3. Kapitel

Die Inter-Rater-Reliabilität von Fehlerbeobachtungen im Feld

Jochen Prümper

Bei der Erhebung von Beobachtungsdaten spielt die Reliabilität als Gütemaß der Beobachterleistung eine zentrale Rolle. Das vorliegende Kapitel dient deshalb dem Zweck, die Reliabilitäten der im Rahmen des Forschungsprojektes FAUST zum Einsatz gekommenen Beobachtungsinstrumente darzustellen. Im einzelnen kommen dabei zur Sprache: 1. die Handlungsorientierte Fehlertaxonomie, 2. die Fehlerentdeckung, 3. die Fehlerbewältigung und 4. die Fehlerkorrekturzeit. Schließlich soll überprüft werden, ob das im Rahmen der Hauptuntersuchung angewendete Rating-Verfahren schriftlicher Ereignisprotokolle ein zur Doppelbeobachtung äquivalentes Verfahren darstellt.

3.1 Problemstellung

Ist man an den Problemen und Schwierigkeiten interessiert, die bei der alltäglichen Arbeit mit Computern auftauchen, so stellt sich die Frage, wie sie sich möglichst zuverlässig erfassen lassen.

Man könnte beispielweise die Benutzer von Softwareprogrammen bitten, über einen bestimmten Zeitraum hinweg ein Fehlertagebuch zu führen; man könnte sie nach Fehlerereignissen befragen oder die Benutzerberatung der Programmhersteller interviewen, mit welchen Anliegen sie konfrontiert wird. Alle diese Methoden haben jedoch ihre Vor- und Nachteile. Sie erfassen lediglich einen begrenzten Ausschnitt der Probleme und Schwierigkeiten, die im täglichen Umgang mit Computern auftreten. Die Benutzer selbst nehmen z.B. manche Fehler überhaupt nicht wahr, oder sie erinnern sich im nachhinein nur noch verzerrt an sie. Auch verfügen die Benutzer, wenn sie einen Fehler bemerken, oft nicht über die zur Selbstreflexion nötige Distanz oder über die sprachlichen Mittel, um einen Fehler adäquat zu beschreiben.

Die Beobachtung weist gegenüber der Befragung oder der Dokumentenanalyse - Methoden, bei denen die Sichtweise der Benutzer im Vordergrund steht, - den Vorteil auf, daß die Schwierigkeiten und Probleme, die bei der alltäglichen Arbeit mit Computern auftreten können, über einen längeren Zeitraum hinweg detailliert, unmittelbar und in ihrem jeweiligen Kontext erfaßt werden.

Doch auch die Verhaltensbeobachtung wirft Probleme auf (vgl. Faßnacht, 1979), und in der einschlägigen Literatur werden verschiedene Ursachen angeführt, die die Reliabilität von systematisch erhobenen Beobachtungsdaten negativ beeinflussen können. Brandstätter (1970) hebt beispielsweise mangelnde Objektivität der Beschreibung und Interpretation, mangelnde Stabilität der Verbindung von beobachteter Anregung und Reaktion und mangelnde Äquivalenz der Auswahl von Situationen in den Vordergrund. Johnson und Bolstad (1973) nennen Beobachterfehler, Beobachtererwartungseffekte und Reaktivitätseffekte als Störfaktoren und Lienert (1978) unterschiedliche Erkennungsschwellen und Wahrnehmungsaspekte der zu beobachtenden Merkmale (für weitere Kategorisierungen möglicher Ursachen mangelnder Reliabilität von Beobachtungsdaten vgl. z.B. Johnson, 1955; Webb, Campbell, Schwartz & Sechrest, 1966; zu einer Diskussion des Beobachterfehlers vgl. Feger & Graumann, 1983).

Darüber hinaus bedarf es bei der Untersuchung von einer Vielzahl von Computerarbeitsplätzen eines hohen Wissens über sämtliche Systeme, die zum Einsatz kommen - von der PC Standardsoftware bis hin zu spezifischen Anwendungen für firmeninterne Fachabteilungen an diversen Großrechnern. Zum anderen ist es vonnöten, einen genügend großen Einblick in die diversen Arbeitsaufgaben unterschiedlichster Benutzer zu gewinnnen (von einfacheren Sekretariatsaufgaben bis hin zu sehr komplexen Sachbearbeitertätigkeiten). Dies sind Forderungen, die in ihrem vollen Umfang von einem einzigen Beobachter nahezu nicht zu erfüllen sind.

Bedenkt man dann noch, daß die Beobachtung von alltäglichen Arbeitsaufgaben an sich sehr schwierig ist, da hochgeübte und damit schnell ablaufende Routinen zum Einsatz kommen, und daß sich Menschen ungern bei Fehlern beobachten lassen, daß diese gerne vertuscht und möglichst schnell korrigiert werden, dann ergeben sich bei einer Feldbeobachtung von Benutzerfehlern am Computer Schwierigkeiten, die von vornherein lediglich geringe Reliabilitäten erwarten lassen (für eine Diskussion der Probleme, die bei Beobachtungsverfahren in der psychologischen Arbeitsanalyse auftreten können, siehe: Frenz & Frey, 1981; Zapf, 1989b).

Nach von Cranach und Frenz (1969) erscheint es dann auch wegen der größeren Meßgenauigkeit aus vier Gründen angeraten, für das Erheben von Beobachtungsdaten technische Einrichtungen (z.B. mechanische Verhaltensschreiber, Sprach-, Film-, Videoaufnahmegeräte etc.) zu verwenden. Erstens können durch diese technischen Mittel die Anforderungen an die Aufmerksamkeit des Beobachters reduziert werden; zweitens können dem Beobachter in der Phase der Datensammlung durch sie Entscheidungen bzw. Beurteilungen abgenommen werden; drittens besteht dann die Möglichkeit, mehr zu registrieren als ein menschlicher Beobachter vermag; und viertens bietet ein solches Verfahren die Möglichkeit, Beobachterdaten mit den auf objektivere Weise erhobenen Daten zu vergleichen und somit ihre Zuverlässigkeit und Gültigkeit zu überprüfen (siehe hierzu auch Frey, Hirsebrunner & Bieri-Florin, 1979; Müller-Holz auf der Heide, Aschersleben, Hacker & Bartsch, 1991).

