen bestimmten Anwendungszweck das software-ergonomisch beste Produkt auszuwählen oder um bei einer Prototyping-Entwicklung zu klären, ob einzelne Modifikationen insgesamt zu einer Verbesserung geführt haben.
- Die Evaluation soll die Frage "How good?" beantworten, d.h. die Ausprägung bestimmter gewünschter bzw. geforderter Systemeigenschaften beurteilen. Eine entsprechende Evaluation kann in fortgeschritteneren Stadien einer Systementwicklung oder auch bei einem bereits fertiggestellten Dialogsystem sinnvoll sein, um einen Überblick über dessen software-ergonomische Qualität zu erhalten.
- Die Evaluation soll Antworten auf die Frage "Why bad?", d.h. Hinweise auf Schwachstellen sowie direkte Gestaltungsvorschläge liefern. Der typische Anwendungsbereich liegt in der Entwicklung bzw. Weiterentwicklung von Systemen. Ihren optimalen Wirkungsgrad kann eine entsprechende Evaluation dann entfalten, wenn sie systematisch in den Entwicklungsprozeß integriert ist und somit frühzeitig und gezielt ansetzt. Dies ist insbesondere beim Prototyping in einem Designteam der Fall (vgl. Holz auf der Heide & Hacker [9]).

3 Kriterien und Mittel

Im folgenden werden diejenigen Kombinationen von Evaluationskriterien und -mitteln skizziert, die wir im Hinblick auf die oben genannten Anforderungen untersucht haben und auf die sich die hier vorgestellten Ergebnisse beziehen. Die Auswahl basiert auf einer umfangreichen Literaturanalyse und der Befragung von

Entwicklern und Entwicklungsleitern mittelständischer Software-Häuser und spiegelt insofern *typische* Vorgehensweisen wider.

Alternative Systeme vergleichen : Für einen globalen Vergleich zwischen Prototypen bzw. fertigen Systemen unter dem Gesichtspunkt "Which is better?" bieten sich demnach insbesondere zwei Arten von Kriterien an:

Zum Einen ist dies die subjektive Zufriedenheit der Benutzer mit dem System bzw. deren Akzeptanz (Helmreich [7]). Diese kann per standardisiertem Fragebogen erhoben werden.

Zum Anderen sind dies objektive Leistungsmaße - insbesondere die Fehler und Zeiten der Benutzer bei der Bearbeitung typischer Arbeitsaufgaben, durch die gewissermaßen die Gesamteignung eines Systems für bestimmte Arbeitsaufgaben und Benutzergruppen bewertet wird. Als Mittel zur Erhebung dieser Leistungsmaße können direkte Beobachtungen mit Protokollierung, Videoaufzeichnungen sowie Rechnerprotokolle benutzt werden (vgl. Abschnitt 4).

Systemeigenschaften bewerten : Kriterien für die Beantwortung der Frage "How good?" sind in erster Linie die gewünschten bzw. geforderten Systemeigenschaften selbst. Bei relativ abstrakten Forderungen - wie z.B. nach "Steuerbarkeit" (DIN 66234, 8 [3]) - besteht hier allerdings das Problem der Operationalisierung und der Übertragbarkeit auf unterschiedliche Benutzergruppen und Aufgabenbereiche. In der Praxis erfolgt die Erhebung üblicherweise durch entsprechende Experten, die das System hinsichtlich der vorliegenden Kriterien einstufen. Für diesen Zweck gibt es einige (halb-)standardisierte Checklisten und Leitfäden, so z.B. das sehr ausführliche "EVADIS II-Verfahren" (Oppermann, Murchner, Reiterer, Koch [13]).

Gewissermaßen eine Variante des üblichen Expertenurteils ist die standardisierte Befragung der Benutzer, bei der diese per Rating *ihr* Expertenurteil zu bestimmten Systemeigenschaften abgeben (Norman & Shneiderman [12], Prümper & Anft [16], Shneiderman [20], Spinas [21]). (Auf die Frage, inwieweit die Beurteilungen durch Experten und Benutzer übereinstimmen, werde ich in Abschnitt 5.3 eingehen.)

Eine weitere Möglichkeit ist die *interaktionsbezogene* Analyse von Benutzerfehlern, anhand derer Rückschlüsse auf bestimmte Systemeigenschaften gezogen werden (z.B. Arnold & Roe [2] Frese & Zapf [5], Rouse & Rouse [18]).

