

1. Kapitel

Planung und Bewertung von wissenschaftlichen Beobachtungen

*Hubert Feger**

1. Übersicht und Systematik

Alle Aussagen über den jeweiligen Gegenstandsbereich einer empirischen Wissenschaft sind direkt oder indirekt auf Beobachtungen zurückzuführen. In einer (nichtapparativen) Beobachtung stellt sich durch und als unmittelbare Anschauung die Beziehung zwischen Beobachter und Beobachtungsgegenstand her. Beobachtung ist ein Prozeß, in dem Beziehungen zwischen Gegebenheiten hergestellt werden. Welcher Art die zu erfassenden Gegebenheiten und Relationen ihrem Inhalt nach sind, bestimmen die theoretischen und praktischen Fragestellungen; aus dem Gehalt der Theorie wäre abzuleiten, wie Beobachtungen als Beziehungen zwischen Elementen in formaler Hinsicht aufgefaßt werden müssen. Daraus ergeben sich Konsequenzen für die Datenanalyse. Die Prüfung von Aussagen über einen Gegenstandsbereich einer empirischen Wissenschaft wird als Prüfung der Übereinstimmung von aus der Theorie abgeleitetem Sachverhalt mit beobachtetem Sachverhalt angelegt sein, wobei der Prüfung logischer Merkmale der Aussagen selbst, etwa ihre Widerspruchsfreiheit der Status einer notwendigen Voraussetzung zukommt. Letzteres im allgemeinen systematisch zu analysieren ist eine der Aufgaben der Wissenschaftstheorie und der Logik.

Dieses Kapitel befaßt sich nicht mit einer psychologischen Theorie des Beobachtens durch Menschen als Beobachtern, weder mit den Prozessen der Wahrnehmung, Beurteilung und des Gedächtnisses, die Beobachtungen ermöglichen und deren Form und Inhalt beeinflussen, noch mit einer allgemeinen und differentiellen Theorie des Beobachters. Einiges zu dieser Thematik findet sich im Kapitel über Verhaltensbeobachtung und Erlebnisbeschreibung (Feger & Graumann in diesem Band). Auch die Auswertung von Beobachtungen ist

* Dank für wesentliche Hinweise schulde ich C. F. Graumann, I. Borg und K. Westhoff.

im Gegensatz zu vielen Kapiteln nicht Gegenstand dieser Darstellung. Vielmehr schildern und diskutieren wir allgemeine Prinzipien der Planung und Bewertung von Beobachtungen als wissenschaftlichen. Die Planung bezieht sich auf alle Fragen, welche Beobachtungen anzustellen und wie sie durchzuführen seien. Die Bewertung umfaßt alle Fragen nach den Kriterien, die eine Beobachtung als wissenschaftliche erfüllen muß, und wie diese Kriterien zu prüfen sind. Empirische Wissenschaften, und in ihnen verschiedene Forschungsgebiete, unterscheiden sich in dem Ausmaß, in dem der Beobachtungsprozeß standardisiert und instrumentalisiert ist. Der Ersatz des Beobachters durch Apparate beseitigt ganz oder teilweise die Probleme der Subjektivität bei der Auswahl und Bewertung des konkret Beobachteten, nicht jedoch die Probleme der Planung und Bewertung sowie der datentheoretischen Interpretation von Beobachtungen. Nicht behandelt werden deshalb Probleme, die sich durch und bei Gebrauch technischer Apparatur, insbesondere Tonband- und Film- oder Videogeräten ergeben (z. B. Clarke & Ellgring, 1978; Longabaugh, 1980), ebenfalls nicht Fragen der technischen Datenverarbeitung, Ablochung und Speicherung. Zur allgemeinen Methodenlehre gehören definitionsgemäß auch nicht Probleme des Beobachtens, die sich aus der speziellen Natur des Beobachteten ergeben, z.B. Beobachtung von Säuglingen, Tieren in freier Wildbahn etc. Ferner würde die Grenze zur Systematik des Faches überschritten, wenn hier jene theoretischen Ansätze besprochen würden, deren empirische Fundierung zumindest gegenwärtig stark auf Beobachtung angewiesen scheint, wie Ethologie oder Humanökologie. Schließlich haben wir uns bemüht, Überschneidungen mit dem in diesem Band folgenden Kapitel und dem Artikel von v. Cranach und Frenz (1969) zu vermeiden.

Ohne auch nur die klassischen oder jeweils neuesten Arbeiten vollständig auflisten zu können, erwähnen wir vorwiegend deutschsprachige Literatur, die eine Einführung und Übersicht ermöglicht: Peak (1953), König (1962), Graumann (1966), Weick (1968), Jahoda et al. (1968), Grüner (1974), Hutt & Hurt (1974), Faßnacht (1979); insbesondere für die Erfassung nonverbaler Verhaltensweisen Scherer (1974), für verbale Manz (1974), zum ethologischen Ansatz McGrew (1972), für Verhalten in „naturalistic settings“ des Kulturvergleichs: Longabaugh (1980), mit Beispielen aus der pädagogischen Psychologie: Medley & Mitzel (1963), für diese in der Entwicklungspsychologie seit langem gepflegte Methodik: Thomae (1959), Wright (1960), speziell die Kleinkindforschung: Simons & Papousek (1978); mit sozialpsychologischen Schwerpunkten: Heyns & Zander (1953), v. Cranach & Frenz (1969), Duncan & Fiske (1977); aus diagnostischer Perspektive: Hasemann (1964), mit Querverbindung zu Verhaltensmodifikation: Keut & Foster (1977), Mees & Selg (1977); s. auch Feger & Graumann in diesem Band.

2. Arten von Beobachtungen

2.1 Allgemeine Übersicht

Bevor wir auf Planung und Bewertung für Beobachtungen allgemein eingehen, geben wir einen Überblick über verschiedene Arten von Beobachtungen, die in der Psychologie und Nachbarwissenschaften eine Rolle spielen. Sie werden unterschieden nach den Bedingungen, unter denen sie zustande kommen, den Verfahren, wie sie gewonnen werden, und nach Besonderheiten des Beobachtungsgegenstandes.

Graumann (1966, S. 86) akzentuiert, wie sich Beobachtung, auch die noch nicht wissenschaftliche, von Wahrnehmung abhebt: „Die absichtliche, aufmerksam-selektive Art des Wahrnehmens, die ganz bestimmte Aspekte auf Kosten der Bestimmtheit von anderen beachtet, nennen wir Beobachtung. Gegenüber dem üblichen Wahrnehmen ist das beobachtende Verhalten planvoller, selektiver, von einer Suchhaltung bestimmt und von vorneherein auf die Möglichkeit der Auswertung des Beobachteten im Sinne der übergreifenden Absicht gerichtet.“ Wenn die übergreifende Absicht ist, eine wissenschaftliche Annahme zu prüfen, und wenn sie in Planung und Bewertung bestimmten Kriterien genügt, geht die vorwissenschaftliche in die wissenschaftliche Beobachtung über.

Alltägliche und wissenschaftliche Beobachtung unterscheiden sich nicht dadurch, wie Information gewonnen wird, nicht durch die Art der Prozesse, die im Beobachter ablaufen, sondern durch die Ziele, derentwegen die Beobachtungen angestellt werden, und durch die Umstände, die wegen dieser Ziele aufgesucht oder hergestellt werden. Beobachtungen lassen sich nach dem Verfahren, wie sie zustande kommen, und dabei wieder nach verschiedenen Gesichtspunkten klassifizieren. Uns erscheint ein Schema von Graumann (1966) das in Tab. 1 wiedergegeben ist, sehr übersichtlich. Das Schema ist lediglich für die kontrollierte direkte Verhaltensbeobachtung ausgeführt. Soweit möglich sollte man sich diese Klassifikationskriterien nicht dichotom, sondern als Pole *eines* Kontinuums vorstellen. Die erste Klassifikation nach nichtkontrollierten und *kontrollierten* Beobachtungen, d.h. solchen, bei denen die Bedingungen, unter denen sie zustande kamen, bekannt sind, ist für manche Autoren gleichbedeutend mit der Unterscheidung zwischen nichtwissenschaftlicher und wissenschaftlicher Beobachtung, doch sollte nach Graumann gerade diese Klassifikation nicht als Dichotomie angesehen werden. Bezeichnet man eine Beobachtung als direkt, so kann dies in der Literatur dreierlei bedeuten: (1) Zwischen Beobachter und Beobachtetem steht kein Hilfsmittel, kein Apparat, Test o.ä. Graumann spricht dann - als fünftem Kriterium - von *vermittelter* vs. *unvermittelter* Beobachtung. Der Einsatz einer Blickbewegungskamera führt also zu direkten, vermittelten Beobachtungen. (2) Zwischen Beobach-

tung einerseits, Beurteilung sowie Registrierung andererseits liegt kein größerer Zeitraum; die retrospektive Rekonstruktion bei Fallstudien wäre demnach als indirekte Beobachtung zu klassifizieren. Graumann jedoch bezeichnet eine Verhaltensbeobachtung dann als indirekt, wenn (3) sie sich nicht auf das Verhalten selbst, sondern auf dessen Spuren, Auswirkungen und Objektivationen richtet. Zu diesen indirekten Verfahren gehört z.B. die systematische Inhaltsanalyse.

Die *teilnehmende* Beobachtung, in der ein Beobachter für die Beobachteten einen erkennbaren Teil der Beobachtungssituation ausmacht, wird im folgenden Abschnitt behandelt. Als *unwissentlich* bezeichnet man Beobachtungsverfahren, bei denen gegenüber den Beobachteten die Tatsache, daß sie beobachtet werden, so weit wie möglich verborgen oder kaschiert wird.

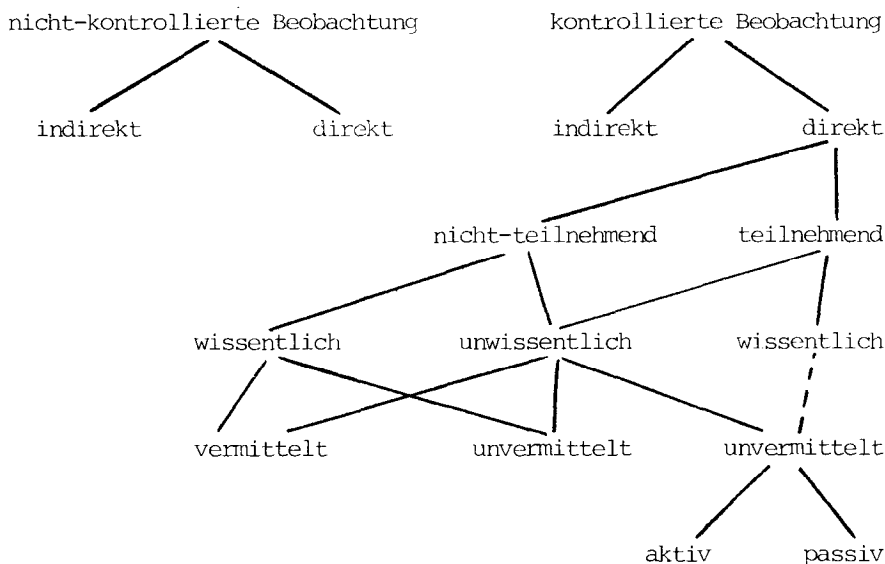


Abb. 1: Schema der Methoden der Verhaltensbeobachtung (nach Graumann, 1966)

Die herkömmliche Unterscheidung zwischen *freier* und *systematischer* Beobachtung greifen v. Cranach & Frenz (1969, S. 269) auf. Während freie Beobachtungen „ohne methodische Einschränkung“ vorgenommen werde, geschehe bei der systematischen Beobachtung die „Anwendung von Beobachtungssystemen als Meßverfahren, welche die quantitative Erfassung operational definierter Variablen mit angebbarer Objektivität, Zuverlässigkeit und Gültigkeit erlauben“.

2.2 Teilnehmende Beobachtung

Da es für Verhaltensbeobachtung wichtig ist anzugeben, auf wen das Verhalten des Beobachteten gerichtet ist, muß die Rolle des Beobachters in der Beobachtungssituation untersucht werden, insofern und insbesondere als er auch für die beobachteten Personen Teil ihrer Situation sein kann. Diese Problematik ist bei der Entwicklung der *Methode der teilnehmenden Beobachtung* (participant observation) durchweg gesehen worden (Whyte 1953 als Klassiker, Schwartz & Schwartz 1955, Kluckhohn 1956, McCall & Simmons 1969, Friedrichs & Lüdtke 1973). Fälle, in denen schon die bloße Anwesenheit eines Beobachters zu anderem Verhalten führte als in Situationen ohne einen Beobachter, sind schon früh berichtet worden, etwa von Polansky et al. (1949) bei verhaltensauffälligen Kindern. Die Frage der Reaktivität ist in jüngster Zeit ausführlicher bei der Erforschung von Verhaltenseinschätzungen untersucht worden und soll im Kapitel von Feger und Graumann dargestellt werden.

Teilnehmende Beobachtung verlangt in der Regel jedoch mehr als bloße Anwesenheit, oft volle Mitgliedschaft des Beobachters in der untersuchten sozialen Gruppierung. Dies wird verständlich, wenn man sich die hauptsächlichsten Anwendungsbereiche dieser Methodik vergegenwärtigt: primitive Kulturen, Subkulturen, Kommunen, Krankenhäuser, Gefängnisse, Fabriken, Bürokratien, Militär, Kulte, Familien, Verbrecherbanden. Grümer (1974) verweist in Anlehnung an Kunz auf drei nach der Rolle des Sprachsystems unterschiedene Forschungssituationen, in denen die teilnehmende Beobachtung eingesetzt werden könne: Erstens Situationen, in denen Beobachter und Beobachtete über ein unterschiedliches Sprachsystem verfügten, wie etwa in der Verhaltensforschung an Tieren; zweitens Situationen, in denen Beobachter und Beobachtete zwar ein unterschiedliches Sprachsystem haben, jedoch wenigstens teilweise lernen können, eine gemeinsame Sprache zu benutzen, beispielsweise bei ethnologischen Untersuchungen in fremdsprachigen Kulturen; drittens Forschungssituationen, in denen zwar ein gemeinsames Sprachsystem besteht, jedoch Unterschiede und Abweichungen vorkommen, die das Verständnis eines Beobachters für eine Situation beeinträchtigen können, z.B. bei Untersuchungen mit Kleinkindern, Geisteskranken und kulturellen Subgruppen.

McCall & Simon (1969, S. 3) beschreiben, was unter teilnehmender Beobachtung bei ihrer Anwendung faktisch verstanden wird:

„... is most sensibly regarded, operationally, as the blend of methods and techniques that is characteristically employed in studies of social situations or complex social organizations of all sorts. These are studies that involve repeated, genuine social interaction on the scene with the subjects themselves as a part of the data-gathering process.“

An dieser Umschreibung sind uns zwei Punkte wichtig, auf die wir weiter eingehen wollen, zum einen der Hinweis, daß teilnehmende Beobachtung typischerweise eine Kombination von Methoden darstellt, und daß sie den Beobachter in eine soziale Interaktion einbindet. Das Wesen der Methode (und die Quelle ihrer Schwierigkeiten) besteht darin, daß der Beobachter im sozialen Feld eine bestimmte Rolle zu spielen hat, und somit nicht nur eine ohnehin oft schwierig als solche zu definierende „repräsentative“ Auswahl des Feldes den Wert der Beobachtungen bestimmt, sondern auch das Geschick, mit dem der Beobachter eine ihm angemessen erscheinende Rolle wählt und diese spielt. Die Rolle muß vom Beobachter so übernommen werden, oder es muß von denen, die solche Rollen gewöhnlich spielen, so ein geeigneter Beobachter rekrutiert werden, daß das soziale Feld nicht verfälscht wird. Ob dies gelungen ist, kann man bisweilen aus den Korrelationen mit Daten erschließen, die mit Hilfe anderer Verfahren, insbesondere durch unwissentliche Beobachtung gewonnen wurde, oder - für den typischen Anwendungsbereich realistischer - aus Übereinstimmung von eigenen Beobachtungen mit den Berichten von Informanten. (Zu spezifischen Problemen z.B. der Informantenauswahl, der Generalisierbarkeit der Befunde, möglicher Veränderung des Kategoriensystems des Beobachters bei langdauernder Interaktion und der Prüfung der internen Validität siehe insb. McCall & Simons und die dort abgedruckten Artikel).

3. Die Planung von Beobachtungen

Bei der Planung muß zunächst entschieden werden, was zu beobachten ist. Für diese Entscheidung muß das Universum der Beobachtungen definiert, die Einheit der Analyse bestimmt und das Kategoriensystem entwickelt werden. Danach sucht man Antworten auf die Frage, welche der unter diesen Vorgaben meist zahlreichen möglichen Beobachtungen tatsächlich realisiert, wie also eine Auswahl aus den Universen von Personen, Situationen, Zeitpunkten etc. getroffen werden soll.

3.1 Das Universum von Beobachtungen

Um die für die folgenden Darlegungen maßgebliche *Facettentheorie* (Guttman 1959, 1971; im folgenden nach Borg, 1977) einzuführen, skizzieren wir zunächst eine Studie von Guttman & Guttman (1976), die sich von den meisten Anwendungen der Facettentheorie dadurch unterscheidet, daß sie sich nicht auf die Konstruktion von Test- oder Fragebogenitems und der Erklärung ihrer korrelativen Zusammenhänge bezieht, und daß sie eine Sekundäranalyse bereits vorliegender Beobachtungen in mehreren Experimenten zu Streßindikatoren bei Mäusen darstellt. Facettentheoretische Analysen von Beobachtungen

finden wir auch bei Canter (1977, 1977b) im Bereich der Architektur- und Umweltpsychologie.

Bei der Durchsicht einer Reihe von Untersuchungen zu emotionalem Verhalten von Mäusen stellten Guttman & Guttman fest, daß die Tiere unter verschiedenartigen Bedingungen beobachtet worden waren. Diese verschiedenen experimentellen Situationen wie „offener Raum“, „Lauftrad“, Schwimmlabyrinth“ usw. stellen verschiedene Elemente einer *Facette* „experimentelle Situationen“ dar, wobei von der Facette vermutet wird, daß ihre Elemente zu Streß in unterschiedlichem Ausmaß führen. Die Beobachtungshinsicht in der untersuchten Forschung waren Maße der Frequenz, der Latenz und der Dauer. Da sich nach Ansicht von Guttman & Guttman das Ausmaß des Streß in diesen „zeitlichen Merkmalen“ unterschiedlich zeigen kann, stellen diese Merkmale eine weitere Facette dar. Schließlich zeige sich Streß in Verhaltensweisen, die entweder der willkürlichen Kontrolle unterliegen, wie z. B. Dauer des Verbleibens im offenen Raum, oder vom autonomen System gesteuert werden, z.B. Harn lassen und Koten. Die Art der Kontrolle stellt somit eine dritte Facette dar. Variation innerhalb einer Facette könne zu hohen oder niedrigen Ausprägungen des Streß führen; diese abhängige Variable stellt die durch die Facetten zu erklärenden Beobachtungen dar, Beobachtungen, an denen hier das unterschiedliche Ausmaß an Streß interessiert.

Verallgemeinert man dieses Beispiel zu dem „general paradigm of empirical research in the social sciences“ (Borg, 1979, S. 65), dann sieht man, daß drei Komponenten analytisch unterschieden werden: Eine Population P, die mit einer Menge Reize S konfrontiert wird und auf diese mit den Reaktionen R antwortet. P stellt in der Methodik der wissenschaftlichen Beobachtung die zu beobachtenden Menschen und Tiere dar, S die Variablen der Beobachtungssituation, R die für die theoretische oder praktische Fragestellung interessierenden Verhaltensweisen. Die Beziehung dieser drei Komponenten zueinander läßt sich als *Abbildungssatz* (mapping sentence) schreiben:

$$P \times S \rightarrow R$$

Das kartesische Produkt $P \times S$ stellt den Definitionsbereich (domain), R stellt den Bildbereich (range) dar; der Pfeil symbolisiert die Abbildung von $P \times S$ in R und kann als „ist verknüpft mit“ gelesen werden. Die Festlegung von Definitions- und Bildbereich stellt das *Universum von Beobachtungen* (Universe of observations) her. Eine konkrete Beobachtung besteht dann - wenn man sie von ihrer Leistung her charakterisiert - in der Zuordnung von Elementen des Definitionsbereiches zum Bildbereich.

Im einfachsten Fall bestehen P, S und R jeweils aus nur einer Menge (Facette), jedoch kann jede Komponente selbst ein kartesisches Produkt sein. Im Beispiel lautet der Abbildungssatz:

„Das Ausmaß, in dem Tier (X) Streß zeigt in

den zeitlichen Merkmalen	A:	<i>zeitliche Merkmale</i>
	$\left\{ \begin{array}{l} a_1 \\ a_2 \\ a_3 \end{array} \right\}$	$\left\{ \begin{array}{l} \text{Frequenz} \\ \text{Latenz} \\ \text{Dauer} \end{array} \right\}$
	B:	<i>Art der Kontrolle</i>
durch Auftreten von	$\left\{ \begin{array}{l} b_1 \\ b_2 \end{array} \right\}$	$\left\{ \begin{array}{l} \text{autonomen} \\ \text{willkürlichen} \end{array} \right\}$
Verhalten in einer streßhaften (neuen)	C:	<i>Experimentellen Situation</i>
Situation im	$\left\{ \begin{array}{l} c_1 \\ c_2 \\ c_3 \\ \vdots \end{array} \right\}$	$\left\{ \begin{array}{l} \text{Labyrinth} \\ \text{offenen Raum} \\ \text{Röhre} \\ \text{etc.} \end{array} \right\}$
$\longrightarrow$	$\left\{ \begin{array}{l} \text{R} \\ \text{hohen} \\ \dots \\ \text{niedrigen} \end{array} \right\}$	Anzeichen von Streß.“

In diesem Beispiel hat S drei Facetten: A, B und C. P, die Variable „Tier (X)“, ist nicht weiter ausgeführt. Dies hätte - wären die Forscher daran interessiert gewesen und hätten entsprechende Arbeiten für die Sekundäranalyse vorgelegen - leicht geschehen können, beispielsweise indem eine Facette verschiedene Zuchtstämme von Ratten unterscheidet, die auf Streß unterschiedlich reagieren.

Ein wesentlicher Gewinn bei der Anwendung der Facettentheorie zur Planung und Analyse wissenschaftlicher Beobachtungen kann in der hierbei geforderten Definition der Facetten bestehen, weil so die tatsächlich berücksichtigte oder mögliche Gesamtheit von Beobachtungen angegeben wird. Daraus können nicht nur Vorschriften für das Ziehen der Stichproben (z.B. für jede Facettenkombination mindestens eine Beobachtung) abgeleitet und begründet werden, vielmehr wird oft erst auf diese Weise eine genaue Formulierung der Fragestellung erzwungen und der Bereich der Verallgemeinerbarkeit festgelegt. Letzteres behandelt die später besprochene Generalisierbarkeitstheorie. Die Prüfung der konvergierenden und diskriminierenden Validität, die später ebenfalls dargestellt wird, stellt facettentheoretisch die Analyse der Facetten „Inhalte“ und „Methoden“ der untersuchten Variablen dar. Ferner kann man bestimmte Prinzipien heranziehen, um die Struktur im Bildbereich aus der Struktur des Definitionsbereiches vorherzusagen. In der erwähnten Reanalyse konnten Guttman & Guttman beispielsweise zeigen, daß die Höhe der Korre-

lationen zwischen den verschiedenen Streßindikatoren davon abhängt, hinsichtlich wievieler und welcher Facetten die Beobachtungen Gemeinsamkeiten aufwiesen. Diese Struktur der korrelativen Beziehungen zwischen den verschiedenen Indizes, dargestellt mit Hilfe einer multidimensionalen Skalierung, erwies sich außerdem als über die verschiedenen Experimente hinweg vergleichbar.

Wir besprechen nun einige in Psychologie und Nachbarwissenschaften häufige spezielle Fälle der Konkretisierung der Mengen P, S und R. Bei der facetten-theoretischen Betrachtung von Tests, Einstellungs- und Persönlichkeitsfragen-bogen werden die Verhaltensrealisationen an verschiedenen Personen betrachtet, die nach demographischen oder zuvor erhobenen differentialpsychologischen Merkmalen geordnet sein können oder lediglich durch ihre im Namen festgehaltene Identität als unterscheidbare Elemente dieser einen oder aller personenbeschreibenden Facetten festgestellt werden. S besteht aus Items, die beispielsweise in einem Intelligenztest so in mehrere Facetten geordnet werden könnten, daß der Guilfordsche Würfel Pate steht. Beobachtet wird, wie eine Person auf ein Item reagiert, dabei können im Bildbereich z.B. bei Leistungstests sowohl Richtigkeit als auch Schnelligkeit als auch Qualitätsmerkmale der Antwort spezifiziert werden.