Bei Betriebsuntersuchungen muß allerdings weitestgehend auf den Einsatz maschineller Aufzeichnungsgeräte verzichtet werden. Die Sensibilität gegenüber dem Schutz persönlicher Daten ist sowohl bei den Beschäftigten als auch bei den verantwortlichen Betriebsräten mittlerweile so ausgeprägt, daß ihr Einsatz in Betrieben nahezu unmöglich geworden ist.

Angesichts dieser Schwierigkeiten würde es sich anbieten, Doppelbeobachtungen durchzuführen. Ein zweiter Beobachter könnte als Kontrollorgan des Fehlergeschehens

fungieren, könnte dort fachliche Ergänzung sein, wo der andere an seine Grenzen stößt, und aufmerksam sein, wenn der andere noch protokolliert.

Obwohl diese Vorgehensweise zunächst optimal erscheinen mag, wirft jedoch auch sie neue Schwierigkeiten auf. Einerseits beeinflußt die Anwesenheit von zwei Beobachtern das Verhalten der beobachteten Personen u.U. noch stärker als dies schon bei einem Beobachter der Fall ist (vgl. Mees, 1977c), und andererseits verbietet sich dieses Vorgehen in vielen Fällen schon allein aufgrund seiner hohen Kosten.

In Anbetracht dieser Vorüberlegungen erscheint uns zur zuverlässigen Beantwortung der Frage, welche Probleme und Schwierigkeiten bei der alltäglichen Arbeit mit Computern auftauchen, die natürliche, passiv-teilnehmende und technisch unvermittelte, systematische Einzelbeobachtung die Methode der Wahl[1].

Die Entscheidung für Einzelbeobachtungen bringt allerdings das Problem mit sich, daß zunächst einmal eine unmittelbare Überprüfung der Inter-Rater-Reliabilität nicht möglich ist. Um diesem Gütekriterium dennoch gerecht zu werden, besteht jedoch die Möglichkeit, die schriftlichen Niederlegungen der Einzelbeobachtungen von zwei unabhängigen Personen kategorisieren zu lassen und ihre Übereinstimmungen dann als Maß der Reliabilität heranzuziehen.

Verfährt man auf diese Weise, so ist dabei freilich nicht sichergestellt, ob die Beurteilung der aus den Einzelbeobachtungen resultierenden Ereignisprotokolle von nicht an der Beobachtung Beteiligten ein äquivalentes Verfahren zu einer tatsächlichen Doppelbeobachtung darstellt, bei der die Beobachter selbst ihre Aufzeichnungen kodieren. Diese Äquivalenz gilt es zu überprüfen, indem die Kodierungen aus tatsächlichen Doppelbeobachtungen mit den Ratings von entsprechenden Beobachtungsprotokollen verglichen werden.

In den Fällen, in denen die Beobachter selbst diejenigen sind, die das Geschehen registrieren, schriftlich fixieren und kodieren, sprechen wir im folgenden von **Primär-Ratings**. Wenn die Ereignisprotokolle von nicht an der Beobachtung Beteiligten kodiert werden, sprechen wir von **Sekundär-Ratings**.

3.2 Methode

Nachdem zunächst die Vorgehensweise bei der Datenerhebung, die Stichproben und die statistischen Kenngrößen beschrieben werden, behandeln die folgenden Ausführungen die Überprüfung der Reliabilität der eingesetzten Beobachtungssysteme und anschließend die Überprüfung der Äquivalenz des in der Hauptuntersuchung angewendeten Rating-Verfah-

---

[1] Die systematische Beobachtung (Synonyme: kontrolliert; strukturiert; standardisiert) ist im Gegensatz zur freien Beobachtung durch Festlegung, Vereinheitlichung und Kontrolle gekennzeichnet. Durch Festlegung eines Kategoriensystems und Vereinheitlichung der Vorgehensweise ermöglicht ein systematisches Beobachtungsverfahren die Kontrolle der Gütekriterien, die auch für andere psychologische Datenerhebungsmethoden gelten; also Prüfungen v.a. hinsichtlich der Validität und der Reliabilität. Natürlich bedeutet in diesem Zusammenhang, daß die Beobachtungen in der natürlichen Umgebung der beobachteten Person und nicht im Labor stattfinden. Passiv-teilnehmend meint, daß der Beobachter so wenig wie möglich mit dem Beobachteten interagiert und eine technisch unvermittelte Beobachtung bedeutet, daß der Beobachter ohne Zuhilfenahmen maschineller Aufzeichnungen direkt als Protokollant fungiert (vgl. Köhne, 1979; Mees, 1977a). Die Ergänzung der systematischen Beobachtung um Einzelbeobachtungen bringt zum Ausdruck, daß lediglich ein Beobachter pro beobachteter Person zum Einsatz kommt.

rens. Das Hauptaugenmerk liegt dabei auf der Darstellung und Diskussion der Fehlertaxonomie.