Systemmängel erfassen : Kriterien für die Beantwortung der Frage "Why bad?" sind in erster Linie die systembezogenen Problemhinweise, Kritikpunkte und Gestaltungsvorschläge durch repräsentative Benutzer, die ihre Anforderungen und Bedürfnisse als Experten für die Arbeitsaufgabe im tagtäglichen Umgang mit dem System vertreten. Für die Erhebung dieser eher qualitativen Daten bietet sich ein prozeß-

bezogenes Vorgehen an. Dieses muß auf den konkreten Einzelfall abgestimmt werden und hängt insbesondere davon ab, ob Software-Prototypen oder bereits in einem Betrieb eingeführte Systeme evaluiert werden sollen.

Bei der Evaluation von *Software-Prototypen* müssen die (potentiellen) Benutzer zunächst in den Umgang mit dem System eingeführt werden. Idealerweise geschieht dies in Verbindung mit einer Systemexploration durch die Benutzer und einer entsprechenden Befragung. Daran anknüpfend können den Benutzern typische Arbeitsaufgaben zur Bearbeitung vorgelegt werden. In einer Nachbesprechung können kritische Stellen im Bearbeitungsablauf mit den Benutzern durchgegangen und deren Ursachen herausgearbeitet werden.

Bei bereits fertigen und in einem Betrieb eingeführten Systemen kann eine schriftliche Benutzerbefragung erste Hinweise liefern. Daran anknüpfend können Beobachtungen am Arbeitsplatz sowie halbstandardisierte Einzelinterviews und/oder Gruppendiskussionen durchgeführt werden, um erkannte Problembereiche und erhaltene Gestaltungshinweise zu konkretisieren.

Neben diesem prozeßbezogenen Vorgehen kann die *systembezogene* Analyse von Benutzerfehlern gewisse Hinweise auf Problembereiche erbringen. Abb. 2 zeigt die hier skizzierten Evaluationsansätze im Überblick.

Ziele	Kriterien	Mittel
Systeme vergleichen "Which is better?"	globale Benutzermeinung (Zufriedenheit, Akzeptanz) globale Leistungsmaße (Fehler, Zeiten)	Fragebogen Beobachtungen, Rechnerprotokolle, Videoaufzeichnungen
Eigenschaften prüfen "How good?"	Systemeigenschaften spezifische Benutzermeinung interaktionsbezogene Fehler	Experten-Checkliste Fragebogen (Rating) s.o.
Mängel feststellen "Why bad?"	Kritik ,Verbesserungs-vorschläge systembezogene Fehler	Exploration, Interviews, Videokonfrontation s.o.

Abb. 2: Die untersuchten Kombinationen von Evaluationzielen, -kriterien und -mitteln

Diese praxisorientierten Evaluationsansätze wurden im Rahmen von Querschnittstudien mit Systemen unterschiedlicher Funktionalität und Komplexität erprobt und weiterentwickelt. Im folgenden stelle ich diejenigen Aspekte der Bewertungsverfahren etwas genauer vor, auf die sich die späteren Ergebnisse beziehen.

Rechnerprotokolle : Durch Rechnerprotokolle werden üblicherweise Tastendrücke und Mausklicks der Benutzer automatisch registriert und mit einer Zeitmarke versehen. Um die per Logfile aufgezeichneten Daten inhaltlich interpretieren zu können, protokollierten wir bei unseren Logfiles zusätzlich Informationen über die Semantik und den Kontext der Benutzereingaben (z.B. welche Systemfunktion in welchem Programmteil ausgewählt wurde).

Beobachterprotokolle : Für die direkte Beobachtung durch den Versuchsleiter wurde ein Beobachtungsprotokoll konzipiert, das es erlaubt, die Beobachtungsdaten (z.B. das Auftreten spezifischer Fehler, Äußerungen der Benutzer etc.) per Graphiktablett direkt mit den Logfile-Daten zu verknüpfen.

Videoaufzeichnung : Die Versuchssitzungen wurden mittels zweier Kameras auf Videoband aufgezeichnet. Hierbei war eine Kamera auf die Benutzer und die andere auf den Computermonitor gerichtet. Über ein Videomischpult wurden beide Aufnahmen auf einem Bildschirm dargestellt sowie mit dem dazugehörigen Rechnerprotokoll synchronisiert.