Eine Facette kann auch nur ein Element aufweisen. Dies wird in der Unterscheidung verschiedener Strategien zum Ziehen von Verhaltensstichproben deutlich (s.u.). Wenn der Merkmalsträger zugleich der Beobachter ist, dann haben wir den Fall der *Selbstbeobachtung* vor uns. Der Beobachter beschreibt dabei, sei es gegenüber sich selbst, oft gegenüber anderen, dem Untersucher etwa, sein Erleben und Verhalten, z.B. seine Reaktionen auf verschiedene Farben. Im Abbildungssatz können dabei durchaus mehrere Personen als Elemente einer Facette vorgesehen sein, etwa vollständig (completely crossed) oder nur teilweise (nested) verschiedenen Untersuchern zugeordnet. Wesentlich ist, daß bei dieser Beobachtungsart die Basis für über Personen verallgemeinernde Aussagen in Äquivalenzannahmen bestehen muß, etwa der Art: Zwei Personen meinen das Gleiche, wenn sie beim Vergleich zweier beleuchteter Flächen eine als „heller“ bezeichnen. Die Menge der Merkmalsträger kann selbst als kartesisches Produkt $P \times P$ geschrieben werden, wenn beispielsweise die (*sozialen*) *Beziehungen* zwischen Personen, Tieren, Gruppen etc. beobachtet werden sollen. Bei Plänen mit wiederholten Beobachtungen kann man in den Definitionsbereich die Facette „*Zeitpunkt*“ aufnehmen; Levy & Guttman (1975) berichten eine facettentheoretisch angelegte Untersuchung mit fünf wiederholten Messungen.

Beobachtungspläne unterscheiden sich unter anderem danach, wie expliziert und vorstrukturiert Definitions- und Bildbereich durch vorausgegangene wissenschaftliche Arbeit sind. Im einen Extremfall, in der Regel bei Datengewinnung in einem *Experiment*, ist vor Beginn der Beobachtung festgelegt, wie alle

Facetten definiert sind, d.h. welche Elemente sie enthalten oder enthalten können. Im anderen Extrem, oft in der *Feldbeobachtung*, besteht bisweilen primär die wissenschaftliche Leistung darin, überhaupt „Beobachtenswertes“ zu finden, d.h. hier: die Facetten zu definieren und einander zuzuordnen. Noch einige Querverweise, um zu erläutern, in welcher Hinsicht wir das facettentheoretische Vorgehen für fundamental halten: Einige Arbeitsrichtungen in der Psychologie befassen sich damit zu bestimmen, mit welchen und wie vorstrukturierten Mengen von Kategorien alltägliche und wissenschaftliche Beobachtungen angestellt werden. Forschungen zur „impliziten Persönlichkeitstheorie“ ermitteln u.a., welche Elemente der Definitionsbereich bei der Beschreibung von Personen enthält, und wie die multivariate Struktur dieses Bereiches sich aus der Kombination der Facetten ergibt. Nosologische Klassifikationssysteme der psychologischen und psychiatrischen Diagnostik sollten m.E. als Abbildungssätze geschrieben werden, so daß Symptomkorrelationen erklärbar erscheinen. Einstellungstheorien (z.B. Fishbein & Ajzen 1975, Feger 1979) spezifizieren verschiedene Mengen von Komponenten und deren Bewertung, wobei z.T. die Vpn selbst die Elemente der Menge generieren.

3.2 Bestimmen der Beobachtungseinheit

Die Genauigkeit, mit der die Komponenten im Abbildungssatz festgelegt werden können, ist nicht unbegrenzt. Oft tritt an die Stelle einer Menge in S ein ausgearbeitetes Kategoriensystem. In ihm, in der Instruktion und in der Beobachtungsschulung versucht der Forscher, die Beobachtungseinheit so präzise wie möglich festzulegen. Dabei tritt zu der inhaltlichen Bestimmung dessen, was beobachtet werden soll - die meistens im Kategoriensystem vorgegeben wird - der Versuch des Forschers, vor allem die zeitliche Dauer, und auch den räumlichen Bereich einer einzelnen Beobachtung festzulegen. Der Forscher versucht, den Beobachter dazu zu bringen, den Verhaltensstrom in möglichst eindeutig angebbarer Weise in Zeitabschnitte zu gliedern. Schließlich will ein Forscher oft nicht nur wissen, was beobachtet wurde, sondern auch, wann und wo. So sehr sich ein Forscher auch bemüht, den Beobachter in dieser Hinsicht so genau wie möglich anzuleiten, er stößt doch auf psychologische Grenzen des Beobachters, die im menschlichen Wahrnehmungssystem begründet sind. Hier überschneiden sich eine allgemeine Methodenlehre der Beobachtung und die Wahrnehmungspsychologie; einige Aspekte dieser Thematik behandeln Feger & Graumann (in diesem Band).

Wie nun im Zusammenspiel von Forscher und Beobachter - auch wenn eine Person beide Rollen spielt - die Definition der Beobachtungseinheit ausgehandelt wird, hat für die Beobachtungsergebnisse weitreichende Folgen. Aus

Arbeiten des Kreises um Newton (1973, 1976, Newton & Enquist 1976, Newton et al. 1977; s.a. Feger & Graumann in diesem Band) wissen wir, daß Beobachter spontan selbst das Beobachtete in Einheiten untergliedern. Daß sich unterschiedliche Ergebnisse einstellen können, wenn entweder die zeitliche Ausdehnung vorgegeben oder die Dauer des beobachteten Phänomens vom Beobachter selbst festgelegt wird, berichten Hayes et al. (1970). Einige Autoren betonen nun, man solle die Wahl der Einheit nicht dem Beobachter überlassen, so z.B. Grümer (1974, S. 41): „Die Definition von Beobachtungseinheiten wird in strukturierten Beobachtungen nicht dem Beobachter selbst überlassen bleiben können, sondern wird zu einer spezifischen Aufgabe eines Forschers oder Untersuchungsleiters werden müssen. Aufgrund seiner Kenntnisse über Strukturen und Verhaltensabläufe in einem Beobachtungsfeld ist er in der Lage, Beobachtungseinheiten festzulegen.“ Und so legen auch v. Cranach & Frenz (1969, S. 286) in ihrer Definition fest: „Als Beobachtungseinheit wird derjenige Bestandteil in einem Verhaltensablauf bezeichnet, der dem Untersucher als kleinstes, nicht reduzierbares Ereignis zur Analyse des Verhaltens notwendig erscheint.“ Die Notwendigkeit, die Einheit durch den Forscher vorzugeben, wird mit den weitreichenden Folgen begründet, die sich aus der Wahl der Einheit ergeben: „Einmal legt man mit der Enge bzw. Breite der Beobachtungseinheit zugleich die Variationsmöglichkeiten fest, die das Beobachtungssystem bei der Abbildung des realen Geschehens zuläßt. Zum anderen muß man alle Fälle, die einer Beobachtungskategorie zugeordnet worden sind, als untereinander invariant ansehen.“ (v. Cranach & Frenz, 1969, S. 286).

Ferner ist die Wahl der Einheit für die später behandelte Frage der Beobachterübereinstimmung relevant. Idealerweise stimmen Beobachter dann überein, wenn sie Gleiches beschreiben. Die Beobachtungseinheit festzulegen bedeutet unter diesem Gesichtspunkt einen Versuch, zur Beobachtung des Gleichen anzuregen. Möglicherweise gibt die zu untersuchende Fragestellung vor, wie aus theoretischen oder praktischen Gründen die Einheit gewählt werden muß. Durch die Wahl der Einheit wird das „Auflösevermögen eines Beobachtungsansatzes“ (Faßnacht, 1979) festgelegt. Das, was bei der Beobachtung als Einheit zusammengefaßt und beschrieben wird, kann in der Datenanalyse nicht mehr differenziert werden. Bales (1950) gibt zur Beschreibung von Interaktion in Problemlösegruppen zwölf globale Kategorien vor, wodurch umfassende Einheiten gebildet werden dürften, während Frey & Pool (1967) allein 125 verschiedene Kopfstellungen einer sitzenden Person unterscheiden. Man sollte jedoch festhalten, daß die Beobachtungseinheit letztlich vom Beobachter gebildet wird, wenn auch sein Spielraum dabei durch genaue Instruktion, Training u.a. eingeschränkt werden kann. Die Übereinstimmung bei der Segmentierung sollte gegebenenfalls geprüft werden. Die Grenze des Auflösevermögens der Beobachter ist erreicht, wenn die Verlässlichkeit der Unterscheidungen zu gering wird.

Wie Einheiten gebildet werden, unterscheiden Kalbermatten & v. Cranach (1980) nach vier Gesichtspunkten. Sie akzentuieren „*natürliche*“ gegenüber „*künstlichen*“ Einheiten. Natürliche Einheiten sind „bereits vor dem wissenschaftlichen Prozeß vorhandene Unterscheidungstendenzen des Beobachters“, z.B. ob eine Person lacht oder nicht. Künstliche sind „theoretisch fundierte, im Forschungsprozeß gewonnene Merkmale“, ob eine Person beispielsweise die Wangenmuskeln aktiviert. Zweitens wird zwischen *sozial bedeutungsvollen* und *physikalisch definierten* Einheiten unterschieden. Bei ersteren werden die Bedeutungen von der sozialen Gemeinschaft zuerkannt. Sie lassen sich feststellen an der Übereinstimmung, mit der Mitglieder der Gemeinschaft sie definieren, und ihre Funktion sei es, den „mehr oder minder reibungslosen Ablauf der Interaktion zu gewährleisten“. Drittens werden *funktionale* gegenüber *strukturellen* Einheiten unterschieden. Funktionale seien „dynamisch konzipiert“, sie hielten „die wechselseitige Auswirkung einer Einheit auf eine andere und auf das Gesamtergebnis des Verhaltens fest“. Drohen wird als Beispiel genannt; die geballte Faust wäre ein Beispiel für strukturelle Einheitsbildung. Viertens wird, wie auch sonst häufig in der Literatur, zwischen *mola-*ren und *molekularen* Einheiten unterschieden, nach der „Höhe des Abstraktionsgrades“ oder dem Umfang der Zusammenfassung.

Kalbermatten & v. Cranach vertreten dann die Auffassung, bei der Analyse menschlicher Handlungssysteme sei es erforderlich, verschiedene, hierarchisch geordnete Organisationsebenen der Einheiten zu unterscheiden. Jede Einheit einer höheren Ebene lasse sich in Untereinheiten der nächsttieferen Ebene aufgliedern; die Einheiten verschiedener Ebenen unterschieden sich auch qualitativ, da die höheren Einheiten in der Regel von ihrer Funktion her und durch soziale Bedeutungsverleihung definiert seien. Eine wesentliche Aufgabe der Forschung bestehe dann in der Untersuchung des genauen Zusammenhanges zwischen Einheiten verschiedener Ebenen (z.B.: Welche Einheiten der tieferen Ebene - etwa Ballen der Faust und Heben der Stimme - werden wie kombiniert, damit die Einheit „Drohen“ entsteht?).

3.3 Kategoriensysteme

Der Bildbereich eines Abbildungssatzes legt fest, welcher Aspekt am Beobachteten untersucht werden soll. Im einfachsten Fall, wie im Streßbeispiel, ist dies lediglich *eine* Variationsdimension, eine Hinsicht, nämlich das Ausmaß des Stresses, das sich in verschiedenen Verhaltensweisen unter verschiedenen Bedingungen zeigt. Eine Kategorie beschreibt Ausprägungen von Reaktionsmerkmalen, die qualitativ oder quantitativ sein können. Variation kann unter den gleichen Bedingungen und an den gleichen Verhaltensweisen in mehr als einer Hinsicht unterschieden und zugleich beobachtet werden. Die Auflistung dieser Hinsichten in ihren qualitativen und quantitativen Abstufungen stellt

das Kategoriensystem dar. Ein Kategoriensystem stellt also die Unterscheidungen zusammen, die der Beobachter treffen kann, und legt fest, bei welchen Zuständen und Prozessen am Beobachteten er sie treffen soll. Wir sprechen von System, weil die Unterscheidungsmöglichkeiten mehr oder weniger geregelt aufeinander bezogen sind, etwa durch die Vorschrift, das Beobachtete solle nur einer Abstufung von mehreren bei einer Kategorie zugeordnet werden können. Zum Kategoriensystem i.e.S. kommen häufig noch Anweisungen hinzu, kodifizierte Anleitungen an den Beobachter, wie er das Beobachtete erfassen, zuordnen und registrieren soll, was das Kategoriensystem i. w. S. ausmacht.

V. Cranach & Frenz (1969) wie auch Medley & Mitzel (1963) unterscheiden drei Arten von „Beobachtungssystemen“, und zwar Zeichen-Systeme, Kategorien-Systeme und Schätzskaleten (rating scale format). Ein einfaches Zeichensystem bestünde beispielsweise darin, daß ein Beobachter immer dann, wenn eine Person bei Rot eine Ampel passiert, einen Registrierknopf drückt. Das Überschreiten der Markierungslinie wäre für ihn ein Zeichen, alles nicht als Zeichen definierte Verhalten bleibt unregistriert. Eines der bekanntesten Kategoriensysteme im Sinne von Medley & Mitzel ist das von Bales (1950, 1968; siehe dazu auch Grüner, 1974, Manz, 1974). In der Regel hat der Beobachter bei diesen Kategoriensystemen nicht nur - wie bei Zeichensystemen - zu entscheiden, ob ein bestimmtes Verhalten aufgetreten ist, sondern auch welches von den einander ausschließenden Möglichkeiten des meist als vollständig konzipierten Kategoriensystems.

Weitere Beispiele für Kategoriensysteme sind das von Caldwell (1969), das sich auf Verhalten von Kindern bezieht, oder das von Kaufman & Rosenblum (1966) für das Sozialverhalten von Primaten; Simon & Boyer (1974) geben eine Übersicht über 99 Systeme. Ausgearbeitete Systeme von Ratingskaleten finden sich z.B. in Thomae (1968, s. auch Rudinger & Feger, 1970).

Viele Systeme sind keine reinen Zeichen-, Kategorien- oder Schätzskaletensysteme. Für ihre Herleitung sind selten theoretisch so konsistent systematisierte Annahmen wie bei Bales vorhanden, doch sollte der Weg von der Fragestellung des Forschers zur Auswahl und Formulierung der Kategorien so einsichtig und zwingend wie möglich sein; die Facettentheorie könnte für die Ableitung von Kategoriensystemen verwendet werden. Viele Kategoriensysteme versuchen in dem Sinne vollständig zu sein, daß alles beobachtete Verhalten einer Kategorie zugeordnet werden kann, und sei es einer Restkategorie, in die alles nicht klassifizierbare eingewiesen wird. Dadurch entstehen logische Abhängigkeiten zwischen den Kategorien, die die statistische Auswertung erschweren. Wenn von zwei Kategorien Kategorie A als „soziales Verhalten“ und B als „nicht-soziales Verhalten“ definiert sind, muß sich perfekter negativer Zusammenhang ergeben. Diese, in umfangreichen Kategoriensystemen

nicht so leicht erkennbaren logischen Zusammenhänge vermischen sich mit den empirischen und müssen bei der Auswertung berücksichtigt werden. Umfangreiche Kategoriensysteme führen bei unabhängiger Auswertung jeder Kategorie oft zu einem erhöhten faktischen α -Niveau der statistischen Signifikanztests, weshalb multivariate Verfahren häufiger als bisher verwendet werden sollten. Ein weiteres Problem ergibt sich bei Systemen, die wie eine check-list angelegt sind, d.h. vom Beobachter lediglich zu registrieren verlangen, wie häufig eine Verhaltensweise auftritt. In der Interpretation wird dann bisweilen unterstellt, die Verhaltensdauer stünde mit der Häufigkeit in direkter proportionaler Beziehung. Für diese Unterstellung fand zwar Adams (1970) bei Schülerverhalten im Unterricht positive Evidenz, aber eine allgemeine Austauschbarkeit von Dauer- und Häufigkeitsmaßen gibt es nicht.

Longabaugh (1980) unterscheidet, was er „Kodiersysteme“ nennt, (1) danach, ob sie aus einer Theorie abgeleitet sind oder auf nicht systematisierten Erfahrungen mit dem Gegenstandsbereich aufbauen, (2) nach der Breite gegenüber der Detailliertheit, mit der sie einen Bereich erfassen, und (3) nach dem Ausmaß, in dem Schlußfolgerungen, Interpretationen durch den Beobachter erforderlich sind: „... to what extent is the coder required to take advantage of the fact that he is a socialized human being, sharing a common culture, in order to characterize the behavior?“ Auf die Folgerungen, die sich aus diesem dritten Gesichtspunkt für die Reliabilitätsfrage ergeben, gehen wir später ein.

Um einzelne Kategorien und Kategoriensysteme zu klassifizieren, scheint es z.B. für die Diskussion ihrer Validität wie für das Training der Beobachter günstig, danach zu fragen, welche Aufgaben sie dem Beobachter stellen. Da Beobachtung ein kontrollierter Prozeß der Wahrnehmung ist, sind die Aufgaben grundsätzlich diejenigen von Wahrnehmung allgemein, die man mit Luce und Galanter (Luce, 1963, Luce & Galanter, 1963a, b) bezeichnen kann als Entdecken (detection), Wiedererkennen (recognition), Unterscheiden (discrimination) und Quantifizierung (scaling). Diese vier Aufgaben lassen sich, wie Luce (1963, S. 105) dies für eine psychophysikalische Behandlung von Wahrnehmungsproblemen getan hat, als vier Fragen formulieren: (1) Liegt das zu beobachtende Verhalten vor? (2) Welche von mehreren möglichen Verhaltensweisen liegt vor? (3) Unterscheidet sich dieses Verhalten von jenem? und (4) Wie verschieden ist dieses Verhalten von jenem? Selbst bei den einfachsten Kategoriensystemen sind Aufgaben des Entdeckens und des Wiedererkennens zu lösen, bei Kategoriensystemen im Sinne von Medley & Mitzel auch Diskriminationsaufgaben, und bei Schätzskalen zusätzlich Quantifizierungsaufgaben. Um diese Aufgaben lösen zu können, definiert der Forscher für den Beobachter durch Instruktion, Kategorien und Schulung *Zuordnungsregeln*, die möglichst explizit vorschreiben, welche Erscheinungen am Beobachtungsgegenstand welchen Kategorien zuzuordnen sind. Duncan & Fiske (1977, S. 15) sprechen in diesem Zusammenhang von recognition rules: „The rules for

the identification of events as instances of the applicable categories will be termed *recognition rules*. Using the investigator's category system, the task of the rater is in the first place to recognize the occurrence of an event specified by a recognition rule.“ Reliabilität und Validität von Beobachtungen hängen davon ab, wie explizit die Zuordnungsregeln formuliert, angewendet und wie gut ihre Anwendung geprüft werden kann.

3.4 Auswahlen aus dem Universum der Beobachtungen

Meistens können nicht alle Beobachtungen, die gemäß dem Abbildungssatz prinzipiell möglich wären, auch tatsächlich angestellt werden, z.B. weil ihre Zahl zu groß ist, oder Kosten und Zeitaufwand zu hoch wären. Deshalb muß fast immer eine Auswahl getroffen, und diese begründet werden. Um Auswahlarten zu beschreiben, ist es hilfreich, den facettentheoretischen Begriff des *Struktupels* einzuführen. Da die Elemente aller Facetten sich vollständig kreuzklassifizieren lassen, d.h. aus jeder Facette jedes Element mit jedem Element jeder anderen Facette kombiniert werden kann, entstehen Kombinationen mit k Elementen, wobei k die Zahl der Facetten ist. Ihre Anzahl entspricht dem kartesischen Produkt der Facetten. Jede Kombination stellt ein Struktupel dar, in dem das erste Element aus der ersten, das zweite aus der zweiten usw. Facette stammt. Die erste Möglichkeit auszuwählen besteht darin, nicht für jedes Struktupel eine Beobachtung zu realisieren, also aus den Struktupeln auszuwählen. So hat beispielsweise Jordan (1971) bei der Konstruktion eines Vorurteils-Fragebogens nicht alle Kombinationen ausgewählt, weil sie seines Erachtens teilweise keinen psychologischen Sinn ergaben. In einem solchen Fall könnte man die Facetten nicht als vollständig überkreuzt, sondern als geschachtelt ansehen. Kann man das nicht, führt die Auswahl aus Facetten zu einer Einschränkung der Allgemeingültigkeit der Aussagen. Im Extremfall untersucht man nur ein Element einer Facette, z.B. berufstätige Frauen aus der Personen-Facette. Eine zweite Möglichkeit auszuwählen stellt die Auswahl *innerhalb* eines Struktupels dar. Bei Testitems etwa gibt es eine Fülle von Formulierungen, in denen der gleiche Inhalt ausgedrückt werden kann, und eine Auswahl ist aus Aufwandgründen erforderlich. Alle Realisationsmöglichkeiten, die durch das gleiche Struktupel beschrieben werden, sind für die Theorie, die zum Abbildungssatz geführt hat, äquivalent. Je mehr Realisationen dann pro Struktupel vorliegen, um so verlässlicher ist die Datenbasis.

3.4.1 Auswahl von Personen

Es gibt verschiedene Gründe, warum man mehr als eine Vp untersucht, selbst wenn man überzeugt ist, das interessierende Verhalten könne in allen Erschei-

nungsformen und aufgrund aller relevanten Bedingungen an jeder beliebigen und somit jeder einzelnen Person beobachtet werden. (1) Wenn der Forscher das interessierende Verhalten nicht herbeiführen kann oder will, oder es zu selten bei einer Person auftritt, beobachtet er mehrere, um die *Auftretenswahrscheinlichkeit* des Verhaltens zu erhöhen. (2) Der Forscher möchte eine Vp unter verschiedenen Bedingungen beobachten, jedoch kann die gleiche Vp nur unter einer Bedingung untersucht werden, z.B. wegen störender carry-over-Effekte (*Kontrollproblem*). Dann ergibt sich insbesondere in Feldstudien das Problem, unter verschiedenen Bedingungen solche Personen zu beobachten, die vergleichbar sind, so daß Verhaltensunterschiede eindeutig den Bedingungen zugeordnet werden können (interne Validität). (3) Der Forscher möchte Aussagen treffen über die *Generalität*, über Verbreitungsgrad und interindividuelle Variabilität des Verhaltens. Nur bei dieser Absicht werden Überlegungen relevant, wie man eine repräsentative Stichprobe aus der interessierenden Population gewinnen kann (dazu gute Einführungen: Kish, 1953; Scott & Wertheimer, 1962; zur Vertiefung: Böltken, 1976; Cochran, 1953; Deming, 1950; Hansen et al., 1953; Yates, 1953).

Ein besonderes Problem ergibt sich aus der Tatsache, daß *freiwillige Vpn* untersucht werden, was oft schon aus ethischen Gründen unvermeidlich ist. Die Arbeiten von Rosenthal (1965) sowie Rosenthal & Rosnow (1969, 1975; dort weitere Literatur) legen zwei Schlüsse nahe: Es gibt bestimmte situative Umstände, die einige Personen zur freiwilligen Vp werden lassen, andere Personen nicht. Und zwischen freiwilligen und zwangsrekrutierten Vpn gibt es einige Unterschiede in Merkmalen wie Höhe der Schulbildung, des sozialen Status, der Intelligenz, der Anpassung und motivationaler Bedürfnisse. Das Problem ist damit im Sinne von Campbell & Stanley (1963) eines der externen Validität, der fraglichen Grenzen der Verallgemeinerbarkeit von Befunden. Allerdings weist Kruglanski (1975) darauf hin, daß Versuchsergebnisse nicht generell von der Freiwilligkeit der Vpn abhängen. Bredenkamp (persönliche Mitteilung) weist darauf hin, nur wenn eine solche Abhängigkeit durchgängig gegeben sei, müsse man diese Artefaktmöglichkeit auch allgemein kontrollieren. Hingegen müßte zum speziellen Nachweis eines Artefaktes gezeigt werden, daß der Faktor „Freiwilligkeit“ disordinal mit der unabhängigen Variablen einer Untersuchung derart interagiert, daß die Relation zwischen unabhängiger und abhängiger Variable für unterschiedliche Gruppen verschieden ausfällt. Die bei Rosenthal & Rosnow aufgeführten Untersuchungen, die diese Interaktion geprüft haben, zeigen eben nicht disordinale Interaktionen (s. Bredenkamp 1980).

