

che Fragen werden beantwortet, sodass er die allgemeine Struktur und den Ablauf versteht und in etwa vorhersehen kann. Gerade wenn Methoden wie (Video-)Beobachtung eingesetzt werden, die zunächst hemmend wirken können, sollte dem Teilnehmer auch deutlich gemacht werden, dass es nicht um die Bewertung seiner Person bzw. seiner Leistung, sondern ausschließlich um die Bewertung des Produktes geht. Danach folgt die eigentliche **Testphase**, in der der Teilnehmer die gestellte Aufgabe bearbeitet und die abhängig von den eingesetzten Methoden sehr unterschiedlich ablaufen kann. Wichtig in dieser Phase ist es, dass die Interaktion zwischen dem Teilnehmer und dem Versuchsleiter auf ein Minimum beschränkt bleibt, um eine Beeinflussung auf ein Mindestmaß zu reduzieren. Nach der Aufgabebearbeitung wird der Teilnehmer häufig gebeten, noch Fragebögen hinsichtlich subjektiver Maße (z. B. Zufriedenheit) oder Personenangaben (z. B. Einstellungen gegenüber Computern oder Belastungen) auszufüllen. Anschließend sollte es eine **Abschlussbesprechung** geben, die dem Teilnehmer Gelegenheit für Kommentare und Anregungen gibt und dem Versuchsleiter die Möglichkeit einräumt, sich während des Tests aufgetretene Fragen erläutern zu lassen.

Die so erhobenen Daten werden in der **Analysephase** ausgewertet. Dies kann zum einen mit recht einfachen Maßen entlang der ISO 9241-11-Kriterien geschehen, aber auch komplexere Auswertungen, z. B. anhand von Fehlertaxonomien (s. Abschnitt „Beispiel für eine formative Evaluation“), enthalten. Das häufigste verwendete Leistungsmaß ist die Zeit, die ein Nutzer benötigt, um eine gegebene Aufgabe mit dem untersuchten System zu bewältigen. Die Effektivität als weiteres Maß der *Usability* setzt die Quantität („Wie viel Prozent der gestellten Aufgabe wurden erfüllt?“) multiplikativ mit der Qualität des Ergebnisses in Verbindung. Hierbei fließt die benötigte Zeit nicht mit in die Berechnung ein. Sie wird jedoch bei einem weiteren Maß – der Effizienz – verwendet, in dem einfach die Effektivität durch die Zeit dividiert wird. Die Zufriedenheit, als letztes in der Definition der

Usability benanntes Maß, wird meist über Fragebögen erhoben und inferenzstatistisch ausgewertet (vgl. Bortz u. Döring 2002). Ergänzend können hierzu auch Beobachtungsdaten hinzugezogen werden. Komplexere Maße der *Usability* sind z. B. Fehlerquoten, deren Auswertung von der verwendeten Taxonomie abhängen. Auch die systematische Analyse von Protokollen des Lauten Denkens ist recht komplex und bedarf spezieller Vorkenntnisse. In der **Berichtsphase** werden dann für den „Auftraggeber“ (z. B. Entwickler, Management) die Methoden und Ergebnisse niedergelegt und Empfehlungen ausgesprochen.

Beispiel für eine formative Evaluation

Als Beispiel für eine formative Evaluation wird im Folgenden die Untersuchung des Pflegeinformationssystems PIK (vgl. Mahler et al. in diesem Buch) dargestellt (ausführliche Ergebnisse bei Büssing u. Herbig 1998 a; Herbig u. Büssing 2002). Basismethoden dieser Evaluation waren das oben bereits dargestellte Laute Denken und eine handlungsorientierte Fehlertaxonomie nach Zapf et al. (1989). In der Literatur herrscht Einigkeit über die Notwendigkeit des Einsatzes von Fehlertaxonomien bei der Untersuchung der Mensch-Computer-Interaktion. Denn nur mit fehlertaxonomisch begründeten Daten können zuverlässige, fehlerfreundliche Systeme geschaffen werden, um die nicht selten folgenschweren Auswirkungen von Fehlern und eine Belastung der Benutzer durch aufwendige und schwierige Fehlerkorrekturen zu vermeiden. Bei der verwendeten Fehlertaxonomie handelt es sich um eine Einteilung, die auf der Handlungsregulationstheorie basiert; sie bietet den Vorteil, konkrete Aussagen darüber machen zu können, an welchem Punkt der Handlungsregulation es zu einem Fehler gekommen ist (Tab. 10-2).