Fragebogen : Wir entwickelten den "Fragebogen zur software-ergonomischen Bewertung von Dialogsystemen" (FBD), mit dem die Benutzer Aspekte der Benutzungsfreundlichkeit eines Dialogsystems aus ihrer Sicht bewerten. So schätzen die Benutzer u.a. ihre allgemeine Zufriedenheit mit dem Programm ein und geben eine differenzierte Beurteilung einzelner Programm-Eigenschaften. Der Fragebogen ist programmunabhängig formuliert und testtheoretisch breit erprobt, so daß er zur Bewertung unterschiedlicher Dialogsysteme eingesetzt werden kann.

Expertencheckliste : Wir konzipierten die "Expertencheckliste zur software-ergonomischen Bewertung von Dialogsystemen" (EBD), durch die arbeitspsychologische und software-ergonomische Aspekte eines rechnergestützten Arbeitssytems bewertet werden. Der Leitfaden soll jedoch nicht in "Konkurrenz" zu aufwendigeren Verfahren wie dem *EVADIS II* (Oppermann et al. [13]) treten, sondern versteht sich als schnell durchzuführendes "Screening-Verfahren". Eingestuft wird sowohl die *Ausprägung* eines Kriteriums als auch dessen *Bedeutung* für das konkrete System. Die programmgestützte Auswertung liefert anschauliche Soll-Ist-Vergleiche.

4 Die Längsschnittuntersuchung

Ich komme nun zur Anwendung der vorgestellten Bewertungsverfahren im Rahmen einer Längsschnittstudie, in deren Verlauf Software-Prototypen einer komplexen Datenbankanwendung entwickelt und evaluiert wurden. Die Entwicklung der Software-Prototypen erfolgte in einem sog. "Designteam", das Prototypen für das zu

entwickelnde Dialogsystem entwarf und diese in einem iterativen Prozeß überarbeitete (vgl. Ortlieb & Holz auf der Heide [14]). Der Prototyp wurde jeweils extern in Hinblick auf seine Benutzungsfreundlichkeit evaluiert, wobei (potentielle) Benutzer typische Arbeitsaufgaben mit dem Prototypen bearbeiteten. Die gewonnen Evaluationsergebnisse wurden in das Designteam rückgemeldet und dort in ein Redesign des Prototypen umgesetzt. Dieser Zyklus aus (Re-)Design und Evaluation wurde insgesamt viermal durchlaufen.

An der Evaluation der Prototypen nahmen 48 Bürokräfte teil (12 Personen pro Iteration), wobei wir Bürokräfte mit unterschiedlicher EDV-Vorfahrung auswählten. Der Ablauf der 3-4 Stunden dauernden Einzelversuche gliederte sich in vier Phasen:

Einführungsphase : Hier wurden die Bürokrafte über Ablauf und Hintergrund der Untersuchung informiert und es wurden demographische Daten, die Erfahrungen im Umgang mit Computern sowie die Einstellung gegenüber EDV erhoben.

Lern- und Explorationsphase : Sie diente dem Kennenlernen des Prototypen. Hierzu wurden die Benutzer in die Arbeit mit dem Prototypen eingeführt und aufgefordert, Systemfunktionen selbständig auszuprobieren und Fragen, Anmerkungen oder Kritik direkt zu äußern. Durch gezieltes Nachfragen wurden die von den Benutzern eingebrachten Hinweise konkretisiert. Im Anschluß erfolgte eine Überprüfung der Kenntnisse über den Prototypen und ggfs. eine Nachschulung, um bei allen Versuchsteilnehmern vergleichbare Eingangsbedingungen für die Testphase zu erreichen.

Testphase : Hier bearbeiteten die Bürokräfte sieben für das System typische Arbeitsaufgaben, die auf der Grundlage einer Aufgabenanalyse erstellt wurden. Die Aufgabenbearbeitung wurde durch *Logfiles*, *Videoaufzeichnung* und *direkte Protokollierung* durch den Evaluationsleiter festgehalten.

Nachbesprechungsphase : Hier schätzten die Bürokräfte zunächst anhand unseres Fragebogens *FBD* ihre allgemeine Zufriedenheit mit dem Prototypen ein und gaben eine differenzierte Beurteilung einzelner software-ergonomischer Eigenschaften. Daran anknüpfend wurden anhand der Videoaufzeichnung auffällige Stellen im Bearbeitungsablauf zusammen mit den Benutzern durchgegangen - d.h. es wurde eine *Videokonfrontation* durchgeführt.