Wenn Verhaltensweisen untersucht werden sollen, die von anderen Personen nicht beeinflußt werden, oder bei denen ein solcher Einfluß nicht interessiert, kann man die Stichprobe so ziehen, daß jede Person unabhängig von jeder anderen eine gleiche oder bekannte Chance hat, gezogen zu werden. Sonst

kann das Paar, die Gruppe, die Kultur als Element der Population gezogen werden, wobei die Facettentheorie zur Populationsdefinition benutzt werden kann. Für kulturvergleichende Studien ringen zwei Kriterien für die Zusammensetzung der Personenstichprobe um gleichzeitige Beachtung „. . . whether a given relationship between variables will be obtained across cultures (irrespective of people) and across people (irrespective of their cultural membership)“, Longabaugh (1980, S. 73). Eine bemerkenswerte Lösung dieser Aufgabe findet sich in Whiting et al. (1966).

3.4.2 Auswahl und Schulung von Beobachtern

Selten hat ein Untersucher überhaupt die Möglichkeit, aus einer Gruppe von mehreren die ihm geeignet erscheinenden Beobachter auszusuchen. Die formalen Kriterien sind bei Auswahlentscheidungen meistens hinreichend hohe Übereinstimmung mit einem Standard, mit dem „mittleren“ oder typischen Beobachter (s.u. Krippendorff, 1970), oder eine hohe oder nicht mehr zu steigernde Übereinstimmung zwischen einem Paar oder einem Team aus der Gesamtmenge der Beobachter. Bock (1956) hat die Auswahl von Beurteilern in Präferenzexperimenten systematisch untersucht. Er weist darauf hin, daß die Auswahl nach interrater agreement die Annahme einschließt, alle Beurteiler bezögen sich auf die gleiche Variable, die nur in einer Dimension variere. Atypische Urteile würden auch dann nicht berücksichtigt, wenn sie reliabel sind (zur Gruppierung von Beobachtern nach der Ähnlichkeit ihrer Beurteilungsstrategien s. Naylor et al., 1967; Naylor & Schenck, 1966).

Wegen der besonderen Bedeutung, die der Verhaltensbeobachtung in der Verhaltenstherapie zukommt, ist im Arbeitskreis von O'Leary (z.B. Romanczyk et al., 1973) der *Schulung von Beobachtern*, meistens unter Reliabilitätsgesichtspunkten, größere Aufmerksamkeit zugewandt worden. Kontrollierte Übung führt fast immer zu einer bemerkenswerten Erhöhung der Übereinstimmung zwischen den Beobachtern. Nay & Kerkhoff (1974) zeigten für ein Kodiersystem mit 22 Symbolen, daß Feedback über Fehler die Reliabilität nach Abschluß des Trainings deutlich gegenüber einer Kontrollgruppe erhöht, die ohne Rückkoppelung am gleichen Videotape-Material lediglich kodieren übte. Training geschieht in jüngster Zeit häufig an Videotapeaufnahmen. Nay & Kerkhoff stellen folgende Vorteile heraus: 1. kann die Reliabilität in bezug auf eine standardisierte Vorlage abgeschätzt werden, 2. kann die Darbietung für Feedback an einen Beobachter unterbrochen werden, 3. können insbesondere schwierige Passagen beliebig oft, bis zur Beherrschung der Kodierung, wiederholt werden, 4. können in natura selten auftretende, jedoch wichtige Phänomene häufiger gezeigt werden; für ihr Auftreten kann sensibilisiert werden - generell läßt sich die Materialauswahl so vornehmen, daß alle Kategorien hinreichend oft geübt werden können.

Beobachter sind jedoch möglicherweise reaktiv: Ihre Einschätzungen werden verlässlicher, wenn man sie vorab darüber unterrichtet, daß Reliabilitätsprüfungen stattfinden. Wenn man Beobachtern Gelegenheit zum „Erfahrungsaustausch“ gibt, ändern sie im Verlauf der Beobachtungszeit ihre Interpretation der vorgegebenen Kategorien, nähern sie einander an und werden so reliabler, wenn auch nicht unbedingt valider (nach Romanczyk et al., 1973; s.a. Reid 1970, sowie Feger & Graumann in diesem Band).

3.4.3 Auswahl des zu beobachtenden Verhaltens

Mit Wright (1960, S. 73) kann man diese Auswahl unter inhaltlichem und unter zeitlichem Aspekt betrachten. „Material coverage . . . refers to what and how much at a time the observer tries to see in the stream of behavior.“ Hingegen bezieht sich „continuum coverage“ auf „. . . the length or parts into which the stream of behavior is divided for purposes of observation“. Diesen zweiten Aspekt bezeichnen wir als *Ziehen von Zeitstichproben* (time sampling). Longabaugh weist auf drei Entscheidungen hin, die beim Ziehen von Zeitstichproben zu fällen sind: 1. Ob man den Plan fixiert, vor Beginn der Beobachtung festlegt, oder offen läßt, wann wie lange etc. zu beobachten sei, und allenfalls Randbedingungen vorgibt, beispielsweise die, jede Person solle nicht öfter als einmal pro Tag beobachtet werden. 2. Wie lange der Zeitraum dauern soll, in dem das Ereignis beobachtet wird. 3. Wie lange das Pausenintervall zwischen Beobachtungen dauern soll. Die Gefahr beim Ziehen von Zeitstichproben liegt allgemein darin, die zeitliche Struktur des beobachteten Phänomens nicht angemessen zu erfassen. Andererseits wird kontinuierliche Beobachtung, z.B. eines gesamten Therapieverlaufs, nicht immer möglich sein (Literatur: Arrington 1939, 1943; Hutt & Hutt 1974; Wright 1960).

Beim Ziehen von Zeitstichproben kann man entweder ein einziges Intervall wählen und es festlegen, indem man Anfangs- und Endzeitpunkt bestimmt. Man kann sich auch für mehrere Intervalle entscheiden, dann sind deren Länge, u.U. auch die Pausenlänge und Startzeitpunkte zu vereinbaren. Ein wesentlicher Gesichtspunkt ergibt sich dabei aus der Art der Beobachtungen, die man erwartet, insbesondere aus deren *vermuteter Dauer und Häufigkeit*. Sackett (1978) stellt folgendes Schema auf (s. S. 19 oben).

Als Optimierungskriterium für die Ausschnittwahl formuliert Sackett (S. 26): „The sampling problems of observational research involve maximizing the chances of actually observing any of these four types of behavior in sequences that are representative of the typical behaviors of the subjects under study.“ Um Verhalten vom Typ II und IV zu erfassen, müssen relativ lange Zeitstichproben mit kontinuierlicher Beobachtung gezogen werden; bei Typ I-Verhalten können die Stichproben kürzer sein, bei Typ III-Verhalten können die Beobachtungen diskontinuierlich vorgenommen werden. Auch die *Auswer-*

	häufig	selten auftretend
Verhalten kurzer und nahezu konstanter Dauer (momentary behaviors)	Z.B. Lidschlag TYP I	Z.B. Niesen TYP II
Verhalten variabler Dauer (duration meaningful)	Z.B. Gesprächsbeitrag in Dyade Typ III	Z.B. Akt physischer Aggression TYP IV

tungsziele, z.B. die Art der beabsichtigten Vergleiche und der zu prüfenden Hypothesen, bestimmen die Ausschnittwahl mit. Wenn z.B. Aussagen über Abfolgen gleicher oder verschiedener Verhaltensweisen oder über Zusammenhänge von Umweltzuständen und Verhalten beabsichtigt sind, liegt in der Regel kontinuierliche Beobachtung in relativ langen Intervallen nahe. Wenn nur Frequenz, nicht aber Dauer interessiert, können die Zeitstichproben unterschiedlich lang sein, sonst muß entsprechend adjustiert werden (ausführlich Sackett, 1978). Wenn mehrere Individuen zusammen auftreten und die Beziehungen zwischen ihnen interessieren, stellt sich die Frage, ob ein Beobachter eine Person oder mehrere zugleich beobachten soll. Eine Möglichkeit, den Ausschnitt festzulegen, besteht darin, ein bestimmtes Individuum auszuwählen (*focal individual coding method*) und während einer bestimmten Beobachtungsperiode alle interessierenden Verhaltensweisen dieses Individuums zu registrieren, insbesondere auch, mit wem es interagiert. Damit diese Methode zu repräsentativen Verhaltensstichproben für das Individuum und die Gruppe führt, sollte 1) jede Person reihum zum Fokus der Beobachtung werden, 2) bei jeder Interaktion bei der das fokale Individuum beobachtet wird, sollte festgehalten werden, mit wem und in welcher Art es interagiert, 3) für jede im Mittelpunkt stehende Person sollte Verhalten in einer genügend großen Anzahl von Situationen erfaßt werden, und 4) die interessierenden Verhaltensweisen sollten über die Gelegenheiten hinweg relativ stabil sein. Sind diese Bedingungen nicht erfüllt, so läßt sich aus Beobachtungen einzelner Individuen das Interaktionssystem der Gruppe nicht unverfälscht konstruieren. Je nach Zielen und Ressourcen des Forschers bleibt dann die Möglichkeit, alle Individuen gleichzeitig zu beobachten, u.U. mit technischen Hilfsmitteln, die Zahl der Kategorien zu vermindern, die Zahl der Beobachter zu vergrößern und deren Beobachtungen zu synchronisieren (Sackett, 1978).

Stehen mehrere Beobachter zur Verfügung, so sollten sie zufällig oder systematisch auf die Vpn oder die Versuchsbedingungen verteilt werden. Eine feste

Zuordnung führt zu der Gefahr einer Konfundierung von Unterschieden zwischen Beobachtern einerseits, Beobachteten oder Bedingungen andererseits.

3.4.4 Übergreifende Auswahlstrategien

Wir bezeichnen Anweisungen zur Auswahl dann als übergreifend, wenn sie sich auf mehr als eine Facette beziehen, z.B. nicht nur auf die Auswahl von Personen, sondern auch auf Situationen. In dem folgenden, leicht abgeänderten Schema von Longabaugh (1980, S. 78ff.) in Tab. 1 können Dauer von Beobachtung und Pausenintervall bei jeder Strategie fixiert sein oder variieren. Die Strategien unterscheiden sich danach, ob die zu beobachtenden Personen, die aufzusuchenden oder herzustellenden Umwelten und das zu erfassende Verhalten fixiert sind oder variieren. Fixiert heißt hier: bei der Planung und vor Beginn der Beobachtung festgelegt. Bleibt die Bedingung variabel, so bestimmt das beobachtete Ereignis, auf welche Situationen, Personen und Verhaltensweisen der Beobachter stößt.

Zur Erläuterung des Schemas besprechen wir einige Strategien. Beim Ziehen von *Ereignisstichproben* (event sampling) wird ein vorbestimmtes Verhalten nur dann aufgezeichnet, wenn bestimmte Personen es in einer zuvor festgelegten Umgebung zeigen. Munroe (1973, hier nach Longabaugh 1980, S. 80) beobachtete bei bestimmten Kleinkindern in zeitlich und räumlich vorher fixierten Umgebungen, wenn die Kinder schrien: wie lange sie dann weinten, nach wieviel Sekunden eine Pflegeperson erschien, wer das war und wie sie sich verhielt, usw. Das Ziel dieser Studie bestand darin, auf Verteilungen bezogene Aussagen zu vergleichen, die durch Ziehen von Ereignisstichproben gegenüber dem Ziehen von Zeitstichproben gewonnen wurden.

Ein Beispiel für das Ziehen von *Personen-Umwelt-Stichproben* ist die Arbeit von Schoggen (1976, s. Longabaugh 1980, p. 80), in der Dreijährige aus drei verschiedenen sozialen Schichten (a) beim Essen und beim freien Spiel, (b) im oder nahe bei dem Elternhaus des Kindes, und (c) in Abwesenheit der Mutter beobachtet wurden. Waren a, b und c gegeben, wurden eine Reihe von Verhaltensweisen aufgezeichnet. Letzteres - mehrere Mengen im Bildbereich - ist im Gegensatz zum Ziehen von Ereignisstichproben wesentlich für diese Strategie, und das Beobachtungsergebnis kann dementsprechend nicht nur aus dem Vergleich über die sozialen Schichten bestehen, sondern neben Aussagen über Verteilungsformen auch solche über korrelative Zusammenhänge der Verhaltensvariablen enthalten. Selten wird die Strategie verwendet, lediglich *Personen-stichproben* zu fixieren. Eine bekannte Ausnahme ist Barkers (1951) „One boys day“, in der zahllose Verhaltensweisen eines bestimmten Jungen in all seinen Umwelten eines einzigen Tages aufgezeichnet wurden. Das Beobachtungsergebnis stellt ein Inventar von Verhaltensweisen, Umwelten und ihren Zusammenhängen dar, bisweilen als *specimen record* bezeichnet (s. Weick 1968, S. 416f.; Wright 1960, S. 86).

Tabelle 1: Übergreifende Auswahlstrategien nach Longabaugh

<i>Bezeichnung der Strategie</i>	<i>Personen</i>	<i>Umgebung</i>	<i>Verhalten</i>
(1) Ziehen von Ereignisstichproben	fixiert	fixiert	fixiert
(2) Ziehen von Personen-Umwelt-Stichproben	fixiert	fixiert	variabel
(3) Ziehen von Personenstichproben	fixiert	variabel	variabel
(4) Ziehen von Umwelt-Verhaltens-Stichproben	variabel	fixiert	fixiert
(5) Ziehen von Umwelt-Stichproben	variabel	fixiert	variabel
(6) Ziehen von Verhaltensstichproben	variabel	variabel	fixiert
(7) Nichtfixiertes Ziehen von Stichproben	variabel	variabel	variabel

4. Die Bewertung von Beobachtungen

Damit Beobachtungen als wissenschaftlich gelten können, müssen sie bestimmte Kriterien notwendigerweise erfüllen. Die Begründung der Kriterien als notwendige, m.E. ein Thema der Wissenschaftstheorie, kann hier nur angedeutet werden. Die Planung von Beobachtungen geschieht auch mit dem Ziel, in der erwarteten Bewertung diesen notwendigen Kriterien zu genügen, wobei in der Vorbereitungsphase oft Planungsschritte und solche der Bewertung rückgekoppelt werden, bis die Kriterien hinreichend erfüllt sind. Von den notwendigen unterscheiden wir differenzierende Kriterien. Dabei gehen wir davon aus, daß alle empirischen Methoden in Psychologie und Sozialwissenschaften als Varianten der wissenschaftlichen Beobachtung aufgefaßt werden können, wobei die wesentlichen Unterschiede zwischen den Methoden sich aus dem Zweck ergeben, dem sie dienen sollen. So ist der Zweck des Tests, inter- und intraindividuelle Unterschiede zu bestimmen, Strukturanalysen wie multidimensionale Skalierung oder Clusteranalysen haben die Aufgabe, Ordnungsstrukturen nachzuweisen, und das Experiment sucht nach Zusammenhängen zwischen Variiertem und Variierendem. Wegen dieser Absicht werden die Bedingungen, unter denen die Variablen des Experiments manipuliert und erfaßt werden, im idealen Experiment so kontrolliert, daß eindeutige Schlüsse von Antezedenzbedingungen auf Beobachtungsvariablen möglich sind. Bei den meisten Tests sind Durchführungssituation und Reaktionsmöglichkeiten in der Form der Standardisierung kontrolliert, wieder mit dem Ziel, Beobachtungen im interindividuellen Vergleich möglichst eindeutig und zutreffend deuten zu können. Test und Experiment sind also nicht, wie bisweilen behauptet wird, Alternativen zu oder Konkurrenten der wissenschaftlichen Beobachtung, sondern Beobachtungen unter zweckspezifischen Kontrollmaßnahmen. Beobachtungsstudien i.e. S. werden oft dann durchgeführt, wenn aus verschiedenartigen Gründen diese Kontrollen nicht durchgeführt werden können, z.B. wenn man die unabhängigen Variablen nicht variieren kann oder darf. Oft auch leisten Beobachtungsstudien die Pionierarbeit, das Universum der Beobachtungen einzugrenzen und zu strukturieren.

Während der Grad der Kontrolle variieren kann und die Art der Kontrolle dem jeweiligen Ziel angepaßt ist, stellt Kontrolliertheit wieder nur Mittel zum Zweck dar: verlässliche und gültige Schlüsse ziehen zu können. Reliabilität und Validität sind jene notwendigen Kriterien, die in der Psychologie das meiste Interesse gefunden haben. Reliabilität spielt als Kriterium in der Psychologie deshalb eine so große Rolle, weil wir oft nicht wissen, welche Variablen wie zu kontrollieren sind, oder diese Kontrollen nicht anwenden können, um jene Bedingungen herzustellen, deren Realisation die Theorie verlangt, und jene Variationsquellen auszuschalten, mit denen sich die jeweilige Theorie nicht befaßt und die insofern gerade irrelevant sind. Reliabilität kann daher allgemein als das Verhältnis von systematischer, d.h. theoretisch erklärter Varianz,

zu Fehlervarianz, d.h. durch die jeweils betrachteten Bedingungen nicht erklärte Varianz, konzipiert werden. Wenn die theoretischen Ideen, die den Abbildungssatz formulierten, zutreffen, dann ist es eine Frage der Kontrolle und ein Ausdruck der Reliabilität, wie stark der empirische Zusammenhang zwischen Definitionsbereich und Bildbereich des Abbildungssatzes ist.

Kategoriensysteme unterscheiden sich, wie erwähnt, nach dem Ausmaß, in dem Beobachtetes interpretiert werden muß, damit es der Beobachter kodieren kann. Longabaugh (1980, S. 102) zeigt, welche Konsequenzen dies für die Übereinstimmung von Beurteilern hat, wenn sie auf den ihnen gemeinsamen kulturellen Hintergrund angewiesen sind, um Verhalten eindeutig zu interpretieren:

„Even within the same culture persons have only partial access to shared meaning and an imperfect understanding of what they have access to. When several observers are asked to categorize a particular unit of behavior, the greater the number of observers, the less the average agreement obtained. . . . As the number of observers increases, the meaning shared by all remains at best constant. Meanwhile the component of meaning shared by fewer than all observers increases as does the idiosyncratic meaning unique to each observer.“

Nicht nur die Verlässlichkeit, sondern auch die Gültigkeit von Urteilen, die auf Verhaltensbeobachtung basieren, hängt von dem gemeinsamen Verständnis der sozialen und kulturellen Umwelt ab. Verständlicherweise wird dies besonders in kulturvergleichenden Untersuchungen deutlich, weshalb wir noch einmal Longabaugh (1980, S. 102) zitieren können:

„Veridical judgments regarding actor intention require that *actor* and *observer* share the same symbols; veridical judgments regarding the effect of the behavior require that *observer* and *target* share the same symbols; veridical judgments concerning the meaning of the behavior for the relationship require that *actor*, *target*, and *observer* share the same symbol. Veridical judgment concerning the cultural significance of the behavior requires that *observer*, *actor*, *target*, and *most* other *cultural participants* (in principle) agree upon the meaning of the symbol.“

5. Die Reproduzierbarkeit von Beobachtungen

Reliabilität hat je nach der mit Beobachtungen konfrontierten Theorie verschiedene Aspekte. Sagt die Theorie nichts über Unterschiede zu verschiedenen Meßzeitpunkten, dann läßt sich die Kontrolliertheit unter dem Gesichtspunkt der zeitlichen Stabilität problematisieren. Spezifiziert die Theorie keine Effekte unterschiedlicher Meßinstrumente, so sollten die Ergebnisse mit Hilfe unterschiedlicher Instrumente wiederholbar sein. Generell beschreibt also Re-

liabilität die Reproduzierbarkeit von Beobachtungen unter theoretisch für das Auftreten des Beobachteten äquivalenten Bedingungen bei Unterschieden in theoretisch irrelevanten Bedingungen. Was als wesentliche oder als irrelevante Bedingungen anzusehen ist muß von der Fragestellung der Untersuchung her begründet werden. Eine Reliabilitätsstudie legt fest, was erfaßt werden soll, welche Bedingungen konstant sind und welche variabel; letztere stellen die Hinsicht dar, in der die Ergebnisse reproduzierbar sein sollen - das können z.B. Meßwiederholungen zu verschiedenen Zeitpunkten, mit verschiedenen Apparaten und durch verschiedene Personen sein. Formal unterscheiden wir zwischen (1) den verschiedenen Meßinstrumenten, für die hier interessierende Reliabilitätsforschung sind das durchweg *Beobachter*, (2) den verschiedenen *Beobachtungsgegenständen*, hier meist verschiedene Personen, (3) den verschiedenen *Beobachtungshinsichten*, hier in der Regel als inhaltlich verschiedene Kategorien konzipiert, die an der gleichen Person erfaßt werden können, und schließlich (4) verschiedenen *Beobachtungsgelegenheiten*, oft in Form von wiederholten Messungen realisiert. Die Reproduzierbarkeit von Beobachtungsergebnissen studiert man über die Variation wenigstens einer dieser vier Mengen; oft interessiert die Vergleichbarkeit der Ergebnisse, die mit verschiedenen Meßinstrumenten - durch diverse Beobachter - gewonnen wurden, oder über unterschiedliche Meßzeitpunkte. Die Reliabilitätsarten der Testtheorie (s. Kap. von Kristof und Fischer in diesem Band) finden hier ihre logische Entsprechung.

Jede Beobachtung und Messung, auch die im Bereich der klassischen Naturwissenschaften vorgenommene, ist mit mehr oder weniger großen Meßfehlern behaftet. Dafür kann man sich zahlreiche Gründe vorstellen, von der falschen Justierung des Meßgerätes bis zur Unaufmerksamkeit des Ablesenden. Da man mit solchen Fehlern rechnet, interessieren die Fragen, wie groß der Fehler ist, ob er konstant, systematisch ist (= bias) oder ob er zufallsmäßig variiert (= error). Häufig geht man davon aus, man sei in der Lage, größere, systematische Fehler zu identifizieren, dann auszuschalten oder zu kontrollieren - meist bei der Konstruktion des Meßgerätes oder der Formulierung von Kategorien und Instruktion. Als Beispiel für diesen Prozeß in der Psychologie sei die Arbeit von Hendel & Weiss (1970) erwähnt, in der ein Nachweis *stabiler* interindividueller Unterschiede in der Konsistenz bei Paarvergleichen gelang. Bestimmte Reliabilitätseinschätzungen können also erhöht werden, wenn die konsistenteren Beurteiler ausgewählt oder die Urteile z.B. durch Schulung konsistenter werden. Was bleibt, sind dann zahlreiche, meist unidentifizierte Quellen, die kleine Zufallsfehler bewirken, über deren Wirkungen man annimmt, daß sie sich additiv kombinieren und einen Erwartungswert von Null haben. Also muß zur Fehlerabschätzung wie zur Bestimmung des richtigen, des „wahren Wertes“ dieser Grenzwert von Messungen approximiert werden, was innerhalb der Naturwissenschaften häufiger getan wird und zu Serien von Hunderten von Meßwiederholungen führt. Als Fehlermaß wird dann die mitt-

lere Abweichung (average error) bestimmt, wie aus der folgenden Tabelle hervorgeht:

Messung	Ergebnis	absolute Abweichung vom Mittelwert in mm
1	100 m m	0
2	101 m m	1
3	104 m m	4
4	99 m m	1
5	96 m m	4
$\Sigma 10; n = 5 \text{ Messungen}$		

mittlere Abweichung = $\bar{x} = 2 \text{ mm}$

Zur besseren Vergleichbarkeit über verschiedene Skalen wird auch der prozentuale Fehler (percentage error) angegeben, der als mittlere Abweichung relativ zur Größe des Objekts definiert ist. Betrachten wir im Beispiel 100 mm als Objektgröße, dann beläuft sich der prozentuale Fehler auf $\frac{2}{100} = 2\%$.