Bei Handlungen kann zwischen dem hierarchischen und dem sequenziellen Aufbau unterschieden werden. Der hierarchische Aufbau beschäftigt sich mit der Ebene der Regulation. Auf

Tab. 10-2 Handlungsorientierte Fehlertaxonomie (modifiziert nach Zapf 1991)

Nutzungsprobleme			
Regulationsgrundlage	Wissensfehler		
Regulationsebene	Schritte im Handlungsprozess		
	Ziele/Planung	Gedächtnis/Monitoring	Rückmeldung
intellektuelle Regulationsebene	Denkfehler	Merk-/Vergessensfehler	Urteilsfehler
Ebene der flexiblen Handlungsmuster	Gewohnheitsfehler	Unterlassensfehler	Erkennensfehler
sensumotorische Regulationsebene	Bewegungsfehler		

der **intellektuellen Regulationsebene** finden sich bewusste Analyseprozesse zur Regulation komplexer Aktivitäten, für die noch keine Handlungsmuster existieren. Auf der **Ebene flexibler Handlungsmuster** müssen bereits bestehende Grundmuster für eine Handlung an die Situation angepasst werden. Auf der **sensumotorische Regulationsebene** finden sich hoch automatisierte Handlungen. Sequenziell lassen sich Handlungen trennen in eine **Planungsphase**, in der Ziele gesetzt und Pläne entwickelt werden, eine **Überwachungsphase**, in der die Planausführung überwacht wird und Ziele und Unterziele erinnert werden müssen, und eine **Rückmeldungsphase**, in der bewertet wird, ob das gesetzte Ziel erreicht wurde. Da eine sequenzielle Trennung auf der sensumotorischen Ebene nicht möglich ist, ergeben sich sieben potenzielle Fehlerarten zuzüglich den so genannten Wissensfehlern, die bei der Grundlage der Handlungsregulation auftreten können.

Um die auftretenden Fehler möglichst genau inhaltlich bestimmen zu können, wurden die Fehlerprotokolle über einen Zeitcode mit dem Lauten Denken zusammengeführt. Die aufeinander folgenden Untersuchungen von drei Versionen des Programms wurden im Labor mit 26 Krankenpflegekräften durchgeführt. Ein ausführlich geschulter Untersuchungsleiter beobachtete die Personen beim Lösen von Pflegeplanungs- und Pflegedokumentationsaufgaben über einen zweiten Monitor und notierte die

auf tretenden Fehler. Das Laute Denken wurde aufgezeichnet und in der Analysephase mit den aufgetretenen Fehlertypen verglichen.

Die Abbildung 10-2 zeigt die Ergebnisse der Fehlertaxonomie für die drei Studien. Exemplarisch wird das Zusammenspiel von Fehlertaxonomie und Lautem Denken anhand der Studie 2 dargestellt. Diese Studie beschäftigte sich mit den Modulen der Stationsorganisation und der Pflegeplanung. Die Versuchspersonen sollten einen Patienten auf Station aufnehmen und eine Pflegeplanung für den Patienten mit Apoplexie basierend auf einem Standardpflegeplan durchführen, der entsprechend den speziellen Problemen des Patienten anzupassen war. In der Fehlertaxonomie fiel eine hohe Anzahl an Urteilsfehlern auf, also Fehlern in der Rückmeldungsphase auf der intellektuellen Regulationsebene. Ein Urteilsfehler besteht beispielsweise aus der falschen Interpretation einer Systemrückmeldung. In den Protokollen des Lauten Denkens zeigte sich schnell, dass ein Teil der Probleme bei dem Übergang zwischen Patientenaufnahme und Pflegeplanung auftrat. Die Personen hatten ein falsches mentales Modell aufgebaut, aufgrund dessen sie erwarteten, dass nach der Eingabe der Diagnose automatisch eine Pflegeplanung generiert werden würde. Sie reagierten verwirrt und verloren die Orientierung im Programm, wenn nach ihrer Eingabe nichts geschah.