5 Ergebnisse

Im folgenden werde ich aus der Fülle der im Rahmen der Längsschnittstudie gewonnenen Ergebnisse einige herausgreifen, die eine in sich geschlossene

Diskussion erlauben. Zum Abschluß werde ich auf den Vergleich der Evaluationsverfahren eingehen.

5.1 Die Erfassung von Benutzerfehlern

Zur Beantwortung der Frage "Which is better", d.h zum direkten Vergleich zwischen den Iterationen des Prototypen werteten wir u.a. die Fehler der Benutzer bei der Bearbeitung der Testaufgaben aus. Dabei interessierte uns die Leistungsfähigkeit der drei Erhebungsmittel *Rechnerprotokoll, Beobachterprotokoll (Logfile)* und *Videoaufzeichnung* hinsichtlich der Erfassung der Benutzerfehler.

Grundlage für einen entsprechenden quantitativen Vergleich bildete eine sog. "Ideale Erhebung" - eine kombinierte Auswertung der Daten aller eingesetzten Erhebungsmittel, quasi nach dem Gestaltprinzip: "Das Ganze ist mehr als die Summe seiner Einzelteile". Dabei wird eine vollständige Erfassung aller Fehler durch die Ideale Erhebung postuliert.

Bezogen auf unsere Längsschnittuntersuchung zeigte sich, daß von den 48 Bürokräften insgesamt 859 Fehler gemacht wurden. Hinsichtlich der Erfassung dieser Fehler erwies sich die Videoaufzeichnung deutlich als das leistungsfähigste Evaluationsmittel, während durch die direkte Beobachtung nur weniger als die Hälfte der aufgetretenen Fehler erfaßt werden konnten (vgl. Abb. 3).

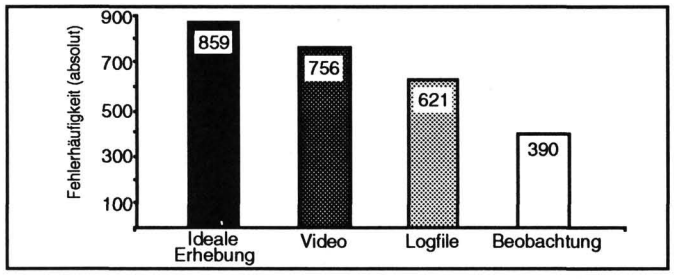

Abb. 3: Die durch die einzelnen Erhebungsmittel erfaßten Benutzerfehler

Wir haben die Benutzerfehler allerdings nicht nur global erhoben, sondern u.a. eine interaktionsbezogene Fehleranalyse im Sinne der Frage "How good" durchgeführt (vgl. Abschnitt 3). In der Literatur finden sich hierzu eine ganze Reihe unterschiedlich fundierter und differenzierter Fehler-Taxonomien (vgl. Arnold & Roe [2]). Unabhängig von den Möglichkeiten und Nutzen einzelner Fehler-Taxonomien ist natürlich die zuverlässige Fehler-Erfassung eine Voraussetzung für die weitere Fehleranalyse. Um diese Zuverlässigkeit zu prüfen, haben wir die drei Erhebungsmittel im Zusammenhang mit verschiedenen Fehlertaxonomien eingesetzt. Die dabei

gewonnenen Erkenntnisse sollen hier stellvertretend am Beispiel der im Projekt "Faust" erarbeiteten Fehler-Taxonomie verdeutlicht werden (Frese & Zapf [5]).

Diese handlungstheoretisch ausgerichtete Fehler-Taxonomie umfaßt insgesamt 15 Kategorien, von denen ich hier exemplarisch zwei Kategorien herausgreife, die ohne Hintergrundwissen verständlich sind. Es handelt sich um die "Wissensfehler" und die "Bewegungsfehler". Wissensfehler zeichnen sich dadurch aus, daß der Benutzer Handlungen nicht planen und ausführen kann, weil ihm dazu notwendige Informationen nicht zur Verfügung stehen. Unter die Bewegungsfehler fallen alle Fehler, die bei hochgeübten sensumotorischen Handlungen auftreten.