Messungen und damit Reliabilitätsprobleme wurden in der Psychologie verständlicherweise dann besonders brisant, als von Testwerten existenzielle Entscheidungen abhingen, so daß im Bereich der Testtheorie und Testkonstruktion eine lange Tradition der Reliabilitätsforschung zu finden ist. Ohne hier auf die zahlreichen Gründe dafür einzugehen, darf man allerdings davon ausgehen, daß eine Person nur selten tausendmal -wie ein Objekt der Naturwissenschaften - gemessen werden kann. Die Lösung dieses Problems in der klassischen Testtheorie besteht darin, statt vieler Messungen an einer Vp zwei oder wenig mehr Erhebungen an zahlreichen Vpn durchzuführen. Um diesen Ausweg beschreiten zu können, muß man annehmen, man könne Zufallsstichproben aus der für eine Messung in Frage kommenden Population von Vpn ziehen und dabei stehe für diese Meßzwecke jede Vp im Prinzip stellvertretend für jede andere. Als Fehlermaß bietet sich analog zum mittleren Fehler die durchschnittliche absolute Abweichung von erster und zweiter Messung an. Dieses Maß wird jedoch in der Psychologie selten angewandt, statt dessen wird ein Produkt-Moment-Korrelationskoeffizient bestimmt, wodurch mögliche Unterschiede zwischen den Mittelwerten und Streuungen beider Meßwertreihen sich nicht auf das Fehlermaß auswirken können. Dieses Vorgehen kommt der Annahme gleich, Differenzen in Skaleneinheit und Skalensprung seien irrelevant, vielmehr seien die relativen Unterschiede zwischen den Personen wesentlich.

Die andere Tradition der Reliabilitätsforschung ist mit der Methodenlehre der wissenschaftlichen Beobachtung verbunden. Hier tritt an die Stelle des Meßinstrumentes oder Tests der Beobachter als Datenquelle, und erhoben wird die Reproduzierbarkeit seiner Aussagen über wiederholte Gelegenheiten als *intra-rater-consistency* und die Vergleichbarkeit der Befunde verschiedener Beobachter als *inter-rater-agreement*. Die Methodik der Reliabilitätsprüfung hat in den letzten Jahrzehnten auf der Auswertungsseite erhebliche Fortschritte gemacht, die wir im folgenden skizzieren. Die veröffentlichten Beobachtungsstudien haben diese Entwicklung oftmals noch nicht rezipiert. Susman et al. (1976, nach Hollenbeck, 1978) untersuchten Beobachtungsstudien in 15 Zeitschriften aus entwicklungspsychologischen, klinischen und pädagogischen Bereichen und stellten fest: 32% der Arbeiten berichteten überhaupt keine Reliabilitätsangaben. Fast alle Schätzungen, die referiert wurden, waren Angaben über prozentuale Übereinstimmung (s.u.), und während der 16 Jahre, die sie diese Zeitschriften verfolgten, fanden die Autoren keine Verbesserung.

Um die große Zahl der Analyseverfahren zu ordnen, *gliedern wir nach dem Skalenniveau*, das den Beobachtungen als Daten zugeschrieben wird. Denn man wird Reproduzierbarkeit nur hinsichtlich jener Datenmerkmale fordern, die man in den Daten als vorhanden annimmt; wenn also nur Rangskaleninformation in den Daten vorhanden ist, wird man nicht auf Übereinstimmung in den Intervallgrößen prüfen (allgemeine Übersichtsliteratur: Asendorpf & Wallbott, 1979; Frick & Semmel, 1978; Landis & Koch, 1975; Lienert, 1973, Kap. 9; Tinsley & Weiss, 1975; Computerprogramme z.B. Cicchetti et al., 1977).

5.1 Übereinstimmungsmaße für nominalskalierte Daten

5.1.1 Prozentuale Übereinstimmung und allgemeine Vorüberlegungen

Von Übereinstimmung sprechen wir, wenn das Beobachtete identischen Kategorien zugeordnet wird.

Zu den in der Literatur sehr häufig berichteten und in der Regel auf dichotome Daten angewandten Übereinstimmungsmaßen gehört der *Prozentsatz*, mit dem (zwei) Beobachter das gleiche Material in die gleichen Kategorien ordnen. Dabei nimmt man implizit an: Die Übereinstimmung bei der Beurteilung des einen Beobachtungsgegenstandes ist genauso (positiv) zu werden wie die Übereinstimmung bei jedem beliebigen anderen und die abweichende Klassifikation bei dem einen Material ist ein gleichschwerer Fehler wie jede andere Nichtübereinstimmung. Unterscheidet sich dann die Materialstichprobe bei der Erprobung des Kategoriensystems in ihrer Klassifikationsschwierigkeit von der Materialstichprobe der eigentlichen Erhebung (z.B. Videotraining vs.

Feldstudie), dann kann man nicht gleiche Reliabilitätskoeffizienten erwarten. Dies gilt auch für den Austausch von Beobachtern, wenn man Unterschiede in der Klassifikationsgüte erwarten muß. Sofern Reliabilitätsstudien weder Material- noch Beobachterparameter berücksichtigen, sind die Grenzen der Verallgemeinerung der Reproduzierbarkeitsangaben unklar.

Tabelle 2: Daten nach Hollenbeck (1978).

Beobachter	Kategorie	Zeitblöcke										
1		1	2	3	4	5	6	7	8	9	10	Σ
	A	0	1	1	0	0	1	0	0	1	1	5
	B	1	0	1	0	0	0	1	0	0	0	3
	C	0	1	1	0	1	0	1	0	1	0	5
2												
	A	0	1	0	0	1	1	1	0	1	1	6
	B	1	0	0	1	0	1	0	0	0	0	3
	C	1	1	1	0	1	0	0	0	1	0	5

Legende: 1 = Verhalten in der entsprechenden Kategorie wurde beobachtet,
0 = sonst.

Hollenbeck (1978) diskutiert die Problematik von Übereinstimmungsprozent-sätzen an den Beobachtungen in Tab. 2, wobei für „Zeitblöcke“ jede Art von Wiederholung stehen kann, beispielsweise auch „Vpn“ oder „Situationen“. - Zunächst ergeben sich nun mehrere Möglichkeiten, prozentuale Übereinstimmung (% $\ddot{U}$) zu berechnen, z.B. für jede Kategorie einzeln, für den Durchschnitt, und für vollständige Übereinstimmung pro Zeitblock:

Kategorie	Anzahl konkordanter Paare	% $\ddot{U}$
A	7 von 10	70
B	6 von 10	60
C	8 von 10	80
Durchschnitt	7 von 10	70
vollständig pro Zeitblock	4 von 10	40

Weitere Möglichkeiten entstehen, wenn man Gewichte für verschiedene Arten von Konkordanz oder Diskordanz einführt (z.B. Übereinstimmung bei positiver Identifikation des Verhaltens sei gewichtiger zu veranschlagen als bei negativer, s.u.), oder statt wie im Beispiel 0/1 zu registrieren, feststellt, wie oft das Verhalten im jeweiligen Zeitblock auftrat, oder die Sequenz berücksichtigt, in der Verhaltensweisen innerhalb eines Zeitintervalles beobachtet wurden.

Die Kritik an % $\bar{U}$ faßt Hollenbeck in folgenden Punkten zusammen: 1. Die Höhe des möglichen % $\bar{U}$ hängt von den Randsummen (letzte Spalte in Tab. 2) ab. Je unterschiedlicher die Randsummen sind, desto geringer ist die maximal mögliche Übereinstimmung. Dies ist solange wünschenswert, wie sich in unterschiedlichen Randsummen ausschließlich Fehler der Beobachter niederschlagen können, etwa zu geringe Sensibilität jener Beobachter mit niedrigen positiven Randsummen. Wie jedoch Randsummen-Differenzen zu deuten sind, läßt sich nur aus der jeweiligen Studie bestimmen, und eine Anwendung der signal detection theory (Green & Swets, 1966, einführend: McNicol, 1972) wäre möglich, wenn man Sensitivitätsunterschiede und unterschiedliche Reaktionstendenzen prüfen möchte 2. Die *Bewertung* des berechneten % $\bar{U}$ hängt von den Randsummen ab, denn aus ihnen läßt sich bestimmen, wie groß die rein zufällig zu erwartende Übereinstimmung ist (s.u.). Für Kategorie A in Tab. 2 ergibt sich als *Erwartung* aufgrund der Randsummen:

		Beobachter 2		
		0	1	
Beobachter 1	0	2	3	5
	1	2	3	5
		4	6	10
		erwartet		

	3	2	
	1	4	
	beobachtet		

Also sind $\frac{4 \times 5}{10} + \frac{5 \times 6}{10} = 5$ Übereinstimmungen aufgrund der Randsummen zu erwarten, somit ist die beobachtete prozentuale Übereinstimmung von 70% genau 20% besser als die Zufallserwartung von 50%, bei Kategorie B ist sie nur 2% besser, bei C 30% besser als der Zufall. 3. % $\bar{U}$ informiert nicht darüber, *wo die Fehler liegen* und welcher Art sie sind. 4. Das durchschnittliche % $\bar{U}$ über alle Kategorien kann - besonders, wenn es nicht extrem hoch ist und über viele Kategorien zusammengefaßt wird - drastische *Unterschiede zwischen den einzelnen Kategorien* verdecken. Wenn inhaltliche Schlüsse auf Beobachtungen in einzelnen Kategorien basieren, muß die Verlässlichkeit für

diese Kategorien bekannt sein. 5. Wenn die Gelegenheiten zur Wiederholung der Beobachtungen als Zeitintervalle geplant werden, variiert % Ü als *Funktion der Länge dieses Intervalles*: Je länger, desto größer die vielleicht nur scheinbare Übereinstimmung, wenn nur das bloße - mindestens einmalige - Auftreten des Verhaltens registriert wird.

Zu den wichtigsten Einsichten der Reliabilitätsmethodik gehört die Erkenntnis, daß Übereinstimmung nicht durch jedes Zusammenhangs- oder Assoziationsmaß beschrieben werden kann. Das allgemeine statistische Problem, den Zusammenhang zwischen zwei oder mehr nominalskalierten Variablen zu messen, wurde ausführlich von Goodman & Kruskal (1954, 1959, 1963, 1972) erörtert (Diskussion dieser Arbeiten bei Bishop et al. 1974, McKinlay, 1975). Übereinstimmung ist jedoch ein Spezialfall von Zusammenhang, wie aus folgendem Beispiel hervorgeht:

		1. Beobachter		
		A	B	C
2. Beobachter	Kategorie A	0	50	0
	Kategorie B	0	0	50
	Kategorie C	50	0	0

Der Zusammenhang ist perfekt, die Übereinstimmung jedoch gleich Null. Damit Übereinstimmung vorliegt, müssen die Fälle in identische Kategorien geordnet werden, für Zusammenhang genügt eine solche Zuordnung, welche die Vorhersage der Klassifikation durch den einen Beurteiler aus der Klassifikation eines anderen Beurteilers gestattet.

5.1.2 Systematik einiger Übereinstimmungsmaße **für** nominalskalierte Daten

Eine weitere, für die Reliabilitätsmethodik wesentliche Einsicht besteht in der Erkenntnis, daß Übereinstimmung sich auch zufällig einstellen kann, wenn Beobachter gar nicht beobachten und so gewonnene Information zur Grundlage ihres Urteils machen, sondern blind - z.B. durch Münzwurf - kategorisieren, oder wenn sie aufgrund verschiedener, voneinander unabhängiger Kriterien zu ihren Resultaten kommen. Wenn jeder von zwei Beurteilern jede von zwei Kategorien gleich oft verwendet, ist zu erwarten, daß sie durchschnittlich in der Hälfte der Fälle übereinstimmen werden. *Die beobachtete Übereinstimmung sollte also in irgendeiner Form gegenüber der zufällig zu erwartenden Übereinstimmung relativiert werden.* Schutz (1952) war in dieser Forschungs-

richtung wohl der erste, der ein Modell für den Urteilsprozeß mit einer Zufallskomponente ausdrücklich formulierte. Er berechnet den Übereinstimmungsprozentsatz, der beobachtet werden muß, damit der Forscher auf einem vom ihm gewählten Konfidenzniveau davon ausgehen kann, daß ein von ihm als mindestens notwendig erachtetes Übereinstimmungsniveau, das frei von Zufallsübereinstimmung ist, tatsächlich erreicht wurde.

Das Zufallsmodell, das Schutz seiner Korrektur zugrunde legte, ist etwas kompliziert und trifft wenig plausible Annahmen, z.B. daß kein Beobachter, wenn er im Zustand der zufälligen Beurteilung ist, eine Präferenz für eine der Kategorien hat. Scott (1955) legte m. W. als erster einen Koeffizienten π vor, in dem die Zufallskorrektur direkt vorgenommen wird, und zwar als Subtraktion der zufälligen Übereinstimmung P_e von der beobachteten Übereinstimmung P_o . Die Differenz $P_o - P_e$ wird dann standardisiert mit Hilfe einer Division durch $1 - P_e$ (das ist der maximal mögliche Wert, den die Differenz annehmen kann), wodurch erreicht wird, daß π bei vollständiger Übereinstimmung = + 1; bei Übereinstimmung, die nicht über die Zufallserwartung hinausgeht, ist $\pi = 0$. Somit

$$\pi = \frac{P_o - P_e}{1 - P_e}$$

Für den Fall von 2 Beobachtern und einer dichotomen Kategorie führen wir folgende Notation ein:

		2. Beobachter		
		1	0	
1. Beobachter	1	a	b	p_1
	0	c	d	q_1
		p_2	q_2	1

a, b, c, d sind die relativen Häufigkeiten in den Zellen, die Randsummen sind ebenfalls als Proportionen dar gestellt. P_o wird als % Ü, also = $a + d$ berechnet. P_e wird berechnet als $\left(\frac{p_1 + p_2}{2}\right)^2 + \left(\frac{q_1 + q_2}{2}\right)^2$, was inhaltlich dem Versuch entspricht, die zufällige Übereinstimmung auf der Basis von Populations-Randwahrscheinlichkeiten und der Annahme zu berechnen, beide Beobachter wiesen die gleiche Randverteilung auf, die der in der Population entspräche (Light, 1971, S. 367, zur Kritik an Scott s. auch Lisch & Kriz 1978).

Cohen (1960) führte einen inzwischen sehr bekannt gewordenen Koeffizienten κ ein, der sich - abweichend von π - auf die beobachteten Randverteilungen bezieht und nicht davon ausgeht, die Randverteilungen seien gleich. Das Rationale für die Zufallskorrektur ist also das gleiche wie bei π , jedoch berechnet

sich hier P_e als $p_1p_2 + q_1q_2$. Light (1971) erweiterte κ , und zwar sowohl für mehr als zwei Beobachter, ein „overall group agreement measure“ (neuere Übersicht hierzu bei Conger 1980) als auch für ein zu den Randverteilungen konditionales übereinstimmungsmaß (Programm: McDermott & Watkins, 1979). Ferner berichtet er einen Signifikanztest für die gemeinsame Übereinstimmung mehrerer Beobachter, wenn sie mit einem Standard - z.B. einer „richtigen“ Klassifikation - oder wenn ein bestimmter Beobachter mit allen anderen verglichen werden sollen (Programm: Watkins & McDermott, 1979). Wackerly et al. (1978) haben diese Thematik für den Fall des Vergleichens eines Beobachters mit einem bekannten Standard weiterentwickelt. Sie unterscheiden zwei Fälle: Dem Beobachter wird die Randverteilung vorgegeben, oder er kann sie selbst bestimmen. Im ersten Fall bestehen Abhängigkeiten zwischen den Urteilen, die man für die Berechnung der Zufallsübereinstimmung berücksichtigen muß. Wackerly et al. berichten eine Möglichkeit zur inferenzstatistischen Prüfung von κ auf Überzufälligkeit auch unter der Annahme intraindividuelle Abhängigkeit der Urteile eines Beobachters.

Um die *Rolle der Randverteilungen* für die Wahl eines Übereinstimmungsmaßes zu verdeutlichen, vergleichen wir zwei Fälle:

Fall I

1. Beobachter

1 0

1	60	20	80
0	20	0	20

80 20

2. Beobachter

Fall II

1. Beobachter

1 0

1	20	60	80
0	0	20	20

20 80

Im ersten Fall scheint die Übereinstimmung mit 60% $\ddot{U}$ größer zu sein als in Fall II mit 40% $\ddot{U}$. Geht man jedoch von den Restriktionen durch die vorgegebenen Randsummen aus, so zeigt sich in I die geringstmögliche, in II die größtmögliche Übereinstimmung. Die deutliche Asymmetrie in der Kategorienbenutzung schafft Probleme für die Interpretation der Übereinstimmung; Light nennt % $\ddot{U}$ die absolute, die randverteilungskorrigierte Übereinstimmung die relative. Wieder hängt es von den Umständen in der jeweiligen Untersuchung ab, ob man in das Übereinstimmungsmaß eine Korrektur für abweichende Randverteilungen einbeziehen soll.

Als Erweiterung hat Cohen (1968) das gewichtete κ eingeführt, das es erlaubt, Nichtübereinstimmung nach der Schwere der Folgen zu gewichten, wobei der Forscher die Gewichte festlegen muß.

Fleiss (1971) verallgemeinerte κ für die Situation, daß jeder Beobachtungsfall auf Nominalskalenniveau von der gleichen Anzahl Beurteiler eingestuft wird, aber die Beurteiler, die den einen Fall beobachten, nicht notwendigerweise die gleichen sind, die einen anderen Fall beobachten. Fleiss schildert auch die Möglichkeit zu bestimmen, wie hoch die Übereinstimmung darüber ist, eine bestimmte Person oder Verhaltensweise einer bestimmten Kategorie zuzuordnen. Fleiss et al., 1972, behandeln den Fall, daß eine Person durch mehr als eine Variable beschrieben wird. - Wenn bei der Berechnung der Übereinstimmung die zeitliche Abfolge berücksichtigt werden soll, also Sequenz und jeweilige Dauer der beobachteten Ereignisse, ergeben sich besondere Schwierigkeiten, auf die Hollenbeck (1978) eingeht, der κ auch in diesem Fall anwendet (s.a. Asendorpf & Wallbott, 1979, sowie Abschn. 5.4). - Um prüfstatistische Fragen behandeln zu können, wurde für κ und gewichtetes κ eine Stichprobentheorie entwickelt (Everitt 1968, Fleiss et al., 1969, Übersicht bei Hubert 1977, für kleine Stichproben Wackerly et al., 1978, man beachte die korrigierten Formeln 'in Fleiss et al., 1979). Die inferenzstatistischen Ansätze gehen durchweg davon aus, daß die wiederholten Beobachtungen eines Beobachters oder verschiedener Beobachter voneinander unabhängig sind. Diese Annahme ist prinzipiell empirisch prüfbar, beispielsweise durch den Vergleich der Übereinstimmung zwischen einzelnen arbeitenden Beobachtern und solchen, die schon dadurch interagieren, daß sie die Beobachtung im gleichen Raum durchführen. Bisweilen führt schon die Planung der Erhebung dazu, daß die aus statistischen Gründen geforderte Unabhängigkeit der Beobachtungen kaum zu erwarten ist, beispielsweise, wenn den Beobachtern gestattet wird, ihre Beurteilung des früher auftretenden X nach Beobachtung des später auftretenden Y zu ändern. Unabhängigkeit der Beobachtungen kann auch bedeuten: Die Ergebnisse der Beobachtungen bei Wiederholungen sind nicht eine Funktion der Wiederholungen. Das erleichtert auch die inhaltliche Interpretation der Befunde, denn im Idealfall ändert wiederholte Beobachtung weder das Beobachtete noch das Beobachtungsinstrument.

Die meisten Übereinstimmungsmaße gehen davon aus, daß die Beobachter die gleichen, vor der Beobachtung definierten Kategorien benutzen. Man kann es jedoch auch den Beobachtern freistellen, welche und wieviele Kategorien sie definieren. Beispiele wären: Beurteiler verwenden verschiedene nosologische Klassifikationen, Lehrer verschiedene Notensysteme, Vpn beschreiben ihr Lösen der gleichen Denkprobleme mit unterschiedlichen Begriffen. Dann muß man dennoch nach Brennan & Light (1974) nicht darauf verzichten, Übereinstimmung zu berechnen. Die Idee besteht darin, Paare von Beobachtungsergebnisse zu betrachten: Ordnen von zwei Beobachtern der eine zwei Beobachtungen in die Kategorie „flüssiger Verhaltensablauf“, der andere beide Beobachtungen in seine Kategorie „geschickter Bewegungsvollzug“, so erfassen beide Kategorien Gleiches trotz inhaltlich verschiedener Bezeichnungen, und eine übereinstimmende Klassifikation eines Paares in dieselbe Kategorie stellt

einen Hinweis auf Übereinstimmung der Beobachter dar wie auch der Fall, daß beide die Elemente eines Paares von Beobachtungen in jeweils andere Kategorien einordnen. Hubert (1977a) leitet Mittelwert und Varianz des Koeffizienten von Brennan & Light ab und ermöglicht so die Prüfung auf signifikante Abweichung von Null und die Bestimmung von Konfidenzintervallen. Hubert zeigt auch, wie dieser Koeffizient auf den Fall geordneter Kategorien ausgedehnt werden kann.

Es gibt inzwischen eine große Zahl von Übereinstimmungskoeffizienten, von denen hier noch der G-Index erwähnt werden soll, den Holley & Guilford (1964) zunächst als Maß der Ähnlichkeit von zwei Personen, charakterisiert über n dichotome Items eingeführt haben. Holley & Lienert (1974) beschreiben ihn als Übereinstimmungsmaß mit einer typischen Anwendung, daß zwei oder mehr Beobachter eine oder mehrere Personen hinsichtlich n dichotomer Merkmale charakterisieren. Weitere Verallgemeinerungen finden sich bei Vegelius (1977, 1977a, 1979) und in der dort referierten Literatur. - Die große Zahl von übereinstimmungsmaßen hat zu statistischen Arbeiten geführt, die die Koeffizienten vergleichen und ihre impliziten Annahmen systematisieren (z.B. Fleiss & Cohen, 1973; Janson & Vegelius, 1979; Hubert, 1979b). Krippendorff (1970a) geht von folgender, häufig verwendeter Formel für Übereinstimmungskoeffizienten aus:

$$(1) \quad \text{Übereinstimmung} = 1 - \frac{\text{beobachtete Nichtübereinst.}}{\text{erwartete Nichtübereinst.}}$$

Der Koeffizient wird 0, wenn die Übereinstimmung rein zufällig ist, 1 wenn sie vollkommen ist, negativ, wenn sie hinter der Zufallserwartung zurückbleibt. Da nur Übereinstimmung, nicht Zusammenhang allgemein interessiert, müssen die Zellen einer Frequenzmatrix für die Beobachter i und j bei der Berechnung der Nichtübereinstimmung gewichtet werden mit d_{ij} , wobei für ungeordnete Kategorien gilt:

$$(2) \quad d_{ij} = \begin{cases} 0, & \text{wenn } i = j, \\ 1, & \text{wenn } i \neq j. \end{cases}$$

Wenn die Kategorien geordnet sind oder durch Skalenwerte mit Intervallskalenniveau repräsentiert werden können (s.u.), wird folgende Gewichtungsfunktion vorgeschlagen:

$$(3) \quad d_{ij} = (i - j)^2, \text{ mit } i, j \text{ als Kennwerten der Kategorien.}$$

Die erwarteten Frequenzen (e_{ij}) werden entweder bestimmt unter Annahme gleicher Randverteilungen:

$$(4) \quad e_{ij} = \frac{1}{n} \left(\frac{n_{i.} + n_{.j}}{2} \right) \left(\frac{n_{i.} + n_{.j}}{2} \right)$$

wobei die folgende Frequenzmatrix die Notation definiert:

		2. Beobachter	
		Kategorie	j
1. Beobachter	i	n_{ij}	$n_{i.}$
	$n_{.j}$		$n = \text{Gesamtzahl der beurteilten Einheiten}$

Dabei bedeutet n_{ij} die Anzahl der Fälle in Zeile i und Spalte j. Ein Punkt im Index zeigt an, daß über die entsprechende Zeile oder Spalte summiert wurde.