3.2.1 Vorgehensweise bei der Datenerhebung

Zusammengefaßt erfolgte die Datenerhebung in vier Schritten:
1. Schritt - Primär-Rating der Einzelbeobachtungen: Im Rahmen der Hauptuntersuchung kam pro beobachteter Person jeweils ein Beobachter zum Einsatz. Die Beobachtungen wurden von ihm selbst schriftlich fixiert und kodiert.
2. Schritt - Sekundär-Rating der Einzelbeobachtungen: Die von den Primär-Ratern schriftlich niedergelegten Fehlerbeschreibungen wurden von zwei weiteren Personen unabhängig voneinander ein weiteres Mal kategorisiert. Lediglich die Fehler, die diese beiden Sekundär-Rater übereinstimmend in dasselbe Taxon einordneten, wurden in der Folge bei weiteren Berechnungen berücksichtigt.
3. Schritt - Primär-Rating der Doppelbeobachtungen: Es wurden zwei kleinere Studien mit tatsächlichen Doppelbeobachtungen durchgeführt. Auch hier wurden die Beobachtungen von den jeweiligen Beobachtern selbst schriftlich festgehalten und kodiert.
4. Schritt - Sekundär-Rating der Doppelbeobachtungen: Um zu überprüfen, ob das Sekundär-Rating der Einzelbeobachtungen eine äquivalente Methode zur Erhöhung der Reliabilität der Einzelbeobachtungen darstellt, fand ein Sekundär-Rating der schriftlich niedergelegten Fehlerbeschreibungen der Doppelbeobachtungen statt.

3.2.2 Stichproben

Die erste Doppelbeobachtung erfolgte bei einer Betriebsstichprobe, die mit der der Hauptuntersuchung vergleichbar ist; die zweite Doppelbeobachtung fand bei einem studentischen Klientel statt (zur Beschreibung der Hauptuntersuchung vgl. Anhang).

3.2.2.1 Doppelbeobachtung I

Bei den 23 Untersuchungsteilnehmern der ersten Doppelbeobachtung handelte es sich um 14 Mitarbeiter einer großen Versicherung (10 weiblich, 4 männlich) und um 9 Sekretärinnen eines EDV-Großunternehmens. Das Durchschnittsalter lag bei 26 Jahren. Die durchschnittliche Erfahrung mit dem benutzten Programm lag zwischen ein und zwei Jahren. Insgesamt waren fünf Beobachter an der Doppelbeobachtung beteiligt. Die Beobachtungen fanden an insgesamt sieben Tagen in einem Zeitraum von zwei Wochen statt. Die Beobachtungsprotokolle und die Klassifikationen wurden von beiden Beobachtern unabhängig voneinander ausgefüllt. Im übrigen war die Vorgehensweise identisch zur Hauptuntersuchung.

3.2.2.2 Doppelbeobachtung II

Bei den 23 Untersuchungsteilnehmern der zweiten Doppelbeobachtung handelte es sich um 17 Studenten und 6 Studentinnen der Betriebswirtschaftslehre, die im Rahmen einer

einsemestrigen Veranstaltung eine Einführung in ein integriertes, menügesteuertes betriebswirtschaftliches Standard-Softwareprogramm (Stammdatenverwaltung, Kundenauftragsverwaltung, Lagerbestandsführung, Lohn/Gehalt, etc.) erhielten. Das Durchschnittsalter lag bei 24 Jahren. 6 Teilnehmer hatten das untersuchte Programm bereits in einem vorhergegangen Semester kennengelernt.

Bei dieser Studie, bei der zwei Beobachter beteiligt waren, handelte es sich um eine Längsschnittuntersuchung zu zwei Meßzeitpunkten (vgl. Kap. 8). Die Beobachtungen fanden jeweils an insgesamt fünfzehn Tagen in einem Zeitraum von drei Wochen statt. Dabei wurden zu jeweils beiden Zeitpunkten 23 Doppelbeobachtungen durchgeführt. Im Rahmen der vorliegenden Darstellung sollen aus ökonomischen Gründen lediglich die Reliabilitäten zum zweiten Zeitpunkt näher untersucht werden.

3.2.3 Statistische Kenngrößen

Tabelle 3.1: Faustregel für die Qualität von kappa (nach Landis & Koch, 1977, S. 165, Übersetzung J.P.).

Kappa Statistik	Übereinstimmung
< 0.00	ungenügend
0.00 - 0.20	mangelhaft
0.21 - 0.40	ausreichend
0.41 - 0.60	befriedigend
0.61 - 0.80	gut
0.81 - 1.00	sehr gut

Bei der **Fehlertaxonomie** und der **Fehlerentdeckung** handelt es sich um nominalskalierte Daten **abhängiger** Urteile, bei der jeder Kodierer je Ereignis lediglich ein einziges Urteil abgibt. Die Relia-bilitätsbestimmung dieser Variablen soll anhand von Cohen's (1960) **kap-pa**[2] bzw. anhand des **kappas** für ein-zelne Kategorien nach Fleiss (1981) erfolgen. Bei der **Fehlerbewältigung** handelt es sich um nominal-skalierte Daten **unabhängiger Urteile**, da jeder Kodierer jeder Untersuchungseinheit mehrere Kategorien zuordnen darf. Die Inter-Rater-Reliabilität dieser Variablen soll anhand des dementsprechenden **kappa**-Maßes von Kraemer (1980) überprüft werden[3]. Die Reliabilitätsbestimmung der ordinal-skalierten **Fehlerkorrekturzeit** erfolgt anhand von Kendall's **tau** (Kendall, 1948) (für eine Übersicht möglicher Verfahren zur Reliabilitätsanalyse unterschiedlicher Skalenniveaus siehe Asendorpf & Wallbott, 1979; Feger, 1983).