Überwanden sie dieses Problem auf der intellektuellen Ebene, zeigte sich ein zweiter Pro-

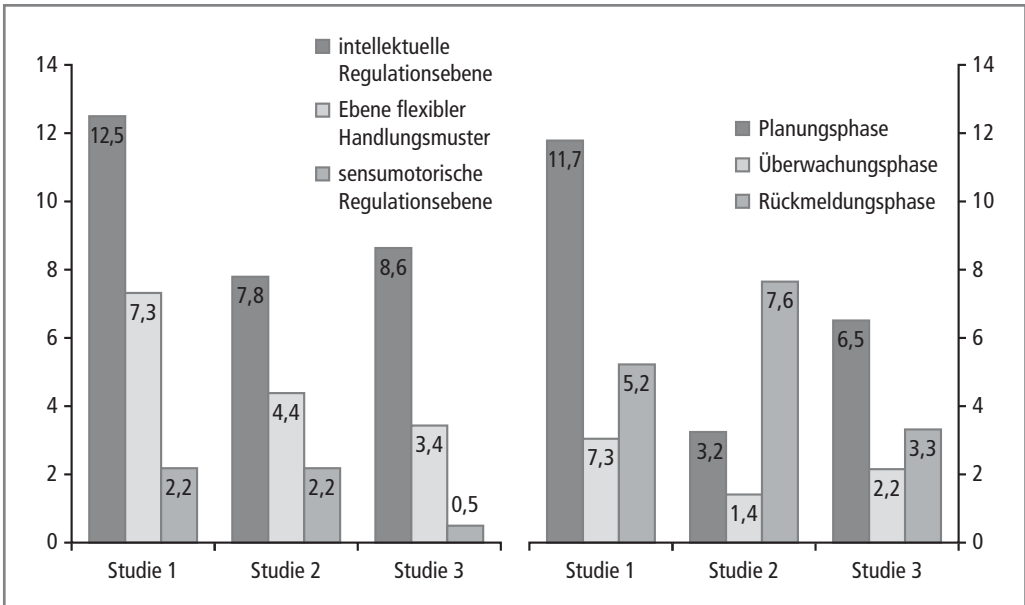


Abb. 10-2 Fehlermittelwerte auf den hierarchischen Regulationsebenen und in den sequenziellen Verarbeitungsphasen

blemkomplex in der Rückmeldungsphase: Obwohl die meisten Personen die Auswahl und die Modifikation von Standardplänen korrekt planen und ausführen konnten, war es ihnen oft nicht möglich nachzuvollziehen, ob sie eine Handlung erfolgreich beendet hatten oder nicht. Die Protokolle des Lauten Denkens zeigten, dass verschiedene Kontrollvarianten in unterschiedlichen Menüs und unzureichende oder fehlende Systemmeldungen, z. B. über Eingabeformate, es für die Nutzer unmöglich machten festzustellen, ob ihre Aktion vom System durchgeführt worden war. In einigen Fällen führte dies sogar dazu, dass dieselbe Handlung wieder und wieder (mit steigender Frustration) durchgeführt wurde, obwohl sie bereits beim ersten Mal erfolgreich ausgeführt wurde. Dieses leicht zu behebbende Problem wurde den Entwicklern zurückgemeldet und in der dritten Studie konnte ein deutlicher Rückgang dieser Fehlerart verzeichnet werden.

Generell zeigt die Abbildung 10-2 den Erfolg dieser Form der formativen Evaluation: Im Verlauf der Untersuchung gelang es, die Anzahl der Nutzungsprobleme generell und signifikant ab-

zusenken. Auch die abschließenden Bewertungen der Teilnehmer bestätigten dies: War der Tenor in Studie 1 noch „Das ist wirklich sehr kompliziert“, fand sich in Studie 2 die generelle Bewertung „Das Programm ist irgendwie knifflig“ und in Studie 3 die bei jeder Evaluation erhoffte, positive Aussage „Das Programm ist auf jeden Fall eine Erleichterung“.

Fazit

Für den erfolgreichen Einsatz von IuK-Technologien im Krankenhaus ist ihre systematische Evaluation vor allem hinsichtlich der Usability unabdingbar. Wie der kurze Überblick in diesem Kapitel zeigen sollte, stehen dafür eine ganze Reihe unterschiedlicher Methoden zur Verfügung, über deren Einsatz und Nutzen jeweils nur aufgrund der spezifischen Gegebenheiten des Krankenhauses, der Abteilung, des Funktionsbereiches usw. entschieden werden kann. Diese Entscheidungen sind wesentlich, um von den Nutzern akzeptierte und auch in einem komple-