Bei Voruntersuchungen hatte sich gezeigt, daß für einen differenzierten Vergleich der drei Evaluationsmittel eine einfache Unterscheidung zwischen *Fehler erkannt* und *Fehler nicht erkannt* nicht ausreicht. Es gibt Benutzeraktionen, die zwar erfaßt werden, bei denen aber eine Zuordnung zur jeweiligen Fehlerkategorie je nach dem von der Methode miterfaßten Kontext schwierig oder unsicher ist. Um diese qualitativen Unterschiede zwischen den Erhebungsmitteln berücksichtigen zu können, unterschieden wir bei der Auswertung zwischen *richtig erkannt, unsicher erkannt, falsch erkannt* und überhaupt *nicht erkannt. Unsicher erkannt* ist ein Fehler immer dann, wenn die vom Erhebungsmittel gelieferten Informationen für eine eindeutige Zuordnung nicht ausreichen, so daß der Auswerter seine Einstufung auf Erfahrungswerte und Plausibilitätsannahmen stützen muß. *Falsch erkannt* ist ein Fehler, wenn der Auswerter seine Einstufung subjektiv sicher vornimmt, sich aber im nachhinein anhand der Idealen Erhebung herausstellt, daß die Einstufung objektiv falsch war. Die Ergebnisse, die sich bei dieser differenzierten Betrachtung ergeben, zeigen Abb. 4 (Wissensfehler) und Abb. 5 (Bewegungsfehler).

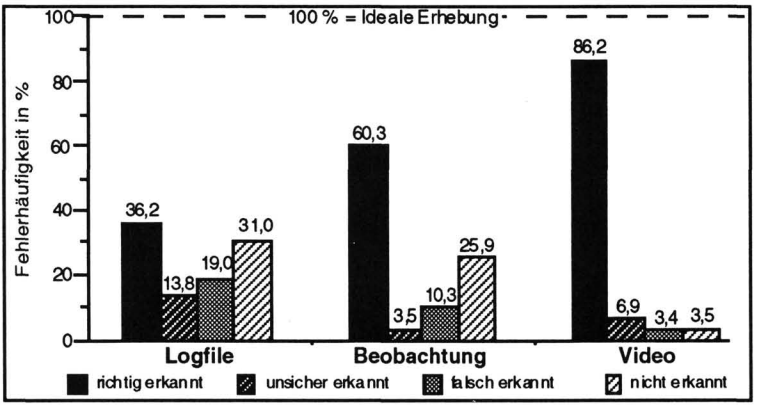

Abb. 4: Die Erfassung von Wissensfehlern

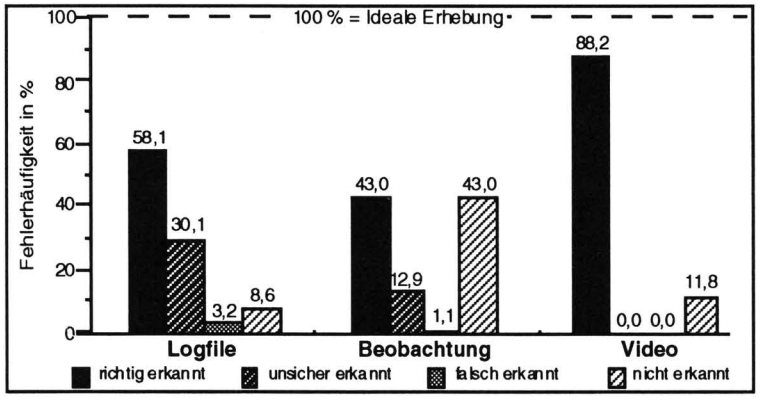

Abb. 5: Die Erfassung von Bewegungsfehlern

Bei den *Wissensfehlern* erweist sich die Videoaufzeichnung mit rund 86 % richtig erkannter Fehler deutlich als das leistungsfähigste Erhebungsmittel. Die beiden anderen Evaluationsmittel schneiden wesentlich schlechter ab, wobei die Beobachtung dem Logfile noch klar überlegen ist. Dieses Ergebnis ist plausibel, denn Wissensfehler zeichnen sich häufig dadurch aus, daß die Benutzer tatenlos vor dem Bildschirm sitzen und quasi nicht mehr weiter wissen. Das Erkennen solcher Situationen aus einem Logfile ist kaum möglich.

Auch bei der Erfassung von *Bewegungsfehlern* ist die Videoaufzeichnung mit rund 88 % richtig erkannter Fehler das leistungsfähigste Evaluationsmittel. Im Gegensatz zu den Wissensfehlern ist hier jedoch das Logfile der Beobachtung hinsichtlich der richtig erkannten Fehler überlegen. Dies ist naheliegend, denn bei Bewegungsfehlern handelt es sich um schnell ablaufende Prozesse, deren Beobachtung und Protokollierung einen Beobachter leicht überfordern können.