Unter Berücksichtigung ungleicher Randverteilungen ergibt sich:

$$(5) \quad e_{ij} = \frac{1}{n} n_{i.} n_{.j} .$$

Wenn als Maß für die beobachtete Nichtübereinstimmung

$$\sum \sum n_{ij} d_{ij}$$

definiert wird und für die zufallsbedingte Nichtübereinstimmung

$$\sum \sum e_{ij} d_{ij} ,$$

so erhält (1) die Form:

$$(6) \quad a = 1 - \frac{\sum \sum n_{ij} d_{ij}}{\sum \sum e_{ij} d_{ij}} .$$

Scotts π erhält man durch Einsetzen von (2) und (4) in (6), Cohens κ durch Einsetzen von (2) und (5) in (6), den Intraklassen-Korrelationskoeffizienten nach Pearson (1901) durch Einsetzen von (3) und (4) in (6) und Spearmans Rangkorrelationskoeffizienten ρ durch Einsetzen von (3) und (5) in (6). Nach Krippendorffs Systematik hat der Forscher also zwei Auswahlentscheidungen zu treffen, ob Unterschiede in den Randverteilungen zum Fehler geschlagen werden sollen oder nicht, und ob die Kategorien eine Ordnung aufweisen oder nicht. In der systematischen Übersicht von Fleiss (1975) kommt als weiterer

Klassifikationsgesichtspunkt die Frage hinzu, ob ein Übereinstimmungsmaß die Proportionen der Übereinstimmung über das Vorliegen eines Merkmals und der Übereinstimmung über Nichtvorliegen symmetrisch behandeln oder nicht. Wenn, wie bei taxonomischen Fragestellungen nicht selten, positive Merkmalsausprägung sehr selten, Fehlen des Merkmals jedoch häufig auftritt, kann es sinnvoll sein, in einem Übereinstimmungsmaß nur die Fälle zu berücksichtigen, in denen von wenigstens einem Beobachter positive Merkmalsausprägung festgestellt wurde (so z.B. Dice, 1945; weiteres bei Fleiss, 1975). Schließlich unterscheiden sich Übereinstimmungsmaße danach, ob in ihnen eine Korrektur für Zufallsübereinstimmung vorgenommen wird oder nicht (s.o.), und Fleiss weist darauf hin, daß die Art dieser Korrektur in π und κ nur eine von mehreren möglichen ist.

Einen etwas anderen Akzent setzen Landis & Koch (1977, 1977a). Sie schätzen nicht nur die Größe von Übereinstimmungen, sondern prüfen an einem oder simultan mehreren Datensätzen mehrere Hypothesen über das Zustandekommen von Daten einer Reliabilitätsstudie mit Hilfe eines allgemeinen Ansatzes zur inferenzstatistischen Analyse von multivariaten kategorialen Daten. Dieses Vorgehen ermöglicht es u.a., gezielt die Ursachen fehlender Übereinstimmung zu identifizieren, wenn z.B. detailliert festgestellt wird, die Beobachter A und B unterschieden sich signifikant im Gebrauch der Kategorien X und Y bei der Teilstichprobe s (vgl. auch Bergen 1980). Abschließend sei auf Trippi & Settle (1976) hin gewiesen, die eine nichtparametrische Variante der internen Konsistenz entwickeln, sowie auf Kaye (1980), dessen Ansatz besonders für die Auswertung *sequentieller* Beobachtungen relevant ist.

Man kann die Frage aufwerfen, welche Relevanz diese Vielzahl von Möglichkeiten für die Forschungspraxis hat, ob also die Unterscheidungen zwar von theoretischem Interesse sind, jedoch - insbesondere bei großen Stichproben - die Schlußfolgerungen gleich sind, die aus verschiedenen Übereinstimmungsmaßen gezogen werden. Zu dieser Frage ist uns keine Literatur bekannt, so daß pragmatische Erwägungen hinter dem Versuch zurückstehen müssen, die theoretisch angemessene Form der Prüfung einzusetzen.

5.2 Übereinstimmungsmaße für ordinalskalierte Daten

Wenn Beobachter ihre Ergebnisse als Rangreihen mitteilen oder Forscher Beobachtungen als Rangreihen auffassen, dann läßt sich nicht prüfen, ob Beobachter in ihren Urteilen über die absolute Merkmalsausprägung übereinstimmen; Unterschiede in Skaleneinheit und Skalenursprung können sich bei allen im folgenden besprochenen Übereinstimmungsmaßnahmen nicht mindernd auswirken. Vielmehr sind alle Transformationen der Skaleneinheiten zugelassen, welche die Ordnung der betrachteten Größen nicht verändern. Berechnet man die Übereinstimmung zwischen 2 Rangreihen, so läßt sich Abweichung

entweder beschreiben als Summe der quadrierten Differenzen, wie dies z.B. in Spearmans Rangkorrelationskoeffizient Q geschieht, oder als Summe der absoluten Differenzen der Wertepaare, so in Spearmans (1906) „footrule“ (vertiefte Darstellung in Hubert, 1979). Schon Spearman wies darauf hin, daß eine Quadrierung größere Abweichungen stärker gewichtet als kleinere. Wenn man erwartet, daß extremere Meßwerte mehr Meßfehler enthalten als mittlere, hat der Meßfehler bei Maßen wie Q einen stärkeren Einfluß auf die Bestimmung der Konsensstärke. Spearmans Q bezieht sich auf Rangdifferenzen, was trivialerweise impliziert, jeder Rangplatz sei von seinem jeweiligen Nachbarn gleich weit entfernt. Kruskal (1958) hat jedoch auf eine Interpretationsmöglichkeit von Q hingewiesen, die ohne diese Annahme auskommt: Wenn man Paare von Tripeln vergleicht (je ein Tripel stammt von einem Beobachter und stellt eine Rangordnung von drei Beobachtungen dar), läßt sich Q als Maß für proportionale Fehlerreduktion interpretieren.

Bei der Auswahl zwischen verschiedenen Maßen für ordinale Übereinstimmung kann man die Frage beachten, wie Rangplatzbindungen (ties) behandelt werden (zur gesamten Diskussion s. Hildebrand et al., 1977). In dem von Goodman & Kruskal (1954) vorgeschlagenen γ werden alle Rangplatzbindungen nicht beachtet, während sie in Somers' d_{xy} berücksichtigt werden. Wilson (1974) schließt nur die Beobachtungspaare aus, die auf beiden Variablen rangplatzgebunden sind.

Will man die ordinale Übereinstimmung zwischen mehr als zwei Beobachtern ermitteln, so berechnet man üblicherweise den Konkordanzkoeffizienten W von Kendall (1948). Er wird Null bei maximaler Nichtübereinstimmung und Eins bei völliger Übereinstimmung, d.h. wenn Rangreihen identisch sind. Er läßt sich auf Überzufälligkeit prüfen. Bei m Beurteilern besteht folgende Beziehung:

$$\text{mittleres } Q = \frac{mW - 1}{m - 1} .$$

Ein niedriges W schließt nicht aus, daß die Gesamtmenge der Beobachter in 2 oder mehr Mengen unterteilt werden kann, die jeweils untereinander stark übereinstimmen. Während W die mittlere Übereinstimmung aller Rangreihen untereinander beschreibt, kann Konsens auch geprüft werden als Übereinstimmung mit einem Kriterium, d.h. einer vorgegebenen Rangordnung (Lyerly, 1952; Cureton, 1958, 1965; Taylor & Fong, 1963; Taylor, 1964). Unterschiede in der Konsensstärke zwischen unabhängigen Beobachtergruppen lassen sich inferenzstatistisch prüfen (s. Stewart et al., 1979, S. 310ff.). Schucany & Frawley (1973) legen ein Prüfverfahren für folgende Fragen vor: Besteht innerhalb einer Gruppe von Beobachtern genügend Übereinstimmung? und wenn ja: Besteht zugleich signifikante Übereinstimmung zwischen zwei Gruppen von Beurteilern? (Für Weinkenner sei in diesem Zusammenhang auf Amerine & Roessler, 1976; hingewiesen.)

5.3 Übereinstimmungsmaße für intervallskalierte Daten

5.3.1 Einfache varianzanalytische Ansätze und Intraklassen-Koeffizienten

Wir behandeln zunächst die von Ebel (1951) eingeführte varianzanalytische Auswertung von Reliabilitätsstudien einschließlich der dabei meistens vorgenommenen Bestimmung eines Intraklassen-Korrelationskoeffizienten (Über-sichten z.B. Haggard, 1958; Maxwell & Pilliner, 1968; Landis & Koch, 1975; Bartko, 1976; Kraemer & Korner, 1976; Werner, 1976; Asendorf & Wallbott, 1979). Wenn jeder von d Beobachtern einmal ein Verhaltensmerkmal an allen n Personen beobachtet, läßt sich folgendes *Modell* für das Zustandekommen einer Beobachtung J_{ij} an Person i durch Beobachter j aufstellen (hier nach Landis & Koch, 1975):

$$(1) \quad J_{ij} = \mu + s_i + e_{ij},$$

wobei μ den Gesamtmittelwert, also die mittlere Beobachtungsausprägung, s_i den Effekt des Merkmalsträgers i und e_{ij} den verbleibenden Residualwert oder Fehler wiedergibt. Für ein inferenzstatistisches Vorgehen muß man die Annahmen eines varianzanalytischen Modells mit Zufallseffekten treffen: Die n Personen stellen eine Zufallsstichprobe aus einer angegebenen Population dar; die s_i sind normalverteilt mit einem Mittelwert = 0 und der Varianz σ_s^2 ; die e_{ij} sind normalverteilt mit einem Mittelwert = 0 und der Varianz σ_e^2 ; die s_i und e_{ij} sind unabhängig voneinander. Selva (1976) diskutiert diese Annahmen und einige Auswege, wenn sie nicht erfüllt sind. Der Intraklassen-Koeffizient ist für (1) definiert als

$$(2) \quad Q = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \quad ,$$

also als Verhältnis der Varianz zwischen den Merkmalsträgern zur Gesamtvarianz. Um Begriff und Namen des Intraklassen-Koeffizienten zu verstehen, kann man von dem Beispiel ausgehen, in dem der korrelative Zusammenhang zwischen den IQs der Personen eines Zwillingspaars bestimmt werden soll.

Ein Produkt-Moment-Korrelationskoeffizient r beruht auf Paaren von Daten, je ein Datum wird dem Vektor X , das andere dem Vektor Y zugeordnet. Welchen Zwilling ordnet man X , welchen Y zu? r kann je nach Zuordnung erheblich schwanken. Werner (1976, S. 489): „Allgemein tritt dieses Problem immer dann auf, wenn Objekte zwar in Klassen einteilbar sind, aber innerhalb derselben nicht weiter unterschieden werden sollen, und man generell ermitteln will, ob Objekte *einer* Klasse einander ähnlicher sind als Objekte *verschiedener* Klassen.“

Modell 1 berücksichtigt nicht explizit Unterschiede zwischen den Beobachtern. Diese kann man als zufällige oder als fixierte Beobachtereffekte einführen. Faßt man die Beobachter als eine Zufallsstichprobe aus einer größeren Population potentieller Beobachter auf, so ergibt sich Modell

$$(3) \quad J_{ij} = \mu + s_i + d_j + e_{ij},$$

wobei d_j zusätzlich den Effekt des Beobachters j repräsentiert. Die d_j sind normalverteilt mit Mittelwert = 0 und Varianz σ_d^2 ; die s_i , d_j und e_{ij} sind voneinander unabhängig. Der diesem Modell entsprechende Intraklassen-Koeffizient ist

$$(4) \quad \rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2 + \sigma_e^2} \quad ,$$

wobei σ_d^2 die Varianz darstellt, die auf Unterschiede zwischen den Beobachtern zurückgeht, also „interobserver bias“ darstellt. Faßt man die Beobachter als fixiert auf und interessieren nur Aussagen über ihre Verlässlichkeit, weil gerade sie in weiteren Studien eingesetzt werden sollen, so ergibt sich ein (3) vergleichbares „mixed model“. Wenn die Varianz zwischen den Beobachtern ausgeklammert werden soll, gibt es zwei Möglichkeiten. Bei Ebel, Haggard und Winer (1971) findet sich die entsprechende Varianzkomponente weder im Zähler noch im Nenner der Reliabilitätsformel, bei Rajaratnam (1960) und Krippendorff (1970, s. Werner, 1976) bleibt die Gesamtvarianz im Nenner; für die sich dann ergebenden Schätzprobleme und für die Bestimmung von Konfidenzintervallen s. Lu (1971) sowie Fleiss & Shrout (1978). Werner weist darauf hin, daß der Intraklassenkoeffizient nur dann dem über alle Beobachterpaare gemittelten Produkt-Moment-Korrelationskoeffizienten gleich ist, wenn die Mittelwerte und Varianzen aller Beobachter gleich sind und bei der Berechnung des Intraklassen-Koeffizienten die Varianz zwischen den Beobachtern nicht zur Fehlervarianz geschlagen wird.

Eine ausführliche Ableitung und Begründung der varianzanalytischen Reliabilitätsprüfung für den univariaten Fall ohne Replikation der Beobachtung findet sich in Winer (1971, S. 283ff.). Winer verdeutlicht die beiden Entscheidungen, die der Forscher für seine Reliabilitätsaussage zu fallen hat: Er kann die Verlässlichkeit entweder einer typischen einzelnen oder der über alle Beobachter gemittelten Messung bestimmen wollen, und zwar in beiden Varianten entweder mit oder ohne Einbezug der Varianz zwischen den Beobachtern in die Fehlervarianz. Die Wahl zwischen diesen beiden Alternativen wird er begründen wollen, wozu er die folgenden Überlegungen heranziehen könnte: Die Ausgangsfrage ist stets, zu welchem praktischen oder wissenschaftlichen Zweck er die Beobachtungen verwendet. Hängt beispielsweise die Einführung von Produkten auf einen Markt von den Schätzurteilen mehrerer Experten ab, und werden diese Urteile zusammengefaßt, ist zweifellos die Reproduzierbar-

keit der durchschnittlichen Schätzwerte gefragt. Wenn der Beurteiler jedoch typischerweise einzeln arbeitet, wie z.B. Lehrer bei der Benotung ihrer Schüler, ist die Verlässlichkeit der individuellen Urteile von Interesse. Auch die Entscheidung über die Einbeziehung der Varianz zwischen den Beurteilern in die Fehlervarianz hängt vom Verwendungszweck ab. Immer dann, wenn Unterschiede zwischen den über alle Beobachtungen berechneten Mittelwerten der einzelnen Beurteiler nicht zu Unterscheiden in den Entschlüssen führen, zu denen diese Beobachtungen die Basis liefern, sollte die Varianz zwischen den Beurteilern nicht zur Fehlervarianz geschlagen werden. Dies sollte hingegen wohl geschehen, wenn -wie etwa bei der zentralen Vergabestelle in ihren Entscheidungen über Studienplätze - Noten verschiedener Schüler, die von verschiedenen Lehrern stammen, gegeneinander aufgerechnet oder Durchschnittswerte verglichen werden, die von verschiedenen Beurteilergruppen stammen.

Die detaillierte Information über die Reproduzierbarkeitsstruktur der Daten wird bei dieser Art Analyse erkaufte durch starke Modellannahmen: (1) der Meßfehler korreliert nicht mit dem wahren Wert; (2) die Stichprobe der beobachteten V_{pn} ist eine Zufallsstichprobe aus der Population der Personen, auf die Schlüsse gezogen werden sollen; (3) die Stichprobe der Meßinstrumente oder Beobachter, welche Replikationen bereitstellen, ist in Modell (3) eine Zufallsstichprobe vergleichbarer Instrumente oder Beobachter; (4) die Schätzung der Fehlervarianz basiert auf der Zusammenfassung der Varianzen innerhalb jeder V_p . Die Datenerhebungssituation muß auf die Bedingungen (2) und (3) zugeschnitten werden, andere Modellimplikationen lassen sich nach der Erhebung prüfen: Da der „wahre Wert“ über alle Beobachter konstant bleibt, muß die Korrelation zwischen den verschiedenen Beobachtern statistisch konstant sein; die entsprechende Varianz-Kovarianzmatrix ließe sich auf Homogenität prüfen.

Wenn jeder Beobachter r wiederholte Einschätzungen des gleichen Verhaltens an den gleichen Personen vornimmt, läßt sich (3) für die k -te Replikation erweitern auf

$$(5) \quad J_{ijk} = \mu + s_i + d_j + (sd)_{ij} + e_{ijk},$$

wobei $(sd)_{ij}$ die Interaktion zwischen Beobachter j und Merkmalsträger i darstellt und ein Maß dafür ist, wie sehr j bei der Beurteilung von i von seinem üblichen Reaktionsmuster abweicht. Will man nicht, wie alle bisher vorgeschlagenen Modelle, annehmen, die Varianzkomponente zu Lasten des Zufallsfehlers sei für alle Beobachter gleich, so kann man auf ein Modell von Grubbs (1948, 1973) zurückgreifen (s.a. Overall, 1968 für die Situation, in der n Personen auf m Bedingungen aufgeteilt und anschließend von zwei unabhängigen Beobachtern beurteilt werden). Beurteilt jeder Beobachter mehr als ein Merkmal, so kommt eine multivariate Erweiterung des varianzanalytischen

Ansatzes von Fleiss (1966) in Frage, allerdings insbesondere auch die später besprochenen Generalisierbarkeitsstudien (zur Diskussion, wie die Residualvarianz zu schätzen sei, s. Huck 1978).

Diesen Abschnitt abschließend sollen noch einige Varianten besprochen werden. Lawlis & Lu (1972) schlagen einen bei Tinsley & Weiss (1975) kritisch besprochenen Index vor, der es erlaubt, Übereinstimmung als fehlerfreien Fall festzulegen, aber auch als Abweichung bis zu einer, bis zu zwei, etc. Maßeinheiten der gewählten Skala. Finn (1970, 1972) weist darauf hin, daß bei varianzanalytischen Reliabilitätschätzungen Beobachtungen über mehr als eine Person oder Verhaltensweise vorliegen müssen, und die Varianz zwischen den Personen oder Verhaltensweisen muß ausgeprägt sein, damit sich ein genügend großer Koeffizient ergeben kann. Situationen sind jedoch nicht selten, in denen nur eine Gegebenheit, diese jedoch mehrfach beobachtet wird, und für diesen Fall ist Finns Methode gedacht (zur Kritik insbesondere an Finns Inferenzstatistik s. Tinsley & Weiss, 1975).

Dem varianzanalytischen Modell folgend hat Krippendorff (1970) einen Ansatz vorgelegt, der insbesondere die während der Konstruktionsphase von Kategoriensystemen wichtige Frage nach Art und Größe der einzelnen Quellen für fehlende Verlässlichkeit zu beantworten gestattet. Globale Gesamtwerte geben keine Information über mögliche Schritte zur Verbesserung. - Detaillierte Informationen können erwünscht sein über

- (1) die geschätzte Reliabilität einer Datenmenge über alle Datenquellen. Man könnte diese Schätzung *Datenreliabilität* nennen und als Maß für das generelle Vertrauen in der Daten interpretieren;
- (2) das geschätzte Ausmaß, in dem sich die Datenreliabilität verbessern ließe, wenn die Urteile einiger Beurteiler transformiert oder neu definiert würden. Diese Schätzung erfaßt den *systematischen Fehler* des Verfahrens und ergibt zusammen mit dem **Zufallsfehler** den Betrag, der bis zu einer völligen Datenreliabilität fehlt;
- (3) für jeden einzelnen Beobachter das geschätzte Ausmaß seiner Verlässlichkeit, also die *individuelle Reliabilität*. Mit dieser Information ließe sich feststellen, ob - und, wenn ja, welche - einzelne Beobachter besonders unzuverlässige, d.h. hier: vom allgemeinen Trend der Beobachtergruppe abweichende Werte liefern. Je nach Fragestellung könnte dann die Schulung dieser Beobachter verbessert oder die Daten dieser Beobachter könnten nicht ausgewertet werden;
- (4) für jeden einzelnen Beobachter eine Schätzung des Ausmaßes, in dem seine Beobachtungen durch Schulung oder Transformation korrigierbar sein würden. Geschätzt werden müßte also der *systematische individuelle Beobachterfehler*, der zusammen mit dem *individuellen Zufallsfehler* die individuelle Unreliabilität ausmacht;
- (5) für jede einzelne Kategorie (recording unit) eine Schätzung des Ausmaßes,

in dem eine Zufallsstichprobe von Beobachtern bei ihren Beobachtungen zu dieser Einheit übereinstimmt. Diese Schätzung der *Reliabilität einer Kategorie* ermöglicht die gezielte Modifizierung oder Eliminierung einzelner Kategorien eines Systems.

Das Verfahren von Krippendorff erlaubt, diese Fragen zu beantworten. Sommerbe (o.J.) hat auf Druckfehler in der Arbeit von Krippendorff hingewiesen und ein Computerprogramm bereitgestellt.

5.3.2 Generalisierbarkeitsstudien

Generalisierbarkeitsstudien (Cronbach et al. 1963, 1972; Gleser et al. 1965) verallgemeinern den varianzanalytischen Ansatz; man kann diese Studien als Anwendung der Prinzipien multifaktorieller und multivariater Varianzanalysen auf die Reliabilitätsproblematik auffassen. Alle Generalisierbarkeitskoeffizienten sind Intraklassen-Koeffizienten; sie sind Maße der Verallgemeinerbarkeit einer Beobachtung über verschiedene Entstehungsbedingungen. Die Generalisierbarkeitstheorie stellt insofern gegenüber der Reliabilitätskonzeption in der klassischen Testtheorie eine „Liberalisierung“ dar (s. Kristof in diesem Band), als nicht gefordert wird, die Beobachtungen unter verschiedenen Bedingungen müßten „parallel“ sein, d.h. gleiche Mittelwerte, Varianzen und Kovarianzen aufweisen.

Für die Untersuchung der Reproduzierbarkeit von Beobachtungen kann man folgende Punkte als wesentliche Fortschritte gegenüber dem einfachen varianzanalytischen Ansatz herausstellen: (1) die ausdrückliche Festlegung der Facetten, auf die verallgemeinert werden soll, (2) die Hinweise auf den Zusammenhang zwischen Aussageabsicht und Definition des Meßfehlers, und (3) die ausdrückliche Unterscheidung zwischen Generalisierbarkeits- oder G-Studien und Entscheidungs- oder D-Studien, womit der Zweck, den die Reliabilitätsanalyse haben soll, thematisiert wird, nämlich eine wissenschaftliche oder praxisbezogene Entscheidung aufgrund der Beobachtungen zu fällen.

Cronbach et al. (1972, S. 15) formulieren den Grundsatz folgendermaßen: „Der Meßwert, auf den die Entscheidung sich stützen soll, ist nur einer von vielen Meßwerten, die den gleichen Zweck erfüllen könnten. Derjenige, der die Entscheidung zu treffen hat, interessiert sich fast nie für die Reaktion, die auf ein ganz bestimmtes Reizobjekt oder auf die speziellen Fragen, gegenüber einem ganz bestimmten Versuchsleiter oder in dem einen besonderen Augenblick der Testdurchführung gegeben werden. Wenigstens einige dieser Meßbedingungen könnten variiert werden ohne den Meßwert weniger akzeptabel für denjenigen zu machen, der die Entscheidung zu fällen hat. D.h., es gibt ein Universum von Beobachtungen, und jede dieser Beobachtungen hätte eine brauchbare Basis für die Entscheidung abgegeben. Der ideale Wert, auf den

man die Entscheidung basieren könnte, wäre der Meßwert einer Person, gemittelt für alle akzeptierbaren Beobachtungen, den wir ihren ‚Universalwert‘ (Universe score) nennen wollen. Der Untersucher benutzt den beobachteten Wert oder eine Funktion des beobachteten Wertes als ob das der Universalwert wäre. Das heißt, er generalisiert von der Stichprobe auf das Universum.“ *Die Frage nach der ‚Reliabilität‘ entwickelt sich somit zu einer Frage nach der Genauigkeit der Generalisation oder der Verallgemeinerbarkeit.* Wenn man die Verlässlichkeit von Beobachtungen prüfen will, wäre es günstig, möglichst im vollständig überkreuzten Versuchsplan die Bedingungen Beobachter, Kategorien, Personen und Zeitpunkte einzuführen. Dann läßt sich auch die von Kraemer & Korner (1976) beklagte Vermischung von interindividuellen Unterschieden, Stabilität des Merkmals, intraindividuelle Konsistenz jedes Beobachters und interindividuellen Unterschieden zwischen verschiedenen Beobachtern entmischen. Ein oft erwünschtes und einfaches Ergebnis läge dann vor, wenn der Haupteffekt zu Lasten der Beobachter und alle Interaktionen, an deren Definition die Beobachter beteiligt sind, keine nennenswerte Varianz auf sich vereinigen. Dann könnte man annehmen, es bestünden keine Unterschiede zwischen den Beobachtern in ihren allgemeinen Reaktionstendenzen oder in ihren Skalenverankerungen, sie benützten die Kategorien gleichartig, beurteilten die Personen nicht idiosynkratisch und schwankten in ihrem Beobachtungsverhalten nicht über verschiedene Zeitpunkte (zur Verbindung von Kosten-Nutzen-Analysen bei der Planung von Beobachtungsstudien mit Generalisierbarkeitsanalysen s. Gleser et al. 1965; für die Weiterentwicklung der multivariaten G-Studien z.B. auf Profilreliabilität oder Verallgemeinerbarkeit von gain scores s. Joe & Woodward, 1976; Anwendungsbeispiele: Rudinger & Feger, 1970; Levy, 1974; Feger, 1978; Weiterentwicklung bei Nußbaum, 1980).