Da die einzelnen Ereignisse aufgrund ihres unterschiedlichen Skalenniveaus und aufgrund ihrer unterschiedlichen Randbedingungen unterschiedlichen Reliabilitätsüberprüfungen unterzogen werden müssen, ist ein unmittelbarer Vergleich der resultierenden Koeffizienten nur schwer möglich. Zudem ist es nur wenig aussagekräftig, ob ein bestimmter Inter-Rater-Reliabilitäts-Koeffizient ein bestimmtes Signifikanzniveau erreicht

---

[2] In Erweiterung der prozentualen Übereinstimmung, welche die Tatsache vernachlässigt, daß nach den Zufallsgesetzen einige Beobachtungen übereinstimmen müssen, wurden im Laufe der Jahre mehrere Vorschläge zur Zufallskorrektur des Übereinstimmungsmaßes entwickelt (für eine Übersicht siehe Asendorpf & Wallbott, 1979). Cohen's (1960) kappa - "the proportion of agreement after chance agreement is removed from consideration" (Cohen, 1960, S. 40) - sollte sich im weiteren Verlauf der Entwicklung des Übereinstimmungskonzeptes bei nominalskalierten Daten als dasjenige Verfahren herausstellen, das sich mathematisch am besten begründen läßt und das für die Praxis am fruchtbarsten ist (vgl. Friede, 1981).

[3] Zu einer Diskussion zufalls- und nicht-zufallskritischer Verfahren zur Bestimmung der Inter-Rater-Reliabilität für abhängige und unabhängige Urteile auf Nominalniveau, siehe: Friede (1981).

oder nicht, da stets die Randbedingungen Berücksichtigung finden müssen, unter denen eine bestimmte Höhe der Übereinstimmung zustande kam (vgl. Cohen, 1960; Hoyt, 1941).

Allerdings liegt für **kappa** zur besseren Beurteilung eine Faustregel von Landis und Koch (1977) vor, die bei der Besprechung der entsprechenden Ergebnisse als Referenz dienen soll.

3.3 Überprüfung der Reliabilität

3.3.1 Inter-Rater-Reliabilitäten der Fehlertaxonomie

Bei der Klassifikation der auftretenden Fehlerereignisse standen den Untersuchern die fünfzehn Kategorien der handlungsorientierten Fehlertaxonomie zur Verfügung (vgl. Kap. 2). Tabelle 3.2 stellt die Reliabilitäten dar, wenn zwei Koder unabhängig voneinander einen Fehler einer dieser Kategorien zuordnen. Bei den Inter-Rater-Reliabilitäten der beiden Doppelbeobachtungsstudien handelt es sich um die Übereinstimmung von zwei unabhängigen Primär-Ratern; bei den Inter-Rater-Reliabilitäten der Hauptuntersuchung um die übereinstimmenden Kategorisierungen, die zwei unabhängige Sekundär-Rater auf Grundlage der Ereignisbeschreibungen eines Primär-Raters anfertigten.
Betrachtet man zunächst die Reliabilitäten der Fehlergesamtheit, so können sie in Anlehnung an Landis und Koch (1977) für alle drei Untersuchungen als "gut", als recht solide, bezeichnet werden. Aufgeschlüsselt nach einzelnen Fehlerklassen bedeutet dies, das die Reliabilitäten sogar in 8 Fällen als "sehr gut", in 20 Fällen als "gut", in 12 Fällen als "befriedigend" und lediglich in jeweils zwei Fällen als "ausreichend" bzw. als "mangelhaft" zu beurteilen sind.

Zusammenfassend läßt sich also festhalten, daß die handlungsorientierte Fehlertaxonomie, auch wenn das eine oder andere Ergebnis noch deutlich verbesserungswürdig ist (für eine ausführlichere Diskussion dieser Ergebnisse siehe: Prümper, in Vorb.), im großen und ganzen zufriedenstellende Reliabilitäten aufweist.

3.3.2 Inter-Rater-Reliabilitäten der Fehlerentdeckung

Bei der Klassifikation der Fehlerentdeckung standen den Untersuchern die sechs Kategorien **"Interner Zielvergleich"**, **"Planbarriere"**, **"Evidente Fehlerinformation"**, **"Forcing Function"**, **"Systemmeldung"** und **"Andere Person"** zur Verfügung (zu den theoretischen Grundlagen vgl. Kap. 4).

Für den kappa-Koeffizienten ergibt sich bei der Fehlerentdeckung für die Doppelbeobachtung I ein Gesamtwert von .46, der nach Landis und Koch (1977) als "befriedigend", als mittelmäßig, als gerade noch befriedigend bezeichnet werden kann (Fehlerentdeckung wurde bei der Doppelbeobachtung II nicht erhoben).

Tabelle 3.2: Inter-Rater-Reliabilitäten der Fehlertaxonomie.

	Untersuchung		
	Doppel I	Doppel II	Haupt
Fehlergesamtheit	.63	.66	.72
Regulationsgrundlage			
Wissensfehler	.87	.56	.73
Intellektuelle Regulationsebene			
Denkfehler	.37	.60	.78
Merk-und Vergessensfehler	.43	.62	.59
Urteilsfehler	.66	.63	.56
Ebene der flexiblen Handlungsmuster			
Gewohnheitsfehler	.59	.82	.65
Unterlassensfehler	.33	.78	.65
Erkennensfehler	.17	.49	.63
Sensumotorische Regulationsebene			
Bewegungsfehler	.79	.92	.81
Ineffizienzen			
Ineffizienz Wissen	.67	.66	.80
Ineffizienz Gewohnheit	.65	.57	.76
Funktionsprobleme			
Handlungsblockade	.00	.71	.66
Handlungswiederholung	.72	.52	.71
Handlungsunterbrechung	.49	---*	.76
Handlungsumweg	.83	.56	.82
Gruppeninteraktion	1.00	.84	.77

Anmerkung: * kein Ereignis aufgetreten

Ein Grund für die - im Vergleich zur Fehlertaxonomie - geringe Reliabilität dürfte darin zu sehen sein, daß mehrere Informationsquellen gleichzeitig auftreten können, sich die Beobachter jedoch für eine Kategorie entscheiden sollten. In manchen Fällen wurde von den Beobachtern also eine mehr oder weniger willkürliche Prioritätensetzung gefordert.