Insgesamt erwies sich bei unseren umfangreichen Untersuchungen eine *geeignete* Videoaufnahme als das leistungsfähigste Erhebungsmittel. Geeignet bedeutet hier, daß sowohl die Versuchsteilnehmer und die Eingabegeräte als auch die Benutzungsoberfläche zeitsynchron dargestellt werden. Die Rechnerprotokolle sind hinsichtlich der Datenerfassung der Beobachtung insgesamt leicht überlegen, während die durch die Beobachtung erhobenen Daten einfacher und sicherer interpretiert werden können. Abb. 6 faßt diese Ergebnisse zu den drei Erhebungsmitteln zusammen.

	Aufwand Realisierung	Aufwand Auswertung*	Daten-Erfassung	Daten-Interpretation
Beobachterprotokolle	gering	x 0,25	+	++
Rechnerprotokolle	mittel	x 0,5 - 0,8	++	+
Videoaufzeichnung	hoch	x 1,5 - 2	+++	+++

*(*Als Faktor der Aufgabenbearbeitungszeit)*

Abb. 6: Überblick über die drei Evaluationsmittel

5.2 Die Erfassung von Gestaltungsmängeln

Ein wichtiges Ziel im Rahmen einer software-ergonomischen Evaluation ist es, direkte Hinweise auf Gestaltungsmängel und entsprechende Verbesserungsvorschläge zu erhalten - also die Frage "Why bad?" zu beantworten. Dazu sind die diskutierten Fehlererhebungen insofern geeignet, als durch eine *systembezogene* Fehleranalyse bestimmte Schwachstellen identifiziert werden können. In der Praxis haben sich jedoch andere, mehr prozeßbezogene Vorgehensweisen besser bewährt: die *Systemexploration* durch die Benutzer in Verbindung mit einer *freien Befragung*, das Mitprotokollieren von Kommentaren und Kritik der Benutzer während der Aufgabenbearbeitung und die *Videokonfrontation.*

Durch diese Verfahren konnten beispielsweise bei der Evaluation der 1. Iteration des Prototypen insgesamt 154 konkrete Gestaltungshinweise von den 12 Bürofachkräften gewonnen werden (z.B. "Man hat einen schlechten Überblick über vorhandene Produktgruppen, eine Auflistung wäre hier besser"). Hierbei zeigten sich insbesondere zwei Ergebnisse:

- Benutzer mit großer EDV-Vorerfahrung gaben im Durchschnitt doppelt so viele Gestaltungshinweise wie unerfahrene Benutzer, wobei sie rund 2/3 ihrer Hinweise bereits in der Explorationsphase äußerten, also *bevor* sie die Testaufgaben bearbeitet hatten. Sie nannten mehr konkrete Verbesserungsvorschläge und kritisierten öfter bestimmte Eigenschaften des Prototypen, wobei sie sich häufig auf ihre Erfahrungen mit anderen Programmen bezogen. Dagegen äußerten Benutzer mit geringer EDV-Vorerfahrung nur etwa 1/3 ihrer Kritik und Gestaltungshinweise in der Lernphase, der überwiegende Teil wurde während und nach der Bearbeitung der Testaufgaben genannt. Die Hinweise waren insgesamt nicht so gestaltungsnah, sondern bezogen sich mehr auf die Schwierigkeiten, die die Benutzer mit dem Prototypen hatten. Man könnte nun meinen, daß man unter diesen Umständen auf die Bearbeitung von Testaufgaben verzichten und nur erfahrene Benutzer befragen sollte. Dies wäre jedoch nicht sinnvoll, da es deutliche *qualitative* Unterschiede zwischen den jeweiligen

Gestaltungshinweisen gibt. Ein Teil der Hinweise ginge deshalb verloren, würde man nur EDV-Fortgeschrittene in die Evaluation einbeziehen.

- Nach etwa 6 Personen (davon 3 mit großer EDV-Erfahrung) wurden kaum noch neue Hinweise genannt. Es ist somit nicht erforderlich, Versuche mit vielen Benutzern durchzuführen, um praxisrelevante Ergebnisse zu erhalten.