5.3.3 Pfadanalytische Modelle für die Reliabilitätsprüfung

Noch „liberaler“, d.h. allgemeiner gegenüber dem einfachen varianzanalytischen Ansatz als die Generalisierbarkeitstheorie ist die pfadanalytische Prüfung von Modellen der Datenstruktur, die Annahmen über den Meßfehler einschließt. Der Ansatz von Jöreskog erfordert nicht einmal die Annahme, verschiedene Beurteiler würden gleiche Meßeinheiten benutzen (Werts et al., 1974). Wesentlicher jedoch als die Flexibilität scheint mir die Möglichkeit zu sein, Voraussetzungen prüfen zu können, die den verschiedenen Ansätzen, auch dem Intraklassen-Koeffizient, zugrunde liegen. So läßt sich die Annahme prüfen, allen Messungen liege der gleiche wahre Wert τ zugrunde. Läßt sich diese Annahme nicht halten, dann ist die Bedeutung einer varianzanalytischen Reliabilitätsschätzung unklar. Möglicherweise liegt nicht nur ein „wahrer“ Faktor zugrunde, und die Annahmen über die Unabhängigkeit der Fehler können falsch sein. Dann läßt sich weiter prüfen, ob die Annahme sowohl der

Intraklassen-Korrelation als auch der Generalisierbarkeitstheorie gilt, die Maße verfügten alle über die gleiche Meßeinheit, d.h. „essentially tau equivalent“ (Lord & Novick, 1968, S. 50) sind. „If this hypothesis is rejected then the Anova formulation is rejected whether used for estimating reliability or for generalizability procedures.“ (Werts et al., 1974, S. 29). Weiter läßt sich die im Intraklassen-Koeffizient implizierte Annahme prüfen, die Meßfehler der einzelnen Messungen seien unabhängig.

Für die pfadanalytische Behandlung der Reliabilitätsproblematik sind folgende Schritte charakteristisch: 1. Man nimmt an, jede Messung sei aus einem wahren Wert und einem Meßfehler zusammengesetzt. 2. Man legt die Art des Zusammenhanges zwischen wahren Wert und Meßfehler jeder Messung mit wahren Wert und Meßfehler jeder anderen Messung in einem Modell für die gesamte Datenmenge explizit fest; das pfadanalytische Vorgehen ist in dieser Hinsicht, wie die Art des Zusammenhanges festgelegt wird, sehr flexibel. 3. Um die Parameter in diesem Modell bestimmen und das Modell auf seine Anpassungsgüte prüfen zu können, genügt nicht eine Messung nur einer Variablen. Auf irgendeine der folgenden Weisen muß repliziert werden. Verschiedene Arten der Replikation über Person oder Beobachtungsgegebenheiten lassen sich auch miteinander verbinden. Wir unterscheiden folgende Möglichkeiten: (A) Eine Variable wird zu zwei oder mehr Zeitpunkten erhoben (Retest-„stability“). (B) Eine Variable wird zum gleichen Zeitpunkt durch verschiedene Operationalisierungen erfaßt (interne Konsistenz, hier oft als „reliability“ bezeichnet). (C) Zwei oder mehr inhaltlich verschiedene Variablen werden an den gleichen Personen erhoben, meistens in Verbindung mit A oder B. Werden A und B verbunden, kann man die Beziehungen zwischen Messungen und Konstrukten differenziert spezifizieren - „measurement specification“; wird C mit A oder B oder beiden kombiniert, so lassen sich die Beziehungen zwischen den Konstrukten spezifizieren - „theoretical specification“. Literatur zu A: Heise, 1969; Wiley & Wiley, 1971; Werts et al., 1971; Wheaton et al., 1977; Anwendungsbeispiel: Feger, 1978. Literatur zu A und B: Costner, 1969, Blalock, 1970; Hauser & Goldberger, 1971. Zu C mit A, teils auch mit B: Bohrnstedt, 1969; Duncan, 1969; Heise, 1970. Probleme der theoretischen und der Meßspezifikation gleichzeitig behandeln: Duncan, 1972, 1975, Hannan et al., 1974; Jöreskog & Sörbom, 1976; Mayer & Younger, 1974; Wheaton et al., 1977.

Bei der pfadanalytischen Untersuchung von Meßfehlern wird oft ausdrücklich unterschieden zwischen dem zu erfassenden Konstrukt, bisweilen als *unobserved variable* bezeichnet, und den - meßfehlerhafteten - Beobachtungen. Jacobson & Lulu (1974) unterscheiden drei Methoden, wie man die durch Meßfehler unverfälschten Beziehungen zwischen Konstrukten bestimmen kann, über einen *einfachen Indikator* (Single indicator), über einen *Index* und über *mehrfache Indikatoren* (multiple indicators). Arbeitet der Forscher nur

mit einem Indikator, so muß er annehmen, dieser erfasse das Konstrukt in dem Sinne gut, als er den größten Teil der Variation der wahren Werte einfange, und ferner bestehe kein Spezifikationsfehler im Modell: „In other words, given two theoretical variables, both of which are measured by single indicators, these two indicators are assumed to be associated *only* through the posited relationship linking the two theoretical variables. If the latter condition cannot be reasonably assumed - and in most practical situations it cannot - estimates of the structural parameters will be biased even when the first assumption holds.“ Ein Index wird aus mehreren einzelnen Indikatoren als Gesamtwert gebildet. Dabei wäre theoretisch begründet zu entscheiden, wieviele und welche Indikatoren kombiniert werden, welches Gewicht jeder einzelne Indikator erhält, und welcher Art, z.B. additiv, die Zusammenfassung sein soll. Arbeitet man mit multiplen Indikatoren, so legt man in einer Hilfstheorie (auxiliary theory) fest, wie die Indikatoren mit bestimmten Konstrukturen verbunden sind; diese Festlegungen sind meistens prüfbar und die Qualität eines jeden Indikators läßt sich bestimmen (s. Costner, 1969; Blalock, 1969; Mayer & Younger, 1974). Beim Vergleich der drei Methoden bevorzugen Jacobson & Lulu aus mehreren Gründen den multiple indicator approach.

5.4 Besondere Erhebungspläne

Keineswegs nur bei intervallskalierten Daten können durch die Art der Erhebung Beobachtungen in einer Ordnung anfallen, die für die Reliabilitätsprüfung besondere Probleme oder Möglichkeiten ergibt. Wir skizzieren zwei Fälle; im ersten liegen die Beobachtungen als Paarvergleichsmatrix über alle Paare von Beobachtungsgegebenheiten vor, im zweiten Fall ist bei der Analyse der Beobachtungen deren Sequenz zu berücksichtigen. Hubert (1979a) hat vier Fragestellungen formuliert, die sich ergeben, wenn *Paarvergleichsmatrizen*, in deren Zellen Ränge stehen, auf Übereinstimmung (concordance) betrachtet werden: 1. Die Übereinstimmung zwischen zwei oder mehr Matrizen - hier: zwischen Beobachtern, die ihre Beobachtungen als Ränge angeben. 2. Der Vergleich mehrerer Matrizen mit einer spezifischen Matrix - hier: z.B. Vergleich aller Beobachter mit einem Standard. 3. Eine Technik zum Vergleich der Übereinstimmung innerhalb und zwischen Teilmengen von Matrizen, z.B. Vergleich von Beobachterteilgruppen untereinander und miteinander. 4. Suche nach einer Matrix, die alle Urteile aller Beobachter am besten repräsentiert. Die Hypothese, die Übereinstimmung sei nur zufällig, wird geprüft, indem man das beobachtete Übereinstimmungsmaß mit der Verteilung derjenigen Übereinstimmungsmaße vergleicht, die bei allen logisch möglichen Permutationen der intakten Zeilen und Spalten einer der verglichenen Matrizen entstehen. Auf diese Weise werden Abhängigkeiten, die zwischen den Werten in den Zellen der Matrizen durch die Rangordnung über alle Zellenwerte einer Matrix entstehen, für die Inferenzstatistik berücksichtigt. Ein Spezialfall ist gegeben,

wenn in den Zellen nicht Ränge, sondern dichotome Werte stehen, wie z.B. bei soziometrischen Wahldaten (dazu: Hubert & Baker, 1978a).

Im zweiten Fall soll bei der Bestimmung der Reliabilität die in den Aufzeichnungen jedes einzelnen Beobachters enthaltene Information über die *Ordnung oder zeitliche Abfolge* der verschiedenen Verhaltensweisen berücksichtigt werden. Dies ist immer dann erforderlich, wenn aus der Reihenfolgeinformation inhaltliche Schlüsse gezogen werden sollen, z.B. in Kausal- und Bedingungsanalysen. Einige Probleme lassen sich an folgender Tab. 3 verdeutlichen (nach Hollenbeck 1978, seine Tab. 6):

Tabelle 3 : Übereinstimmung in segmentierten Protokollen

verschiedene Zeiteinheiten (sec)	<i>Beobachter 1</i>		<i>Beobachter 2</i>		<i>Beobachter 3</i>	
	Zeit	Ereignis	Zeit	Ereignis	Zeit	Ereignis
1	A		A	A	kein Wert	
2	A	A	A		A	A
3	A		B		A	
4	B		B		B	
5	B	B	B	B	B	B
6	B		B		B	
7	A	A	A	A	C	
8	A		C		C	C
9	C	C	C	C	C	
10	C		C		C	

Das erste Problem, *Synchronisation von Protokollen* (protocol alignment), zeigt sich beim Vergleich des dritten mit den übrigen Beobachtern; er hat den Start verpaßt. Wie läßt sich sein Protokoll mit den übrigen synchronisieren? Und es können auch Ausfälle während der laufenden Beobachtung eintreten. Das zweite Problem besteht in der *Bestimmung der Fehlerart*: Beobachter 3 registrierte in Sek. 7 Verhaltensweise A nicht. Wenn dies sein Fehler war: Hat er das kurze Auftreten übersehen (error of omission, Fehler des Übersehens) oder hat er A mit C verwechselt (error of commission, Verwechslungs-Fehler)? Ein drittes Problem ergibt sich aus dem Vergleich der Zeit- und der Ereignisnotierung für die ersten beiden Beobachter: Die Ereignissequenzen beider Protokolle stimmen völlig überein, die Verhaltensweisen in sec. 3 und 8 nicht. Welche Art Fehler sind diese; und wie soll man sie für die Reliabilitätsbestimmung berücksichtigen? Hollenbeck diskutiert Lösungen für jedes Problem, von denen auch nach seiner Meinung keine zufriedenstellend ist. Der For-

scher, der diesen Erhebungsplan benutzt, sollte zumindest berichten, ob diese Schwierigkeiten in seinen Daten aufgetreten sind und wie er sie gelöst hat (s. auch Kaye 1980).

5.5 Die Berücksichtigung von Reliabilitätskenntnissen bei der weiteren Datenauswertung

Es dürfte selten sein, daß Beobachtungen nicht weiter ausgewertet werden, nachdem sich herausstellt, daß sie ein (vorher festgesetztes) Minimum an Reliabilität nicht aufweisen. Man wird auch nicht allgemein, sondern höchstens im einzelnen Fall festlegen können, wie hoch das erforderliche Minimum anzusetzen wäre. Eine realistische Strategie im Umgang mit Reliabilitätsinformation zeigt sich in den Veröffentlichungen der letzten Jahrzehnte: Die Frage lautet, wie durch Meßfehler und Unzuverlässigkeit der Daten der Schluß - z.B. über die Richtung einer Mittelwertsdifferenz, ihre Signifikanz oder die Höhe einer Korrelation - verfälscht werden könnte, den man aus den Beobachtungen ziehen möchte. Im folgenden können wir nur einige Beispiele erwähnen, die untereinander zwar verschieden sind, aber die Thematik nicht vollständig abdecken.

Unreliable Kategorien und Beobachter führen zu falschen Klassifikationen. Keys & Kihlberg (1963, hier nach Fleiss, 1973 S. 135ff.) behandeln die Frage, wie groß der Fehler ist und in welche Richtung er die Analyse verfälscht: Gegeben seien ein Merkmal, das relativ eindeutig feststellbar ist, z.B. Lungenkrebs, und ein anderes, das weniger sicher zu klassifizieren ist, beispielsweise Häufigkeit des Rauchens. P_L sei die Proportion von Lungenkrebspatienten, die häufig rauchten, somit ist $1-P_L$ die Proportion der Patienten, die nicht rauchten. E_L sei die Proportion der Patienten, die zwar rauchten, aber fälschlich als Nichtraucher klassifiziert wurden, und F_L die Proportion der Patienten, die fälschlich als Raucher eingestuft wurden. Statt wie gewünscht P_L bestimmen zu können, ist nur p_L , die beobachtete Proportion der rauchenden Patienten, aus den Angaben der Personen oder Beobachter zu erhalten. Dabei ist

$$p_L = (1-E_L)P_L + F_L(1-P_L).$$

Also setzt sich p_L additiv aus zwei Teilen zusammen, aus wahren Rauchern und wahren Nichtrauchern. Ob $p_L = , >, < P_L$ hängt von den beiden Fehlerraten F_L und E_L ab:

$$p_L > P_L \text{ wenn } \frac{F_L}{E_L + F_L} > P_L \quad ,$$

$$p_L = P_L \text{ wenn } \frac{F_L}{E_L + F_L} = P_L ,$$

$$p_L < P_L \text{ wenn } \frac{F_L}{E_L + F_L} < P_L .$$

Nicht nur die Schätzung der wahren Proportion wird verständlicherweise durch Unreliabilität beeinträchtigt, sondern auch der *Vergleich zwischen zwei* Proportionen: Ein Forscher möchte prüfen, ob die Proportion der rauchenden Lungenkrebspatienten $P_B < P_L$ ist, und bestimmt die Differenz D zwischen den wahren Proportionen:

$$D = P_L - P_B$$

Beobachtet ist jedoch

$$d = D + (F_L - F_B) + P_B(E_B + F_B) - P_L(E_L + F_L),$$

und d kann größer, gleich oder kleiner D sein. Wenn

$$E_L = E_B = E \text{ und } F_L = F_B = F,$$

ist $d = D(1 - (E + F))$.

Hieraus ergibt sich: (1) Die beobachtete Differenz kann nicht gleich der wahren Differenz sein, wenn auch nur eine Fehlerrate ungleich Null ist. (2) Wenn beide Fehlerraten unter 50% liegen, hat d die gleiche Richtung wie D , ist jedoch numerisch kleiner. Dies ist der seit Bross (1954) bekannte Minderungseffekt, der jedoch - wie die eben skizzierte Analyse ergab - keineswegs immer auftritt. (3) Die Annahme ist falsch, Unreliabilität könne lediglich Unterschiede mindern; sie kann auch die Richtung eines Zusammenhanges fälschen (weitere Literatur zur Auswirkung von Fehlklassifikationen auf Zusammenhangsmaße und χ^2 : Rogot, 1961; Mote & Anderson, 1965, Assakul & Procter, 1967; Koch, 1969). Fleiss (1973) behandelte ausführlich und mit zahlreichen Literaturverweisen die Frage, wie der Meßfehler statistisch kontrolliert und wie Korrekturfaktoren bestimmt werden können.

Ein weiterer Forschungsschwerpunkt in diesem Bereich befaßt sich mit der Frage, wie Unreliabilität den Typ II oder β -Fehler oder die Prüfstärke ($1 - \beta$, power, s. Cohen, 1969; Bredenkamp, 1972) eines Inferenztests beeinflusst. Nicewander & Price (1978) haben kürzlich den Stand der Diskussion zusammengefaßt. Die Annahmen über den Meßfehler sind dabei in der Regel die der klassischen Testtheorie, die über den Stichprobenfehler jene des allgemeinen linearen Modells, speziell der Varianzanalyse. Der Fehlerterm der Varianzana-

lyse besteht dann aus der Abweichung des wahren Wertes vom Bedingungs-mittelwert plus dem Meßfehler. Ändert man die Reliabilität der abhängigen Variablen, so kommt es für die Auswirkung auf die Prüfstärke darauf an, ob dadurch die Varianz der wahren Werte, der Fehler, oder beide geändert wurden. Nimmt man wie Sutcliffe (1958) und Cleary & Linn (1969) an, nur die Varianz der Fehler werde verringert, dann muß die Prüfstärke ansteigen. Sutcliffe (1980) zeigt erneut den nachteiligen Effekt der Unreliabilität von Differenzen und der abhängigen Variablen für die power einer statistischen Prüfung.

Abschließend seien noch einige weitere Beispiele aufgezählt: Werts & Linn (1971) zeigen, wie man die Unreliabilität einer Kovariaten in einer Kovarianzanalyse berücksichtigen kann. Rock et al. (1977) behandeln den Fall des linearen Modells, wenn (auch) die unabhängigen Variablen fehlerbehaftet sind. Stroud (1974) diskutiert Möglichkeiten des Posttest-Vergleichs bei zwei Gruppen, wenn fehlerbehaftete Pretest-Werte vorhanden sind, und Alternativen zur üblichen Kovarianzanalyse. Nur bei Unreliabilität können Regressionsartefakte auftauchen, die Hunter & Cohen (1974) bei der Analyse nichtlinearer Modelle der Einstellungsänderung berücksichtigen. Huber (1973) hat die Bedeutung der testspezifischen Fehlervarianz für die zufallskritische Einzelfalldiagnostik herausgearbeitet. In der Entscheidungsforschung hat Fischer (1976) zwei verschiedene Fehlertheorien, eine für rating, die andere für Paarvergleiche inhaltlich begründet und in ihren Auswirkungen auf die Prüfung des multidimensionalen Nutzenmodells analysiert. Cochran (1968) gibt eine gute Einführung in den gesamten Bereich.

6. *Validität von Beobachtungen*

Man kann zur Beobachtung Apparate verwenden und ihre Daten statistisch analysieren, ohne daß ein Mensch dazwischentritt; ein Beobachter hat also keinen Einfluß auf die Daten. Dann könnten die Beobachtungen völlig objektiv und reliabel sein; das Validitätsproblem stellt sich dennoch, denn aus den Beobachtungen werden Schlüsse gezogen. Diese Schlüsse können sich von einer Menge von Beobachtungen auf eine andere Menge beziehen oder auf theoretische Vorstellungen, und als Vorhersagen oder Korrespondenzaussagen formuliert sein. Solche Schlüsse können für jeweils bestimmte Beobachtungen und bestimmte theoretische Vorstellungen gültig sein. Validität - oder synonym Gültigkeit - bezieht sich zunächst auf Schlußfolgerungen: Beobachtungen oder bestimmte Verfahren, Beobachtungen zu gewinnen, kann man dann als valide bezeichnen, wenn die auf ihnen basierenden Schlüsse zutreffen. Die Methodenlehre der Validierung, d.h. der Prüfung von Validität, befaßt sich mit der Frage nach den Kriterien (und ihrer Anwendung) der Prüfung, ob diese Schlüsse zutreffen. Hierzu wurden verschiedene Kriterien und Verfahrensweisen entwickelt, die in den folgenden Abschnitten geschildert werden.

Wenn Beobachtungen von menschlichen Beobachtern stammen, kommen nicht nur Täuschung, Selbsttäuschung und systematisch verfälschende Beobachtungstendenzen als problematisch für die Gültigkeit hinzu. Vielmehr wird man menschlichen Beobachtungen, läßt man die Möglichkeit von Täuschungen und Fälschungen einmal beiseite, im Prinzip auch Gültigkeit zusprechen müssen, ohne daß sich die Gültigkeit auf einen logischen Schluß bezieht. Der Autofahrer - um bei Alltagsbeobachtungen zu bleiben - sieht, daß die Ampel rot ist, der Vordermann doch noch durchfährt, schlägt mit der Hand an seine Stirn und ärgert sich. Die Beobachtungen und Schilderungen des eigenen und fremden Verhaltens und des eigenen Erlebens kann man als valide bezeichnen in dem Sinn, daß sie sich auf die unmittelbare Anschauung des Beobachters beziehen. Diese Basis, die jede menschliche Beobachtung konstituiert, wird im folgenden nicht weiter besprochen. Zwar kann eine Methodenlehre Bedingungen und Prüfmöglichkeiten angeben, wann und wie Fremd- und Selbstbeschreibungen verfälscht werden; sie muß die grundsätzliche Möglichkeit gültiger Beobachtungen durch Menschen jedoch voraussetzen. Validitätstheorie befaßt sich also nicht mit den philosophischen Bedingungen der Möglichkeit von - zutreffenden - Beobachtungen durch Menschen und der näheren Bestimmung des Wortes „zutreffend“, und auch nicht mit der Fülle physiologischer, psychologischer, sozialer und anderer Faktoren, die verzerrend und störend auf die Prozesse der Wahrnehmung und des Beurteilens einwirken können. Validitätstheorie innerhalb der Methodenlehre befaßt sich mit der Prüfung der Gültigkeit von Schlüssen aus Beobachtungen.

Die Prüfung, ob Schlüsse von Beobachtungen auf Konstrukte und auf andere Beobachtungen zutreffen, kann nach verschiedenen Kriterien geschehen. Diese Kriterien führen zur Unterscheidung verschiedener *Validitätsarten*. Üblicherweise werden folgende Validitätsarten unterschieden: Kriteriumsvalidität mit den Varianten prädiktive und gleichzeitige (concurrent) Kriteriumsvalidität, Inhalts- oder Kontentvalidität und Konstruktvalidität (vgl. Davis, 1974). Bei *Kriteriumsvalidität* interessiert der Schluß vom gerade beobachteten Verhalten auf anderes Verhalten, das gleichzeitig oder zeitlich später gezeigt wird, sei es nun gleichartiges oder verschiedenes Verhalten. Zweck ist die Vorhersage (im statistischen Sinn) einer Menge von Beobachtungen aufgrund einer anderen Beobachtungsmenge.

Die *Inhaltsvalidität* bezieht sich auf einen Repräsentationsschluß, der von einer Stichprobe von Beobachtungen auf ein definiertes Universum gerichtet ist (zur Vertiefung und für weitere Literatur s. Klauer, 1978). Wegen ihrer besonderen theoretischen Wichtigkeit widmen wir der Konstruktvalidität eigene Abschnitte.

6.1 Konstruktvalidierung

Für zahlreiche Variablen, die durch Beobachtung erfaßt werden sollen, läßt sich inhaltliche oder kriterienbezogene Validität jedoch (noch) nicht nachweisen, weil man nicht über ein plausibles Kriterium verfügt oder man das Universum der Verhaltensweisen, Situationen, Aufgaben o.ä. nicht erschöpfend und nach theoretisch fundierten Taxonomien beschreiben kann. Für solche Fälle wird der langwierige Prozeß der *Konstruktvalidierung* erforderlich, dessen Grundgedanken erstmals umfassend von Cronbach & Meehl (1955) systematisiert wurden: „Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim.“ Zwischen verschiedenen theoretischen Konstrukten und zwischen Konstrukten und Beobachtungen bestehen Beziehungen, die in der Konstruktvalidierung expliziert und - soweit möglich - empirisch geprüft werden.