3.3.3 Inter-Rater-Reliabilitäten der Fehlerbewältigung

Bei der Klassifikation der Fehlerbewältigung standen den Untersuchern die fünf Kategorien **"eigenes Wissen"**, **"Handbuch"**, **"Kollege"**, **"Beratung"**, **"Hilfesystem oder Menü"** zur Verfügung. Da sich diese fünf Kategorien zwar logisch-begrifflich gegenseitig ausschließen, jedoch praktisch gemeinsam auftreten können, durfte hier jeder Koder jeder Untersuchungseinheit mehrere Kategorien zuordnen.

Für den kappa-Koeffizienten ergibt sich für die Fehlerbewältigung für die erste Doppelbeobachtung ein Wert von .71 und für die zweite Doppelbeobachtung ein Wert von .59 -

ein gutes bis befriedigendes Ergebnis, das nur leicht unter dem Ergebnis der Fehlertaxonomie liegt, obwohl bei der Abgabe unabhängiger Urteile, bei der mehrere Katgeorien gleichzeitig auftreten können, die Kodierleistung komplexer ist als bei der Kategorisierung abhängiger Urteile.

3.3.4 Inter-Rater-Reliabilitäten der Fehlerkorrekturzeit

In der Einleitung wurde bereits darauf hingewiesen, daß bei Betriebsuntersuchungen auf den Einsatz maschineller Aufzeichnungsgeräte weitgehend verzichtet werden muß. Dies trifft auch für die Verwendung von Zeitmessern zu. Dadurch waren wir gezwungen, die Dauer der Fehlerkorrektur zu schätzen. Als Einheiten wurden folgende Zeiten festgehalten: sofort (kodiert: 15 Sekunden), bis 2 Minuten (kodiert: 1 Minute), bis 5 Minuten (kodiert: 4 Minuten), bis 10 Minuten (kodiert: 8 Minuten), mehr als 10 Minuten (kodiert: 12 Minuten).

Für die Fehlerkorrekturzeit resultiert für die erste Doppelbeobachtung ein Kendall's tau von .69 und für die zweite Doppelbeobachtung ein tau von .56. Wenn man bedenkt, daß in diese Werte nicht nur die Reliabilität der Zeitschätzung an sich eingeht, sondern auch die Reliabilität der Bestimmung, wann ein Fehler beginnt bzw. wann er beendet ist (ein Problem, das sich auch unter der Zuhilfenahme von Stoppuhren nicht umgehen ließe), dann sind die Reliabilitäten der Fehlerkorrekturzeit zwar nicht ausgezeichnet, aber durchaus als akzeptabel zu bewerten.[4]

3.3.5 Zusammenfassung der Überprüfung der Reliabilität

Möchte man die Ergebnisse der vorliegenden Reliabilitätsstudien abschließend beurteilen, so gilt zu bedenken, daß sich eventuell ergebende hohe Übereinstimmungen umso aussagekräftiger sind, je mehr Unterscheidungen zu treffen sind und je schneller der Wechsel der Ereignisse ist (Mees, 1977b) und daß Beobachtungen im Felde immer sehr kompliziert sind (Semmer, 1984) - und dies, wie oben ausgeführt, in besonderem Maße für die Beobachtungen von Fehlern gilt. Vor diesem Hintergrund liefern die Inter-Rater-Reliabilitäten der Beobachtungsinstrumente im großen und ganzen zufriedenstellende Werte.

3.4 Überprüfung der Äquivalenz

Nach der Bestimmung der Inter-Rater-Reliabilität stellt sich nun die Frage, ob die Methode des Sekundär-Ratings, so wie sie im Rahmen der Hauptuntersuchung Anwendung

---

[4] In einem Laborexperiment (Lang, 1991) wurden neben dem Einsatz der systematischen Beobachtung zur Schätzung der Fehlerkorrekturzeit zusätzlich Keystroke-Protokolle der Zeiten mit Hilfe des Computers erfaßt. Wenn man diese beiden Methoden zur Erfassung der Fehlerkorrekturzeit miteinander vergleicht, so resultiert eine Korrelation von $r = .75$ ($N = 707$, $p < .001$).

fand, nicht nur ein ökonomisches, sondern auch zur Doppelbeobachtung ein äquivalentes Verfahren darstellt.

Diese Überprüfung der Äquivalenz kann auf unterschiedliche Weise erfolgen (vgl. Prümper, in Vorb.). An dieser Stelle sollen zwei Fragen im Vordergrund stehen.

Die erste Frage lautet, ob durch das Primär- und Sekundär-Rating äquivalente Fehlerstichproben gezogen werden oder ob die unterschiedlichen Herangehensweisen zur Folge haben, daß bestimmte Arten von Fehlern zu Lasten anderer präferiert werden. Diese Frage zielt also darauf ab, ob die einzelnen Primär-Rater ebenso häufig eine bestimmte Fehlerart identifizieren wie die Sekundär-Rater. Ist dies nicht der Fall, dann bestünde die Gefahr, daß die Forschungsergebnisse nicht durch den Forschungsgegenstand an sich, sondern durch die Art des Rating-Verfahrens bestimmt würden. Zur Beantwortung dieser Frage sollen die relativen Fehlerhäufigkeiten der Primär- und Sekundär-Ratings pro Kategorie miteinander verglichen werden.

Doch selbst wenn durch das Primär- und Sekundär-Rating äquivalente Fehlerstichproben gezogen werden, so ist damit noch nicht sichergestellt, daß beide Methoden mit äquivalenten Inter-Rater-Reliabilitäten aufwarten können. Dies wäre beispielsweise dann nicht der Fall, wenn die Sekundär-Rater zwar eine bestimmte Fehlerart annähernd ebenso häufig identifizieren wie die Primär-Rater, sie aber zu geringeren Übereinstimmungen in ihren Urteil untereinander kommen.