5.3 Vergleich der eingesetzten Evaluationsverfahren

Zum Abschluß dieses Beitrages möchte ich kurz auf die Frage eingehen, inwieweit sich die hier vorgestellten Verfahren auf eine *gemeinsame* Dimension beziehen, die gemeinhin als "Benutzungsfreundlichkeit" bezeichnet wird. Um eine erste Antwort auf diese Frage zu erhalten, haben wir die durch die einzelnen Evaluationsverfahren gewonnenen Daten in prozentuale Werte umgerechnet, wobei die Anzahl der Gestaltungshinweise, die Bearbeitungszeiten und die Fehlerzahlen der *1. Iteration* jeweils als "100%-Wert" festgelegt wurden. Auf diese Weise wurde einerseits eine anschauliche Erfolgskontrolle unserer Software-Entwicklung über vier Iterationen und andererseits ein direkter trendorientierter Vergleich zwischen den Ergebnissen der Evaluationsverfahren möglich (vgl. Abb. 7).

Bezogen auf den *Entwicklungsprozeß* zeigte sich eine stetige Verbesserung des Prototypen über den Entwicklungszeitraum: Die Bearbeitungszeiten und Fehler typischer Benutzer reduzierten sich drastisch, während sich die Zufriedenheit der Benutzer mit dem Prototypen und die Bewertung durch Experten für Software-Ergonomie stetig verbesserte. Die größte Verbesserung wurde dabei durch die Überarbeitung des Prototypen von der ersten zur zweiten Iteration erzielt.

Bezogen auf die *Evaluationsverfahren* wird der starke (statistisch abgesicherte) Zusammenhang zwischen den Ergebnissen der einzelnen Evaluationsverfahren deutlich. Dieses Ergebnis bejaht insofern die oben gestellte Frage nach einer gemeinsamen Dimension von Benutzungsfreundlichkeit.

Hinsichtlich der Benutzerbefragung muß hier jedoch eine gewisse Einschränkung vorgenommen werden: Im Gegensatz zur Erfassung von Gestaltungsmängeln sollte sich die globale Bewertung eines Dialogsystems nicht auf Benutzer mit geringen EDV-Vorerfahrungen stützen, da deren Bewertung oftmals einen starken "Trend zur Mitte" aufweist. D.h. EDV-Anfänger bewerten - gemessen an anderen Daten - software-ergonomisch ungenügende Dialogsysteme tendenziell zu gut, und software-ergonomisch ausgereiftere Systeme tendenziell zu schlecht.

Abgesehen von dieser Einschränkung hat das vorliegende Ergebnis meines Erachtens wichtige Implikationen für die software-ergonomische Forschung und

Praxis. So deutet es u.a. an, daß bei etlichen Fragestellungen auf die Durchführung aufwendiger Verfahren zugunsten ökonomischerer Vorgehensweisen verzichtet werden kann. Nicht zuletzt deshalb erscheint es ratsam, die sinnvollen Einsatzbereiche und -bedingungen der einzelnen Evaluationsverfahren konsequent zu berücksichtigen.

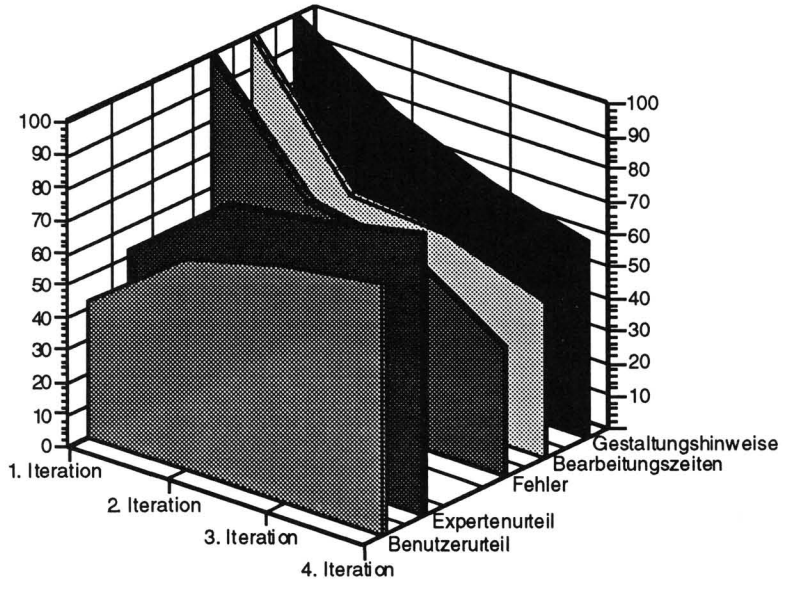

Abb. 7: Ergebnisse der Evaluationsverfahren über vier Iterationen des Prototypen

6 Literatur

[1] Apple (1987). Human Computer Interface Guidelines. The Apple Desktop Interface. Addison Wesley.