Eine Beziehung stellt, formal betrachtet, den Zusammenhang von mindestens zwei Variablen dar. Der Zusammenhang kann statistisch auf verschiedene Weise analysiert werden, z.B. als Vergleich von Mittelwertdifferenzen bei der einen Variablen, wenn die andere die Gruppierung bedingt, als Trendhypothese, etc. Besondere Bedeutung hat in der Validierungsforschung die Formulierung des Zusammenhanges als Korrelation oder Kovariation gefunden. Besteht zwischen zwei reliablen Meßwertreihen kein Zusammenhang, so ergibt sich kein Problem für die Schlußfolgerung, auf die sich ein Validitätsurteil stützen könnte. Besteht jedoch ein signifikanter und substantieller Zusammenhang, so sind mehrere Schlüsse möglich. 1. Es besteht ein inhaltlich begründeter Zusammenhang zwischen A und B; 2. die experimentelle Manipulation, die B variieren sollte, variierte statt dessen A; 3. zwar wurde B experimentell variiert oder zeigte ohne Eingriff des Forschers Variation, das Instrument jedoch, das A erfassen sollte, registriert in Wirklichkeit B (Tesser & Krauss, 1976). Wenn A und B beispielsweise Ängstlichkeit und Selbstwertgefühl darstellen, könnten alle drei Deutungen a priori gleich plausibel sein. Den Fall, daß eine dritte Variable A und B beeinflußt, verfolgen Tesser & Krauss zunächst nicht weiter. Ihre entscheidende Frage ist dann, unter welchen Bedingungen man schließen kann, der beobachtete Zusammenhang zwischen zwei Variablen bedeute etwas anderes als daß beide Indikatoren des gleichen Konstrukts sind. Dafür stellen sie eine nicht erschöpfende Liste von Kriterien auf. Ein wichtiges Kriterium besteht darin zu beobachten, ob die Veränderung einer dritten Größe die Beziehung zwischen A und B ändert; wenn ja, können sie nicht (nur) Indikatoren des gleichen Konstruktes sein. Ein weiteres Kriterium prüft, ob die Effekte von A auf B und von B auf A asymmetrisch sind, ein drittes Kriterium die Konstanz der Beziehung zwischen A und B über verschiedene Erfassungsmethoden. Die Frage, wie ein Zusammenhang zwischen

zwei Variablen, die unterschiedliche Konstrukte erfassen sollen, zu deuten ist, sollte logisch vor einer Prüfung der konvergenten und diskriminanten Validität geklärt werden.

Die Möglichkeiten der Konstruktvalidierung, sofern diese korrelationsanalytisch vorgeht, wurden durch den Ansatz der *konvergierenden und diskriminierenden Validierung* mit Hilfe einer multitrait-multimethod matrix (MTMM) von Campbell & Fiske (1959) wesentlich vorangetrieben und systematisiert. Der Ansatz geht von folgenden Erwägungen aus: (1) Wenn unabhängige, in ihrer Operationalisierung des Konstruktes sich nicht überlappende Meßverfahren das Gleiche erfassen sollen und gültig sind, so müssen sie, auf die gleichen Meßgegebenheiten angewandt, zu gleichen Ergebnissen kommen; die Ergebnisse müssen konvergieren. (2) Beobachtungen und Messungen, die inhaltlich verschiedene Konstrukte erfassen sollen, dürfen nicht höher korrelieren als aufgrund der Theorie des jeweiligen Gegenstandsbereiches zu erwarten ist, auch dann nicht, wenn sie mit Verfahren gewonnen werden, die technisch identische Operationalisierungsschritte aufweisen. Auch von der Erhebungsmethodik her gesehen gleiche Verfahren müssen inhaltlich verschiedene Konstrukte diskriminieren, wenn die Beobachtungen valide sein sollen. (3) Bei jeder konkreten Beobachtung oder Messung wird eine „trait-method unit“, eine Einheit von spezifischem Inhalt und bestimmtem, allerdings nicht an diesen Inhalt gebundenem Meßvorgehen erfaßt. So liegt nun der Gedanke nahe, die systematische Varianz zwischen Beobachtungen aufzuspalten in einen Teil zu Lasten der Reaktionen auf unterschiedliche Eigentümlichkeiten verschiedener Erhebungsverfahren und einen Teil zu Lasten der Reaktionen auf inhaltliche Unterschiede zwischen den Variablen, die verschiedenen Konzepten zugeordnet sind. Damit wird es notwendig, bei der Prüfung der konvergenten und diskriminierenden Validität sowohl mehr als ein Konstrukt zu untersuchen als auch mehr als eine Erhebungsmethode zur Erfassung des gleichen Konstruktes zu benutzen, und wiederholte Messungen an den gleichen Vpn sind in der Regel erforderlich.

Die Aufbereitung der Daten geschieht übersichtlich in einer multitrait-multimethod Matrix, in der Korrelationen nach dem in Tab. 4 benutzten Schema angeordnet werden können. Die in der Hauptdiagonalen stehenden, eingeklammerten Werte stellen Reliabilitätskoeffizienten dar. Jeder dieser Koeffizienten bezieht sich auf eine Methode und ein Konzept (= „*monotrait-monomethod value*“). Die sich anschließenden Dreiecke mit den durchgezogenen Linien - „*heterotrait-monomethod triangle*“ - ergeben zusammen mit den zugehörigen Werten der Reliabilitätsdiagonale einen Einmethodenblock (*monomethod block*). Zwischen den gestrichelten Dreiecken befinden sich „Validitätsdiagonalen“, darin stellt jeder Koeffizient einen „*monotrait-heteromethod*“ Wert dar. Zwei angrenzende, gestrichelte Dreiecke, die mit möglicherweise verschiedenen Werten besetzt sein können und als „*heterotrait-heteromethod*“

triangles“ bezeichnet werden, bilden zusammen mit der zugehörigen Validitätsdiagonale einen Mehrmethodenblock.

Tabelle 4: Fiktive Korrelationen in einer multitrait-multimethod Matrix (nach Campbell & Fiske, 1959)

		Methode I			Methode II			Methode III		
<i>Konzepte</i>		A ₁	B ₁	C ₁	A ₂	B ₂	C ₂	A ₃	B ₃	C ₃
Methode I	A ₁	(.89)								
	B ₁	.51	(.89)							
	C ₁	.38	.37	(.76)						
Methode II	A ₂	.57	.22	.09	(.93)					
	B ₂	.22	.57	.10	.68	(.94)				
	C ₂	.11	.11	.46	.59	.58	(.84)			
Methode III	A ₃	.56	.22	.11	.67	.42	.33	(.94)		
	B ₃	.23	.58	.12	.43	.66	.34	.67	(.92)	
	C ₃	.11	.11	.45	.34	.32	.58	.58	.60	(.85)

Damit Konstruktvalidität als nachgewiesen gelten kann, sollten folgende Forderungen erfüllt sein:

- (1) Die Koeffizienten in den Validitätsdiagonalen sollten signifikant von Null verschieden und substantiell sein; dies rechtfertigt den Schluß auf konvergierende Validität.
- (2) Ein Wert in der Validitätsdiagonale sollte größer sein als die Koeffizienten in den Zeilen und Spalten der gleichen heterotrait-heteromethod Dreiecke. D.h., die Validität der Variablen sollte größer sein als die Korrelation dieser Variablen mit irgendeiner anderen Variablen, die mit ihr weder das Konzept noch die Meßmethode gemeinsam hat (im Beispiel: $A_1A_2 = .57$ sollte größer sein als $A_1B_2 = .22$, $A_1C_2 = .11$, $A_1B_3 = .23$, $A_1C_3 = .11$ sowie $A_2B_1 = .22$, $A_2C_1 = .09$, $A_2B_3 = .43$, $A_2C_3 = .34$).
- (3) Eine Variable sollte höher mit einem unabhängigen Meßversuch korrelieren, der sich auf das gleiche Konzept bezieht, als mit Messungen, die sich „zufällig“ der gleichen Methode bedienen, jedoch zur Erfassung eines anderen Meßobjektes durchgeführt werden. Verglichen werden also Koeffizienten in der Validitätsdiagonale mit den Korrelationen in den zugehörigen heterotrait-monomethod Dreiecken (z.B. einerseits $A_1A_2 = .57$, andererseits $A_1B_1 = .51$; einerseits $A_1A_3 = .56$, andererseits $A_1C_1 = .38$. Für A₁ ist im Beispiel Kriterium 3 annähernd erfüllt, nicht jedoch z.B. für A₂).

- (4) Das gleiche Muster der korrelativen Beziehungen zwischen den Meßobjekten sollte sich in den heterotrait-Dreiecken sowohl der monomethod als auch der heteromethod Blöcke zeigen; die Beziehungen zwischen den Variablen sollten methoden-invariant sein (dies ist im Beispiel der Fall, trotz deutlicher Unterschiede in der absoluten Stärke der Korrelationen). Die Kriterien (2) bis (4) gestatten den Schluß auf die diskriminierende Validität.

Die grundlegende Auffassung der Validität, die den verschiedenen Ansätzen zu ihrer Bestimmung im Vorschlag von Campbell & Fiske in Abhebung zur Reliabilität gemeinsam ist, ist die *Übereinstimmung der Ergebnisse von unabhängigen Meßprozeduren am gleichen Objekt*. Die Beziehung zwischen den Konzepten Reliabilität und Validität erscheint aus dieser Perspektive über die Ähnlichkeit der Meßverfahren gegeben. Reliabilität ist die Übereinstimmung zwischen den Ergebnissen zweier Versuche, das gleiche Meßobjekt durch maximal ähnliche Verfahren zu erfassen. Validität drückt sich in der Übereinstimmung zwischen den Ergebnissen zweier Versuche aus, das gleiche Meßobjekt durch maximal verschiedene Methoden zu erfassen. Die Ähnlichkeit von Methoden kann man als variierend zwischen den Enden „maximal“ und „minimal“ eines Kontinuums auffassen; je stärker wir uns dem Pol maximale Ähnlichkeit nähern, desto stärker muß die Information als auf Reliabilität bezogen interpretiert werden. In diesem Sinne deuten Campbell & Fiske den Split-half Koeffizienten als in stärkerem Maße validitätsrelevant als ein Test-Retest-Koeffizient, da einzelne Items nicht völlig identisch sein können.

Es liegt nahe, Zeilen und Spalten der multitrait-multimethod Matrix inhaltlich anders als Campbell & Fiske zu definieren, wenn dies durch die Problemstellung erforderlich wird. In einer Untersuchung inhaltlich unterschiedlicher Erlebensdimensionen in einem intraindividuellen Konflikt verwendete Feger (1971; 1978, Kap. 2) für jedes der drei Konzepte drei verschieden formulierte Schätzskalen, insgesamt also neun Skalen. Der Versuchsplan entspricht somit einem design mit nested effects:

Konzept	A	B	C
Skala	1, 2, 3	4, 5, 6	7, 8, 9

Centra (1971) setzt an die Stelle unabhängiger Methoden diskrete soziale Gruppen; verschiedene Konzepte werden realisiert als Skalen, die Gruppenreaktionen auf verschiedene Aspekte eines multidimensionalen Wahrnehmungsraumes erfassen sollen. Ziel ist somit eine Einschätzung der Übereinstimmung der Gruppen; je ähnlicher die Urteile bei inhaltlich gleichen Skalen, desto höher die Validität, wobei der Gedanke im Hintergrund steht: „Wenn . . . diskrete Gruppen in gleicher Weise auf Skalen reagieren, die ihre Umwelt erfassen sollen, dann wird die Annahme einleuchtender, daß die Skalen Charakteristika oder Bedingungen dieser Umwelt wiedergeben.“

Krause (1972) hat die logischen Implikationen der konvergierenden und diskriminierenden Validierung untersucht. Er betont, daß andere Gründe als die Erfassung des gleichen Merkmals zur Konvergenz verschiedener Methoden führen können, und zeigt, daß sehr spezifische Bedingungen erfüllt sein müssen, damit nach den Kriterien von Campbell & Fiske auf Validität geschlossen werden kann. Er führt u.a. aus, die Meßverfahren müßten dafür bereits ein Mindestmaß an a priori-Validität besitzen, läßt aber unklar, was mit a priori-Validität gemeint ist und wie man nun ihr Vorliegen prüfen kann.

6.2 Neuere Entwicklungen zur Analyse von multitrait-multimethod Matrizen

Die Kritik, die an Campbell & Fiskes Ansatz geübt wurde, betont, die Entscheidung, ob die vier Kriterien oder auch jedes einzeln erfüllt sei, ließe sich im konkreten Fall nicht eindeutig treffen, weil die Kriterien nur vage verbal und nicht formal, exakt prüfbar definiert seien. Diese Kritik greifen Hubert & Baker (1978) auf und definieren die nur leicht modifizierten Kriterien formal, und zwar so, daß man inferenzstatistisch prüfen kann, ob sie erfüllt sind. Die gewählte Inferenzstatistik ist nonparametrisch, und man kann sie auch auf Korrelationskoeffizienten anwenden, die lediglich Ordinalskalenniveau in den Daten voraussetzen, wie etwa τ oder γ , sofern man τ oder γ selbst Intervallskalenniveau zuerkennt. Die später zu beschreibenden Weiterentwicklungen setzen durchweg Intervallskalenniveau der Daten voraus und gehen von Produkt-Moment-Korrelationskoeffizienten aus. Wie schwerwiegende Folgen Verstöße gegen diese impliziten Annahmen haben, ist nicht bekannt.

Mit anderen Argumenten hat Krause (1972, p. 183) infrage gestellt, ob Produkt-Moment-Korrelationen die geeigneten Koeffizienten seien, um Konvergenz und Diskriminanz von Verfahren festzustellen: „Whether or not two methods can be valid for the same trait is a question of codimensionality rather than (linear) prediction, and differences between methods in measurement distributions over the same n subjects can yield low correlations even when the pair of measurement sets is perfectly ordinally consistent, i.e., shows perfect Scalogram reproducibility, which is the necessary topological condition for codimensionality.“

Hubert & Baker definieren vier Indizes: (1) den Durchschnitt der Korrelationen zwischen den gleichen Merkmalen, (2) die Differenz zwischen dem ersten Index und dem Durchschnitt der Korrelationen zwischen verschiedenen Merkmalen, gemessen mit verschiedenen Methoden, (3) die Differenz zwischen dem ersten Index und dem Durchschnitt der Korrelationen zwischen gleichen Methoden, und (4) einen Index, der anzeigt, wie gleichartig das Korrelationsmuster der Merkmale über verschiedene Methoden hinweg ist. Um

die Bedeutung der numerischen Werte für die verschiedenen Indizes zu beurteilen, entwickeln Hubert & Baker eine Stichprobentheorie für die Zufallsverteilung der Indizes. Die allgemeine Nullhypothese besagt, daß die Korrelationen nicht die Struktur aufweisen, die einen Schluß auf konvergente und diskriminante Validität rechtfertigen könnte, daß die Zuordnung eines Meßverfahrens zu einer bestimmten Kombination von Methode und Merkmal eine zufällige ist. Die Indizes werden dann gegen die Zufallsverteilung auf einem vom Forscher gewählten Konfidenzniveau verglichen, und es lassen sich objektive Entscheidungen darüber fallen, welche Kriterien erfüllt sind.

Die erste kritische Weiterentwicklung des MTMM-Ansatzes stammt von Campbell & O'Connell (1967). Sie untersuchen die Frage, wie die Varianzquellen zusammenwirken, um die Höhe einer Korrelation in einer MTMM zu bestimmen. Die beiden angenommenen Varianzquellen sind (1) der Zusammenhang zwischen den Merkmalen und (2) der Zusammenhang zwischen den Meßmethoden. Für die untersuchten Datensätze mußte die Annahme eines additiven Zusammenwirkens der Varianzquellen zugunsten eines multiplikativen aufgegeben werden: Je höher eine heterotrait-heteromethod Korrelation ausfällt, desto stärker steigt sie dadurch an, daß die gleichen Variablen mit der gleichen Erhebungsmethode erfaßt werden. Als Konsequenz ergibt sich die Forderung, spezifische faktoranalytische Modelle zu prüfen, die nicht von einem additiven Zusammenwirken von Methoden- und Merkmalsfaktoren ausgehen.

Jackson (1969) kann als der eigentliche Begründer des faktoranalytischen Auswertungsansatzes von MTMM-Daten angesehen werden. Um das Vorgehen von Campbell & Fiske zu verbessern, geht er von folgenden Überlegungen aus: (1) Korrelationen sind in ihrer Höhe auch abhängig von Stichproben- und Meßfehler, was zum Fehlurteil führen kann, diskriminante Validität liege nicht vor. Daher sollte ein ideales Auswertungsverfahren diese Fehlerquellen, die Unreliabilität der Variablen und Korrelationen berücksichtigen. (2) Die Replizierbarkeit von Ergebnissen aus Validitätsstudien läßt sich vergrößern und Effekte von Stichprobenfehlern lassen sich mindern, wenn nicht einzelne Korrelationen, sondern das gesamte Korrelationsmuster der MTMM der Analyse zugrundegelegt wird. (3) Man kann nicht, wie Campbell & Fiske, generell davon ausgehen, daß heteromethod-Validitätskoeffizienten höher ausfallen als monomethod-heterotrait-Korrelationen. Das Analyseverfahren sollte die Beziehung zwischen Methodenvarianz und Merkmalsvarianz spezifizieren. (4) Zu welchen Schlüssen man über die konvergente und diskriminante Validität von Beobachtungen kommt, hängt von der Anzahl und relativen Ähnlichkeit der untersuchten Merkmale und Methoden ab. Wie Reliabilitätsaussagen auf den jeweiligen Erhebungsplan und die verwendete Stichprobe relativiert werden müssen, so sind Aussagen zur Konstruktvalidität darüber hinaus zu beziehen auf die mituntersuchten Merkmale und Methoden.

Jacksons eigenes Vorgehen, „multimethod factor analysis“ berücksichtigt nur Korrelationen zwischen Merkmalen, die mit verschiedenen Methoden erfaßt wurden, und gestattet nicht, Methodenfaktoren zu identifizieren. Außerdem können bei dieser Methode mathematisch-statistische Probleme auftreten (s. Schmitt et al. 1977, Conger, 1971). Jackson (1975) hat deshalb seinen Ansatz neu formuliert, wobei er an die Hauptkomponentenanalyse von Golding & Seidman (1974) anknüpfte. Die verschiedenen Verfahren werden von Schmitt et al. (1977) verglichen; auf den Ansatz von Boruch & Wolins (1970), der neben einem hypothetischen allgemeinen Faktor je einen pro Merkmal und pro Methode vorsieht, sei lediglich verwiesen. Hoffman & Tucker (1964), später z.B. auch Schmitt et al. (1977) haben Tuckers (1966) dreimodale Faktoranalyse zur Validitätsprüfung angewendet. Abgesehen von technischen und interpretativen Problemen, welche dieser Ansatz bietet, betont Jackson (1969, S. 35), diese Methode versuche, eine gemeinsame Faktorenstruktur der Merkmale zu identifizieren, die in mehr als einer Meßmethode wiederkehrt, und das Ziel sei nicht, für ein einzelnes Erhebungsverfahren oder ein bestimmtes Merkmal konvergierende und diskriminierende Validität zu untersuchen.

Das *pfadanalytische Vorgehen* bei Validierungsstudien haben meines Wissen Werts & Linn (1970) in die Psychologie eingeführt. Als Anwendungsbeispiele seien die Arbeiten von Kalleberg & Kluegel (1975), Ray & Heeler (1975) sowie Schmitt (1978) erwähnt. Die Pfadanalyse (zur Einführung: Heise, 1975) stellt ein System dar, mit dessen Hilfe Korrelationen in bezug auf ein vorgegebenes theoretisches Modell der Beziehungen zwischen den untersuchten Variablen interpretiert werden können. Campbell & Fiskes Kriterien werden im pfadanalytischen Modell im allgemeinen als die Hypothese interpretiert, jede Beobachtung für eine MTMM komme als Wirkung einer latenten Methodenvariablen und einer latenten Merkmalsvariablen zustande, plus einer Fehlerkomponenten. Methoden und Merkmale werden also als latente, nicht direkt gemessene, sondern aus verschiedenen Beobachtungen erschlossene Variablen aufgefaßt, wobei (zunächst) pro Methode und pro Merkmal eine spezifische latente Variable oder Faktor veranschlagt wird, weshalb dieses Vorgehen auch als „konfirmatorische Faktoranalyse“ bezeichnet wird. Das pfadanalytische Modell muß weiter die Beziehung zwischen Beobachtungen, Methoden- und Merkmalsfaktoren spezifizieren, z.B. festlegen, ob Methodenfaktoren als unabhängig von Merkmalsfaktoren gedacht werden, wie Campbell & Fiske annehmen. Die Auswertung geschieht oft mit Verfahren, die durch Jöreskog entwickelt wurden (Jöreskog 1969, 1970; Jöreskog & Sörbom, 1978). Ob Modell und Daten übereinstimmen, läßt sich inferenzstatistisch prüfen, auch verschiedene Modelle für die gleichen Beobachtungen lassen sich so vergleichen. Eine wichtige Validitätsfrage, ob bestimmte Methoden verschiedene Merkmale unterschiedlich verzerren, läßt sich für alle Methoden und Merkmale beantworten. Schmitt et al. (1977, S. 460) nennen folgende Vorteile einer pfadanalytischen Validierungsstudie: (1) sie zwingt zu einer exakten Formulie-

rung der Annahmen und deren logische Konsequenzen, (2) sie erlaubt, die Korrelationen zwischen Merkmalen, zwischen Methoden sowie zwischen Merkmalen und Methoden zu schätzen, (3) sie gestattet für jede Beobachtung die Einwirkung der Methoden- und Merkmalsfaktoren sowie der spezifischen Varianz abzuschätzen (Weiterentwicklung bei Bien 1980).

Man kann, wie Costner, ausdrücklich betonen, daß Beobachtungen nur den Charakter und die Funktion von *Indikatoren* für erschlossene, latente Variablen haben und explizit eine „Hilfstheorie“ (Costner, 1969) formulieren, deren Aufgabe es ist, die Beziehungen zwischen empirischen Indikatoren und Konstrukten zu beschreiben, und so die Grundlage für Urteile über die Güte der einzelnen Indikatoren bereitzustellen. Wichtig ist dabei auch, daß die inhaltliche Theorie nicht als an die empirische Prüfung durch nur eine spezielle Auswahl von Beobachtungen gebunden erscheint. Vielmehr sind im allgemeinen mehrere Hilfstheorien für eine inhaltliche Theorie möglich. Für Validitätsstudien haben Althausen & Heberlein (1970) sowie Costner & Schoenberg (1973) diese Zugangsweise vorgeschlagen, die sich mit konfirmatorischen Faktoranalysen verbinden läßt.

Insbesondere richtet sich in diesem Ansatz die Aufmerksamkeit auf die Frage, ob Meßfehler beim Erfassen des einen Konstrukts (oder verbunden mit einem bestimmten Indikator) mit Meßfehlern bei anderen Konstrukten oder Indikatoren korrelieren, ob also „differential bias“, „systematic measurement error“ oder „correlated measurement error“ vorliegt (Avison, 1978). Campbell & Fiske schließen auf konvergente Validität, wenn die monotrait-heteromethod Korrelationen hoch sind. Aus der Sicht der Indikator-Analyse impliziert diese Annahme zwei weitere: (1) die „epistemischen Koeffizienten“, die den Zusammenhang zwischen Indikator und Konstrukt beschreiben, sind substantiell hoch, (2) die Erhebungsmethoden sind unkorreliert. Ohne die zweite Annahme könnten hohe monotrait-heteromethod Korrelationen auf Zusammenhänge zwischen den Methoden zurückzuführen sein. Auf weitere, oft unrealistische Annahmen, die Campbell & Fiskes Kriterien für diskriminante Validität zugrunde liegen, machen Althausen & Heberlein, Althausen et al. (1971) sowie Avison (S. 438f.) aufmerksam. Althausen & Heberlein entwickeln eine eigene Vorgehensweise, die im Vergleich verschiedener Modelle für eine MTMM besteht. Costner & Schoenbergs Verfahren bewahrt den Forscher nicht nur vor falschen Schlüssen über den kausalen Zusammenhang von Variablen, der in Wirklichkeit auf korrelierten Meßfehlern beruht. Es erlaubt auch eine Diagnose von Schwächen bestimmter Indikatoren und eine genaue Abschätzung sowohl der epistemischen Koeffizienten als auch der „wahren“ Zusammenhänge zwischen den Konstrukten (s. auch Alwin 1974, Althausen 1974, Feger 1978, S. 77f.).

Die Informationen in den Beobachtungen einer Validitätsstudie lassen sich auch *varianzanalytisch* betrachten, wie dies Stanley (1961) vorgeschlagen hat.

Tabelle 5: Vergleich von Methoden für Studien über konvergierende und diskriminierende Validität
(nach Schmitt et al. 1977, S. 475).