Die zweite Frage lautet nun, ob die Primär- und Sekundär-Ratings äquivalente Inter-Rater-Reliabilitäten erzeugen. Die Bedeutsamkeit der Beantwortung dieser Frage liegt in dem Umstand begründet, daß lediglich die übereinstimmend kategorisierten Ereignisse die Grundlage für weitere Berechnungen darstellen. Zur Beantwortung dieser Frage sollen die Übereinstimmungskoeffizienten des Primär-Ratings zu den entsprechenden Werten des Sekundär-Ratings in Beziehung gesetzt werden.

3.4.1 Methode

Zum Zweck der Äquivalenzüberprüfung wurden die Ereignisbeschreibungen der Beobachter aus Doppelbeobachtung I ein zweites Mal von zwei weiteren Untersuchern, die an dieser Untersuchung nicht beteiligt waren, unabhängig voneinander kategorisiert. Diesen beiden Sekundär-Ratern lag dabei weder das Urteil der Primär-Rater noch das ihres Counterparts vor. Insgesamt lagen damit für jeden Fehler sechs Kategorisierungen vor.

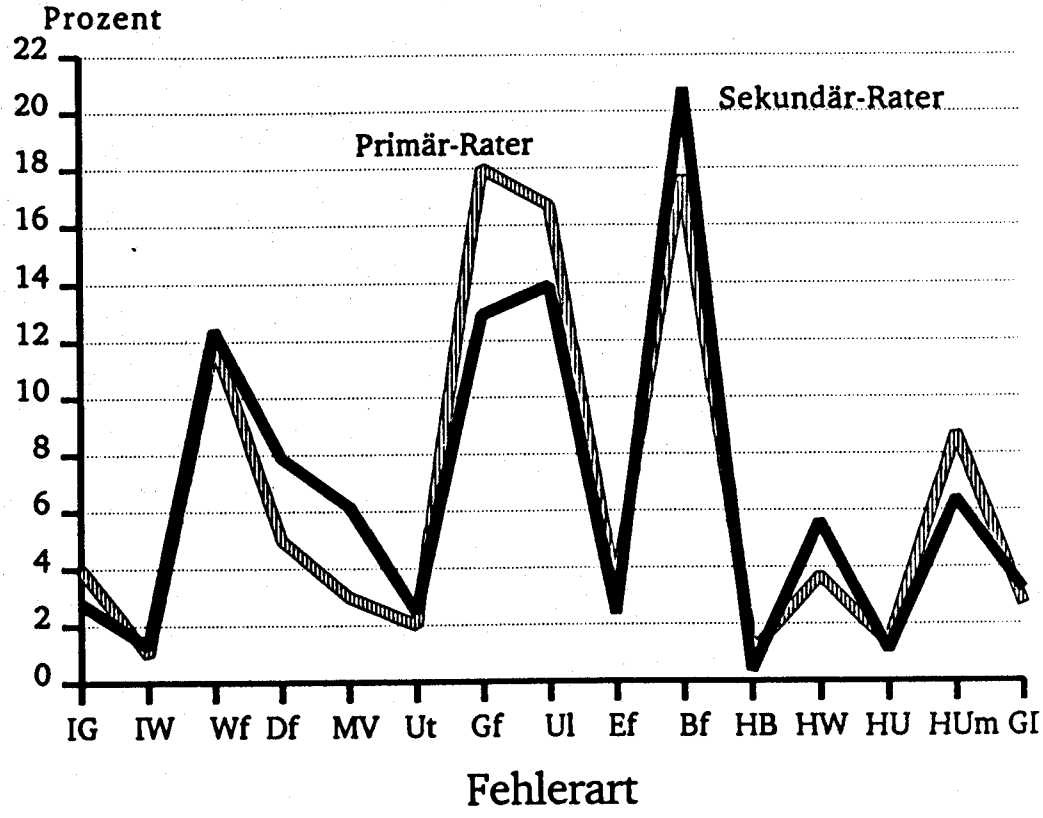
Zur Beantwortung der ersten Frage, ob sich durch das Sekundär-Rating eine äquivalente Fehlerverteilung ergibt wie durch das Primär-Rating, wurde zum einen die relative Häufigkeit der einzelnen Fehlerkategorien über die vier Kodierungen der zwei Sekundär-Rater und zum anderen über die zwei Kodierungen der beiden Primär-Rater gemittelt.

Zur Beantwortung der zweiten Frage, ob das Sekundär-Rating ein zur Doppelbeobachtung äquivalentes Verfahren zur Bestimmung der Inter-Rater-Reliabilitäten darstellt, wurden zu den entsprechenden Koeffizienten der Primär-Rater die gemittelten kappa-Werte der einzelnen Fehlerkategorien über die vier Kodierungen der zwei Sekundär-Rater berechnet (vgl. ausführlich Prümper, in Vorb.).

3.4.2 Ergebnisse

In Abbildung 3.1 werden die relativen Fehlerhäufigkeiten, die von den Primär-Ratern und von den Sekundär-Ratern registriert wurden, gegenübergestellt. Dabei zeigt sich, daß die beiden Verteilungen einen weitgehend kongruenten Verlauf aufweisen.

Korreliert man die gemittelten relativen Häufigkeiten der Primär- und Sekundär-Ratings, so resultiert ein Koeffizient von $r = .93$ ($N = 15$, $p < .001$). Auch dieses Ergebnis spricht dafür, daß durch die beiden Verfahren äquivalente Fehler registriert werden.