[2] Arnold, B. & Roe, R. (1987). User errors in human-computer interaction. In: M. Frese, E. Ulich & W. Dzida (Eds.), Psychological issues of human-computer interaction in the work place (pp. 203-220). Amsterdam: North Holland.

[3] DIN 66234 Teil 8. (1988). Bildschirmarbeitsplätze, Grundsätze ergonomischer Dialoggestaltung.

[4] Floyd, C. (1984). A Systematic Look at Prototyping. In: R. Budde, K. Kuhlenkamp, L. Matiassen & H. Züllighoven (Eds.), Approaches to Prototyping. Heidelberg: Springer.

[5] Frese, M. & Zapf, D. (Hrsg.) (1991). Fehler bei der Arbeit mit dem Computer. Ergebnisse von Beobachtungen und Befragungen im Bürobereich. Bern: Huber.

[6] Gould, J. D. & Lewis, C. (1984). Designing for usability - key principles and what designers think. In: Human-Computer Interaction, Proceedings of the ACM, 50-53.

[7] Helmreich, R. (1984). Geplante Akzeptanz im Büro. Benutzer, Organisation und Technik - Aspekte für die Gestaltung von Bürosystemen. In: data report 19, Heft 3.

[8] Hampe-Neteler, W. & Rödiger, K. H. (1992). Software-Ergonomie - Verfahren der Evaluierung und Standards zur Entwicklung von Benutzeroberflächen. Forschungsbericht des Studiengangs Informatik der Universität Bremen, Bericht Nr. 2/92.

[9] Holz auf der Heide, B. & Hacker, S. (1991). Prototyping in einem Designteam: Vorgehen und Erfahrungen bei einer Software-Entwicklung unter Benutzerbeteiligung. In D. Ackermann & E. Ulich (Hrsg.), Software-Ergonomie '91, (S. 108 - 118). Stuttgart: Teubner.

[10] IBM (1991). Systems Application Architecture. CUA. Advanced Interface Design Guide.

[11] ISO 9241 Part 10. Ergonomic requirements for office work with visual display terminals (VDTs). Dialogue principles - Second committee draft, June 1992.

[12] Norman, K., & Shneiderman, B. (1989). Questionaire for user interface satisfaction Vers. 5.0. Maryland: University of Maryland, HCI-Lab.

[13] Oppermann, R., Murchner, B., Reiterer, H. & Koch, M. (1992). Software-ergonomische Evaluation. Der Leitfaden EVADIS II. Berlin: DeGruyter.

[14] Ortlieb, S. & Holz auf der Heide, B. (1993). Benutzer bei der Software-Entwicklung angemessen beteiligen - Erfahrungen und Ergebnisse mit verschiedenen Konzepten. In diesem Band.

[15] Piepenburg, U. & Rödiger, K.-H. (1989). Mindestanforderungen an die Prüfung auf Konformität nach DIN 66234, Teil 8 (Werkstattbericht Nr. 61 der Reihe "Mensch und Technik"), Druckerei Hartmann, Nordrhein-Westfalen, 1989

[16] Prümper, J. & Anft, M. (1993). Die Evaluation von Software auf Grundlage des Entwurfs zur internationalen Ergonomie-Norm ISO 9241 Teil 10 als Beitrag zur partizipativen Systemgestaltung - ein Fallbeispiel. In diesem Band.

[17] Rauterberg, M. (1992). Läßt sich die Gebrauchtauglichkeit interaktiver Software messen? Und wenn ja, wie? Ergonomie & Informatik, 16, 3-18.

[18] Rouse, W. B. & Rouse, S. (1983). Analysis and classification of human error. IEEE Transactions on Systems, Man, and Cybernetics, SMC13 (4), 539-549.

[19] Smith, S. L. & Mosier, J. (1986). Guidelines for Designing Interface Software, MITRE, Bedford

[20] Shneiderman, B. (1992). Designing the user interface: Strategies for effective human-computer interaction. Reading, Massachusetts: Addison-Wesley.

[21] Spinas, P. (1987). Arbeitspsychologische Aspekte der Benutzerfreundlichkeit von Bildschirmsystemen. Zürich: ADAG.

Bernd Holz auf der Heide
Technische Universität München
Lehrstuhl für Psychologie
Lothstraße 17
8000 München 2