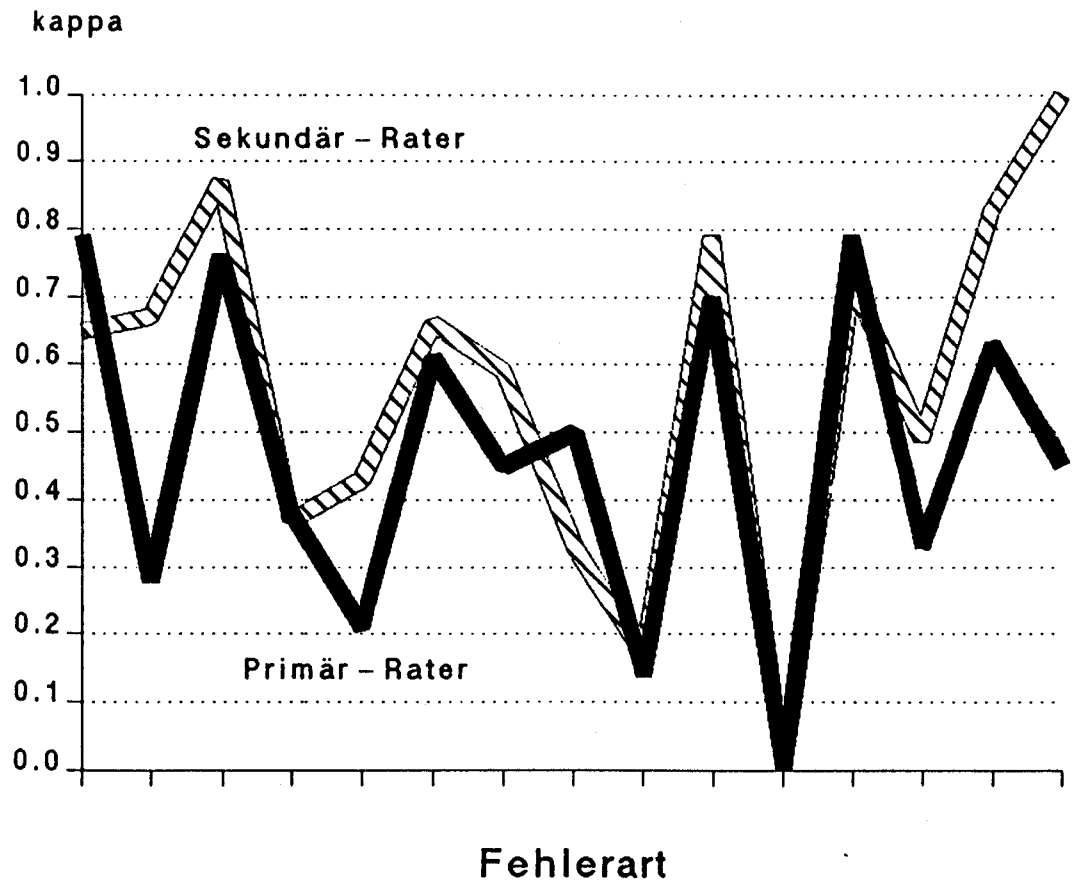
IG Ineffizienz Gewohnheit
IW Ineffizienz Wissen
Wf Wissensfehler
Df Denkfehler
MV Merk-/Vergessensfehler
Ut Urteilsfehler
Gf Gewohnheitsfehler
Ul Unterlassensfehler
Ef Erkennensfehler
Bf Bewegungsfehler
HB Handlungsblockade
HW Handlungswiederholung
HU Handlungsunterbrechung
HUm Handlungsumweg
GI Gruppeninteraktion
(vgl. die Fehlertaxonomie in Kap. 2)

Abbildung 3.1: Vergleich der relativen Fehlerhäufigkeiten der Primär- und Sekundär-Ratings.

In Abbildung 3.2 werden die Inter-Rater-Reliabilitäten der Primär- und Sekundär-Ratings verglichen. Bezüglich des Gesamtwertes resultiert für die Sekundär-Rater ein kappa von .55, der damit niedriger liegt als der entsprechende Koeffizient der Primär-Rater (.63, vgl. Tabelle 3.2). Jedoch zeigt sich auch hier, daß die beiden Kurven im großen und ganzen

einen ähnlichen Verlauf aufweisen. Größere Abweichungen ergeben sich lediglich bei ineffizientem Verhalten aufgrund fehlenden Wissens und bei den Interaktionsproblemen (zu einer ausführlicheren Diskussion der Schwierigkeiten, auf die Primär-Rater bzw. Sekundär-Rater bei ihren Urteilen stoßen siehe Prümper, in Vorb.).

Eine Korrelation zwischen den gemittelten kategorialen kappa-Koeffizienten der Primär-Rater und den kategorialen kappa-Koeffizienten der Sekundär-Rater von r = .74 (N=15, p<.001) bestätigt diesen Eindruck. Auch dieses Ergebnis spricht dafür, daß die Primär- und Sekundär-Rater zu äquivalenten Übereinstimmungen bei der Beurteilung der Fehler gelangen.



IG Ineffizienz Gewohnheit
Df Denkfehler
Gf Gewohnheitsfehler
Bf Bewegungsfehler
HU Handlungsunterbrechung
(vgl. die Fehlertaxonomie in Kap. 2)

IW Ineffizienz Wissen
MV Merk-/Vergessensfehler
Ul Unterlassensfehler
HB Handlungsblockade
HUm Handlungsumweg

Wf Wissensfehler
Ut Urteilsfehler
Ef Erkennensfehler
HW Handlungswiederholung
GI Gruppeninteraktion

Abbildung 3.2: Vergleich der Inter-Rater-Reliabilitäten der Primär- und Sekundär-Ratings.

3.4.3 Zusammenfassung der Überprüfung der Äquivalenz

Zusammenfassend läßt sich festhalten, daß die Sekundär-Rater sowohl Fehlerereignisse registrieren, die zu denen der Primär-Rater äquivalent sind, als auch, daß sie in ihren Beurteilungen zu äquivalenten Übereinstimmungen gelangen. Somit kann davon ausgegangen werden, daß das im Rahmen der Hauptuntersuchung angewendete Sekundär-Rating ein durchaus äquivalentes Verfahren zur Doppelbeobachtung darstellt.

Allerdings liegt das Gesamt-kappa der Sekundär-Rater unter dem der Primär-Rater. Offensichtlich führt also die Methode des Sekundär-Ratings eher zu einer Unterschätzung der Inter-Rater-Reliabilitäten von Doppelbeobachtungen. Damit ließe sich auch für die Hauptuntersuchung vermuten, daß auch hier, unter der Bedingung tatsächlicher Doppelbeobachtungen, höhere Inter-Rater-Reliabilitäten zu erwarten gewesen wären.

Dieser niedrigere kappa-Wert des Sekundär-Ratings könnte darin begründet sein, daß die Sekundär-Rater aufgrund der Nicht-Teilnahme am Fehlergeschehen über weniger Informationen als die Primär-Rater zur Kategorisierung verfügen oder - anders ausgedrückt - daß es den Primär-Ratern nicht durchgängig gelungen ist, Ereignisprotokolle von ausreichend großer Präzision und Anschaulichkeit anzufertigen.

3.5 Zusammenfassung und Ausblick

Im ersten Teil des vorliegenden Beitrags wurden die im Rahmen des Forschungsprojektes FAUST zum Einsatz gekommenen Beobachtungsinstrumente auf ihre Inter-Rater-Reliabilitäten hin untersucht. In Anbetracht der eingangs diskutierten Schwierigkeiten von Fehlerbeobachtungen im Feld, die von vornherein lediglich geringe Reliabilitäten erwarten ließen, darf man also mit den Ergebnissen der vorliegenden Analysen zufrieden sein. Allerdings vermitteln diese Reliabilitätsuntersuchungen lediglich einen ersten, deskriptiven Einblick in die Beurteilung von Fehlern und gehen kaum über den Status eines Rechenschaftsberichtes hinaus. Weiterführende Analysen, die beispielsweise der Frage nachgehen, warum manche Fehler hohe und manche niedrige Reliabilitäten aufweisen oder ob bestimmte Fehler zu typischen Verwechlungen mit anderen führen, stehen noch aus.

Im Anschluß an die Reliabilitätsüberprüfungen wurde das in der Hauptuntersuchung angewendete Verfahren des Sekundär-Ratings dahingehend überprüft, ob es eine äquivalente Methode zu dem Primär-Rating der Doppelbeobachtung darstellt. Dies geschah sowohl bezüglich der Frage, ob das Sekundär-Rating eine äquivalente Fehlerverteilung zum Primär-Rating aufzeigt als auch bezüglich der Äquivalenz der Inter-Rater-Reliabilitäten. In beiden Fällen konnten zufriedenstellende Ergebnisse berichtet werden. Dabei zeigte sich, daß das Sekundär-Rating eher zu einer Unterschätzung der durch die Primär-Ratings erzielten Inter-Rater-Reliabilitäten führt. Damit rückt die Frage in den Blickpunkt des Interesses, ob es bestimmte Ereignisse gibt, bei denen die Beobachter des Fehlergeschehens mehr Schwierigkeiten haben als bei anderen, ihre Beschreibung so klar und deutlich zu formulieren, daß die Leser der Fehlerprotokolle sie eindeutig zuorden können. Auch für die Beantwortung dieser Frage müssen weitere Analysen durchgeführt werden.

Allerdings ist durch die Beantwortung der beiden Fragen nach der Äquivalenz der beiden Rating-Verfahren noch nicht sichergestellt, daß von den Primär- und Sekundär-Ratern

auch **identische** Fehler kategorisiert werden. Um die Frage der Identität zu beantworten, müssen in einer weiteren Analyse die Übereinstimmungen zwischen den Primär- und Sekundär-Ratings bestimmt werden. Eine zufriedenstellende Übereinstimmung hätte den praktischen Wert, daß sich neben der üblichen Methode der Doppelbeobachtung, bei der (mindestens) zwei Primär-Rater im Feld tätig sein müssen und der im Rahmen unseres Projektes vorgeschlagenen Methode des Sekundär-Ratings, bei der ein Primär-Rater und zwei Sekundär-Rater beschäftigt sind, eine weitere, noch ökonomischere Alternative anbietet. Dann würde nämlich bereits schon die Übereinkunft zwischen einem Primär-Rater und einem Sekundär-Rater eine zuverlässige Basis für weitere Berechnungen liefern.

Die handlungsorientierte Fehlertaxonomie wurde in erster Linie zu dem Zweck entworfen, ein in der Praxis einsetzbares Fehleranalyseinstrument zur Verbesserung von Software und Training zu sein. Damit werden auch Fehleranalytiker auf den Plan gerufen, die nicht an der theoretischen Konzeptualisierung der Taxonomie beteiligt waren und die auch nicht unbedingt in die handlungstheoretischen Tiefen vorzudringen wünschen. Gerade für diese Personengruppe dürften diese weiterführenden Analysen von besonderem Gewinn sein, da sie Aufschluß über die Fallen geben, die bei der Beschreibung und Taxonomierung von Fehlern lauern.

Ein erster Schritt in diese Richtung wurde bereits mit der Anwendung und Äquivalenzüberprüfung der Methode des Sekundär-Ratings schriftlicher Ereignisprotokolle gemacht. Es handelt sich dabei um ein einfaches und ökonomisches Verfahren zur Bestimmung der Inter-Rater-Reliabilität, das insbesondere den Anforderungen und Restriktionen betrieblicher Studien entgegenkommt.

Quelle:
Prümper, J. (1991). Die Inter-Rater-Reliabilität von Fehlerbeobachtungen im Feld. In M. Frese & D. Zapf (Hrsg.), *Fehler bei der Areit mit dem Computer - Ergebnisse von Beobachtungen und Befragungen im Bürobereich* (S. 47-59). Bern: Huber.