ANALYSE-METHODE					
ZIELE	Campbell & Fiske	Stanley (ANOVA)	Pfadanalyse	Golding & Seidman	Jackson (1975) Tucker (dreimodale FA)
1. Trennung von NEIN Merkmals- und Methodenvarianz.	NEIN	NEIN	Bei genügend großem N können verschiedene Modelle für Varianzquellen geprüft werden	Gemeinsame und spezifische Varianz sind in der ersten Transformation konfundiert	Gemeinsame und spezifische Varianz sind in der ersten Transformation konfundiert
2. Evaluationskriterien für einzelne Merkmale und Methoden statt für Faktoren oder Komponenten	JA	NEIN, es gibt nur eine Gesamtaussage für die untersuchten Merkmale und Methoden	JA, jede Beobachtung wird als Resultat von Merkmals-Methoden- und spezifischer Varianz postuliert	NEIN, Konvergenz der Merkmalsfaktoren über Methoden ist das einzige Ziel	Konvergenz und Diskrimination geschehen auf dem Niveau der Faktorscores; Rotation gezielt auf einzelne Variablen
					Ziel ist Schätzung der Interkorrelationen von Personen-, Methoden- und Merkmalsfaktoren

3. Basis für Bewertung des Ausmaßes an konvergenter und diskriminanter Validität	Prozentsatz, mit dem die Beziehungen zwischen Korrelationen den Kriterien entsprechen	Signifikanztest für Konvergenz von Merkmal und Methode	Chi ² -Tests für die Anpassungsgüte der angenommenen Modelle	Auftauchen eindeutiger Merkmalsfaktoren im zweiten Analyseschritt	Auftauchen von Merkmals- und Methodenfaktoren im zweiten Analyseschritt	Höhe der Korrelationen zwischen Personen-, Methoden- und Merkmalsfaktoren in der Kernmatrix
4. Bestimmung der Position der Merkmalsträger auf einem zugrundeliegenden Merkmal	NEIN	NEIN	JA	JA	JA	JA

Benutzt wird eine dreifaktorielle Varianzanalyse mit den systematischen Varianzquellen Methoden, Merkmale und Personen bzw. Merkmalsträger (Beispiele und Erweiterungen: Boruch et al. 1970, Kavanagh et al. 1971). Wie in den zuvor geschilderten Ansätzen liegt auch hier das allgemeine lineare Modell zugrunde, allerdings wird der Meßfehler nicht thematisiert. Zwar wird eine Validitätsaussage über alle verwendeten Methoden und Merkmale zugleich angestrebt, nicht jedoch für eine beliebige, bestimmte Erhebungsmethode für ein spezifisches Merkmal. Aus der Größe der Varianz zulasten der Merkmalsträger wird auf konvergierende Validität geschlossen. Dies ist, wie Schmitt et al. (1977) zu Recht betonen, nicht konvergierende Validität im Sinne von Campbell & Fiske, sondern das Ausmaß, in dem ein allgemeiner Faktor die gesamte MTMM zu erklären vermag. Insbesondere aus der Varianz zulasten der Interaktionen von Personen und Merkmalen sowie zwischen Personen und Methoden wird auf die diskriminierende Validität geschlossen.

Schmitt et al. (1977) haben verschiedene Methoden verglichen, mit deren Hilfe man konvergierende und diskriminierende Validität beurteilen kann. In ihrer Tab. 16 geben sie dazu eine Übersicht, zugleich die verschiedenen Ziele, die man mit einer Validitätsstudie verfolgen kann. Mit JA oder NEIN notieren sie, ob sich ein bestimmtes Ziel durch die betrachtete Methode erreichen läßt. Wir geben die übersetzte Tabelle als Tab. 5 wieder.

Literatur

Mit * gekennzeichnete Angaben werden nicht im Text zitiert, sind jedoch als interessante Quellen angeführt.

- Adams, R. S. 1970. Duration and incident frequency as observation indices. *Educational and Psychological Measurement*, 30, 669-674.
- Althauser, R. P. 1974. Inferring validity from the multitrait-multimethod matrix: Another assessment. In: Costner, H. L. (ed.): *Sociological Methodology*, 1973/74. San Francisco: Jossey-Bass, 106-127.
- Althauser, R. P. & Heberlein, T. A. 1970. Validity and the multitrait-multimethod matrix. In: Borgatta E. F. & Bohrnstedt, G. W. (eds): *Sociological methodology* 1970. San Francisco: Jossey-Bass, 151-169.
- Althauser, R. P., Heberlein, T. A. & Scott, R. A. 1971. A causal assessment of validity: The augmented multitrait-multimethod matrix. In: Blalock, H. M. (ed.): *Causal models in the social sciences*. London: Macmillan Press.
- Alwin, D. F. 1974. Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In: Costner, H. L. (ed.): *Sociological Methodology*, 1973/74. San Francisco: Jossey-Bass, 79-105.
- Amerine, M. A. & Roessler, E. B. 1976. *Wines: Their sensory evaluation*. San Francisco: Freeman & Co.

- Arrington, R. E. 1939. Time sampling studies of child behavior. *Psychological Monographs*, 51, 2.
- Arrington, R. E. 1943. Time sampling studies of social behavior: a critical review of techniques and results with research suggestions. *Psychological Bulletin*, 40, 81-124.
- Asendorpf, J. & Wallbott, H. G. 1979. Maße der Beobachterübereinstimmung: Ein systematischer Vergleich. *Zeitschrift für Sozialpsychologie*, 10, 243-252.
- Assakul, K. & Procter, C. H. 1967. Testing independence in two-way contingency tables with data subject to misclassification. *Psychometrika*, 32, 67-76.
- Avison, W. R. 1978. Auxiliary theory and multitrait-multimethod validation: A review of two approaches. *Applied Psychological Measurement*, 2, 431-447.
- Bales, R. F. 1950. *Interaction process analysis*. Cambridge, Mass.: Addison-Wesley.
- Bales, R. F. 1968. Die Interaktionsanalyse: Ein Beobachtungsverfahren zur Beobachtung kleiner Gruppen. In: König, R. (Hrsg.): *Beobachtung und Experiment in der Sozialforschung*. Köln: Kiepenheuer & Witsch, 6. Aufl., 148-170.
- Barker, R. 1951. *One boy's day*. New York. Harper & Row.
- Bartko, J. J. 1976. On various intraclass correlation reliability coefficients. *Psychological Bulletin*, 83, 762-765.
- Bergan, J. R. 1980. Measuring observer agreement using the quasi-independence concept. *Journal of Educational Measurement*, 17, 59-69.
- Bien, W. 1980. Ein Vergleich von Datenerhebungsmethoden zu kognitiven sozialen Strukturen durch die Analyse einer Multitrait-Multimethod-Matrix. Dissertation, RWTH Aachen.
- Bishop, Y. M., Fienberg, S. & Holland, P. 1974. *Discrete multivariate analysis: Theory and practice*. Boston: M. I. T. Press.
- Blalock, H. M. Jr. 1969. Multiple indicators and the causal approach to measurement error. *American Journal of Sociology*, 75, 264-272.
- Blalock, H. M. Jr. 1970. Estimating measurement error using multiple indicators and several points in time. *American Sociological Review*, 35, 101-111.
- Bock, R. D. 1956. The selection of judges for preference testing. *Psychometrika*, 21, 349-366.
- Böltkén, F. 1976. *Auswahlverfahren. Eine Einführung für Sozialwissenschaftler*. Stuttgart: Teubner (Studienskripten).
- Bohrnstedt, G. W. 1969. Observations on the measurement of change. In: Borgatta, E. F. (ed.): *Sociological Methodology 1969* San Francisco: Jossey-Bass, 113-133.
- Borg, I. 1977. Some basic concepts of facet theory. In: Lingoes, J. C. (ed.): *Geometric representations of relational data*. Ann Arbor, Mich.: Mathesis Press, 65-102.
- Boruch, R. F. & Wolins, L. 1970. A procedure for estimations of trait, method, and error variance attributable to a measure. *Educational and Psychological Measurement*, 30, 547-574.

- Boruch, R. F., Larkin, J. D., Wolins, L. & Mac Kinney, A. C. 1970. Alternative methods of analysis: Multitrait-multimethod data. *Educational and Psychological Measurement*, 30, 833-853.
- *Brannigan, C. R. & Humphries, D. A. 1972. Human non-verbal behavior, a means of communication. In: Blurton Jones, N. (ed.): *Ethological studies of child behavior*. Cambridge: University Press, 37-64.
- Bredenkamp, J. 1972. *Der Signifikanztest in der psychologischen Forschung*. Frankfurt: Akademische Verlagsgesellschaft.
- Bredenkamp, J. 1980. *Theorie und Planung psychologischer Experimente*. Darmstadt: Steinkopf.
- Brennan, R. L. & Light, R. J. 1974. Measuring agreement when two observers classify people into categories not defined in advance. *The British Journal of Mathematical and Statistical Psychology*, 27, 154-163.
- Bross, I. 1954. Misclassification in 2×2 tables. *Biometrics*, 10, 478-486.
- Caldwell, B. M. 1969. A new „approach“ to behavioural ecology. In: Hull, J. P. (ed.): *Minnesota Symposia on Child Psychology*, Vol. 2, Minneapolis: Univ. of Minnesota Press.
- Campbell, D. T. & Fiske, D. 1959. Convergent and discriminant Validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Campbell, D. T. & Stanley, J. C. 1963. Experimental and quasi-experimental designs for research in teaching. In: Gage, N. L. (ed.): *Handbook of research in teaching*. Chicago: Rand McNally, 171-246.
- Campbell, D. T. & O'Connell, E. J. 1967. Method factors in multimethod-multitrait matrices: Multiplicative rather than additive? *Multivariate Behavioral Research*, 2, 409-426.
- Canter, D. 1977. Is there a mapping sentence for architectural semiotics? Paper prepared for the Semiotics Symposium at Edra 8, Spring.
- Canter, D. 1977. Children in hospital: A facet theory approach to person/place synonymy. *Journal of Architectural Research*, 6, 20-32 (b).
- Centra, J. A. 1971. Validation by the multigroup-multiscale matrix: an adaptation of Campbell and Fiske's convergent and discriminant validation procedure. *Educational and Psychological Measurement*, 31, 675-683.
- Cicchetti, D. V., Aivano, S. L. & Vitale, J. 1977. Computer programs for assessing rater agreement and rater bias for qualitative data. *Educational and Psychological Measurement*, 37, 195-201.
- Clarke, A. H. & Ellgring, J. H. 1978. Verfahren zur halbautomatischen Bearbeitung von Videoaufzeichnungen. In: Heimchen, H. & Rengardt, E. (Hrsg.): *Fernsehen in der Psychiatrie*. Stuttgart: Thieme, 107-115.
- Cleary, T. A. & Linn, R. L. 1969. Error of measurement and the power of a statistical test. *British Journal of Mathematical and Statistical Psychology*, 22, 49-55.
- Cochran, W. G. 1953. *Sampling techniques*. New York: Wiley.

- Cochran, W. G. 1968. Errors of measurement in statistics. *Technometrics*, 10, 637-666.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- Cohen, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.
- Cohen, J. 1969. *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Conger, A. J. 1971. An evaluation of multimethod factor analysis. *Psychological Bulletin*, 75, 416-420.
- Conger, A. J. 1980. Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88, 322-328.
- Costner, H. L. 1969. Theory, deduction, and rules of correspondence. *American Journal of Sociology*, 75, 245-263.
- Costner, H. L. & Schoenberg, R. 1973. Diagnosing indicator ills in multiple indicator models. In: Goldberger, A. S. & Duncan, O. D. (eds): *Structural equation models in the social sciences*. New York: Seminar Press, 167-199.
- Cranach von, M. & Frenz, H. G. 1969. Systematische Beobachtung. In: Graumann, C. F. (Hrsg.): *Handbuch der Psychologie*, Band 7/1, Sozialpsychologie. Göttingen: Hogrefe, 269-331.
- Cronbach, L. J. & Meehl, P. E. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Cronbach, L. J., Rajaratnam, N. & Gleser, G. C. 1963. Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology*, 16, 137-163.
- Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. 1972. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley, 1972.
- Cureton, E. E. 1958. The average Spearman rank criterion correlation when ties are present. *Psychometrika*, 23, 271-272.
- Cureton, E. E. 1965. The average Spearman rank correlation when ties are present: A correlation. *Psychometrika*, 30, 377.
- Davis, F. B. 1974. *Standards for educational and psychological tests*. Washington: American Psychological Association.
- Deming, W. E. 1950. *Some theory of sampling*. New York: Wiley.
- Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26, 297-302.
- *Dickman, H. R. 1963. The perception of behavioral units. In: Barker, R. G. (ed.): *The stream of behavior*. New York: Appleton-Century-Crofts.
- Duncan, O. D. 1969. Some linear models for two-wave, two-variable panel analysis. *Psychological Bulletin*, 72, 177-182.

- Duncan, O. D. 1972. Unmeasured variables in linear models for panel analysis. In: Costner, H. L. (ed.): *Sociological Methodology 1972*. San Francisco: Jossey-Bass, 36-82.
- Duncan, O. D. 1975. Some linear models for two-wave, two variable panel analysis, with one-way causation and measurement error. In: Blalock, H. M. Jr., Aganbegian, A., Borodkin, F. N., Bondon, R. & Capeccch, V. (eds): *Quantitative Sociology: International perspectives on mathematical and statistical modeling*. New York: Academic Press, 285-306.
- Duncan, S. 1969. Nonverbal communication. *Psychological Bulletin*, 72, 118-137.
- Duncan, S. Jr. & Fiske, D. W. 1977. *Face-to-face interaction: Research, methods, and theory*. Hillsdale: Erlbaum.
- Ebel, R. L. 1951. Estimation of the reliability of ratings. *Psychometrika*, 16, 407-424.
- Everitt, B. S. 1968. Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 21, 97-103.
- Faßnacht, G. 1979. *Systematische Verhaltensbeobachtung*. München: Reinhardt (UTB Nr. 889).
- Feger, H. 1971. Probleme der Reliabilität und einer konvergierenden und diskriminierenden Validierung von Selbstbeobachtungsdaten. *Archiv für Psychologie*, 123, 1-16.
- Feger, H. 1978. *Konflikterleben und Konfliktverhalten*. Bern: Huber.
- Feger, H. 1979. Einstellungsstruktur und Einstellungsänderung: Ergebnisse, Probleme und ein Komponentenmodell der Einstellungsobjekte. *Zeitschrift für Sozialpsychologie*, 10, 331-349.
- Finn, R. H. 1970. A note on estimating the reliability of categorical data. *Educational and Psychological Measurement*, 30, 71-76.
- Finn, R. H. 1972. Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement*, 3.5, 255-265.
- Fischer, G. W. 1976. Multidimensional Utility models for risky and riskless choice. *Organizational Behavior and Human Performance*, 17, 127-146.
- Fishbein, M. & Ajzen, I. 1975. *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, Mass.: Addison-Wesley.
- Fleiss, J. L. 1966. Assessing the accuracy of multivariate observations. *Journal of the American Statistical Association*, 61, 403-412.
- Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382.
- Fleiss, J. L. 1973. *Statistical methods for rates and proportions*. New York: Wiley.
- Fleiss, J. L. 1975. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 31, 651-659.

- Fleiss, J. L. & Cohen, J. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33, 613-619.
- Fleiss, J. L. & ShROUT, P. E. 1978. Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika*, 43, 259-262.
- Fleiss, J. L., Cohen, J. & Everitt, B. S. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72, 323-327.
- Fleiss, J. L., Nee, J. C. M. & Landis, J. R. 1979. Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86, 974-977.
- Fleiss, J. L., Spitzer, R. L., Endicott, J. & Cohen, J. 1972. Quantification of agreement in multiple psychiatric diagnosis. *Archives of General Psychiatry*, 26, 168-171.
- Frey, S. & Pool, J. 1976. A new approach to the analysis of visible behavior. Forschungsbericht aus dem Psychologischen Institut der Universität Bern.
- Frick, R. & Semmel, M. I. 1978. Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research*, 48, 157-184.
- Friedrichs, J. & Lüdtkke, H. 1973. *Teilnehmende Beobachtung*. Beltz, Weinheim.
- Gleser, G. C., Cronbach, L. J. & Rajaratnam, N. 1965. Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 30, 395-418.
- Golding, S. L. & Seidman, E. 1974. Analysis of multitrait-multimethod matrices: A two step principal components procedure. *Multivariate Behavioral Research*, 9, 479-496.
- Goodman, L. A. & Kruskal, W. H. 1954. Measures of association for cross classification. *Journal of the American Statistical Association*, 49, 732-764.
- Goodman, L. A. & Kruskal, W. H. 1959. Measures of association for cross classification. II: Further discussion and references. *Journal of the American Statistical Association*, 54, 123-163.
- Goodman, L. A. & Kruskal, W. H. 1963. Measures of association for cross classification. III: Approximate sampling theory. *Journal of the American Statistical Association*, 58, 310-364.
- Goodman, L. A. & Kruskal, W. H. 1972. Measures of association for cross classification. IV: Simplification of asymptotic variances. *Journal of the American Statistical Association*, 67, 415-421.
- Graumann, C. F. 1966. Grundzüge der Verhaltensbeobachtung. In: Meyer, E. (Hrsg.): *Fernsehen in der Lehrerbildung*. München: Manz Verlag, 86-107.
- Green, D. M. & Swets, J. A. 1966. *Signal detection theory and psychophysics*. New York: Wiley.
- Grubbs, F. E. 1948. On estimating precision of measuring instruments and product variability. *Journal of the American Statistical Association*, 43, 243-264.
- Grubbs, F. E. 1973. Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics*, 15, 53-66.
- Grüner, K.-W. 1974. *Beobachtung*. Stuttgart. Teubner (Studienskripte).

- Guttman, L. 1959. A structural theory for intergroup beliefs and action. *American Sociological Review*, 24, 318-328.
- Guttman, L. 1971. Measurement as structural theory. *Psychometrika*, 36, 329-347.
- Guttman, L. & Guttman, R. 1976. A theory of behavioral generality and specificity during mild stress. *Behavioral Science*, 21, 469-477.
- Haggard, E. A. 1958. *Intraclass correlation and the analysis of variance*. New York: Dryden.
- Hannan, M. T., Robinson, R. & Warren, J. T. 1974. The causal approach to measurement error in panel analysis: Some further contingencies. In: Blalock, H. M. Jr. (ed.): *Measurement in the social sciences: Theories and strategies*. Chicago: Aldine, 293-323.
- Hansen, M. H., Hurwitz, W. N. & Madow, W. G. 1953. *Sample survey methods and theory*. New York: Wiley.
- Hasemann, K. 1964. Verhaltensbeobachtung. In: Heiss, R. (Hrsg.): *Handbuch der Psychologie*, Band 6, *Psychologische Diagnostik*. Göttingen: Hogrefe, 807-836.
- Hauser, R. M. & Goldberger, A. S. 1971. The treatment of unobservable variables in path analysis. In: Costner, H. L. (ed.): *Sociological Methodology 1971*. San Francisco: Jossey-Bass, 81-117.
- Hayes, D. P., Meltzer, L. & Wolf, G. 1970. Substantive conclusions are dependent upon techniques of measurement. *Behavioral Science*, 15, 265-268.
- Heise, D. R. 1969. Separating reliability and stability in test-retest correlation. *American Sociological Review*, 34, 93-101. (Nachdruck in: Blalock, H. M. (ed.): *Causal models in the social sciences*. London: MacMillan, 1971.)
- Heise, D. R. 1970. Causal inference from panel data. In: Borgatta, E. F. & Bohrnstedt, G. W. (eds): *Sociological Methodology 1970*. San Francisco: Jossey-Bass, 3-27.
- Heise, D. R. 1975. *Causal analysis*. New York: Wiley.
- Hendel, D. D. & Weiss, D. J. 1970. Individual inconsistency and reliability of measurement. *Educational and Psychological Measurement*, 30, 579-593.
- Heyns, R. W. & Zander, A. F. 1953. Observation of group behavior. In: Festinger, L. & Katz, D. (eds): *Research methods in the behavioral sciences*. New York: Dryden Press, 381-417.
- Hildebrand, D. K., Laing, J. D. & Rosenthal, H. 1977. *Analysis of ordinal data*. Beverly Hills: Sage Publications.
- Hollenbeck, A. R. 1978. Problems of reliability in observational research. In: Sackett, G. P. (ed.): *Observing behavior*. Vol. II. Baltimore: Univ. Park Press, 79-98.
- Holley, J. W. & Guilford, J. P. 1964. A note on the G index of agreement. *Educational and Psychological Measurement*, 24, 749-753.
- Holley, J. W. & Lienert, G. A. 1974. The G index of agreement in multiple ratings. *Educational and Psychological Measurement*, 34, 817-822.
- Huber, H. P. 1973. *Psychometrische Einzelfalldiagnostik*. Weinheim: Beltz.

- Hubert, L. 1977. Kappa revisited. *Psychological Bulletin*, 84, 289-297.
- Hubert, L. 1977. Nominal scale response agreement as a generalized correlation. *British Journal of Mathematical and Statistical Psychology*, (a), 30, 98-103.
- Hubert, L. J. 1979. Comparison of sequences. *Psychological Bulletin*, 86, 1098-1106.
- Hubert, L. J. 1979. Generalized concordance. *Psychometrika*, (a), 44, 135-142.
- Hubert, L. 1979. Matching models in the analysis of cross-classifications. *Psychometrika*, (b), 44, 21-41.
- Hubert, L. J. & Baker, F. B. 1978. Analyzing the multitrait-multimethod matrix. *Multivariate Behavioral Research*, 13, 163-179.
- Hubert, L. J. & Baker, F. B. 1978. Evaluating the conformity of sociometric measurements. *Psychometrika*, (a), 43, 31-41.
- *Hubert, L. J. & Baker, F. B. 1979. A note on analysing the multitrait-multimethod-matrix: An application of a generalized proximity function comparison. *British Journal of Mathematical and Statistical Psychology*, 32, 179-184.
- Huck, S. W. 1978. A modification of Hoyt's analysis of variance reliability estimation procedure. *Educational and Psychological Measurement*, 38, 725-736.
- Hunter, J. E. & Cohen, S. H. 1974. Correcting for unreliability in nonlinear models of attitude Change. *Psychometrika*, 39, 445-468.
- Hutt, S. J. & Hutt, C. 1974. *Direct Observation and measurement of behavior*. Springfield: Thomas, 2. Aufl.
- Jackson, D. N. 1969. Multimethod factor analysis in the evaluation of convergent and discriminant validity. *Psychological Bulletin*, 72, 30-49.
- Jackson, D. N. 1975. Multimethod factor analysis: A reformulation. *Multivariate Behavioral Research*, 10, 259-275.
- Jacobson, A. L. & Lalw, N. M. 1974. An empirical and algebraic analysis of alternative techniques for measuring unobserved variables. In: Blalock, H. M. Jr. (ed.): *Measurement in the social sciences*. Chicago: Aldine, 215-242.
- Jahoda, M., Deutsch, M. & Cook, S. W. 1968. Beobachtungsverfahren. In: König, R. (Hrsg.): *Beobachtung und Experiment in der Sozialforschung*. Köln: Kiepenheuer & Witsch, 6. Aufl., 77-96.
- Janson, S. & Vegelius, J. 1979. On generalizations of the G-index and the phi-coefficient to nominal scales. *Multivariate Behavioral Research*, 14, 255-269.
- Joe, G. W. & Woodward, J. A. 1976. Some developments in multivariate generalizability. *Psychometrika*, 41, 205-217.
- Jöreskog, K. G. 1969. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183-202.
- Jöreskog, K. G. 1970. A general method for analysis of covariance structures. *Biometrika*, 57, 239-251.
- Jöreskog, K. G. & Sörbom, D. 1976. Statistical models and methods for analysis of longitudinal data. In: Aigner, D. J. & Goldberger, A. S. (eds): *Latent variables in socioeconomic models*. Amsterdam: North Holland.

- Jöreskog, K. G. & Sörbom, D. 1978. LISREL - IV: A general computer program for estimation of linear structural equations by maximum likelihood methods. User's Guide. Chicago: International Educational Service.
- Jordan, J. E. 1971. Construction of a Guttman facet designed crosscultural attitude-behavior scale toward mental retardation. *American Journal of Mental Deficiency*, 76, 201-219.
- Kalbermatten, U. & v. Cranach, M. 1980. Hierarchisch aufgebaute Beobachtungssysteme zur Handlungsanalyse. In: Winkler, P. (Hrsg.): *Methoden zur Analyse von Face-to-Face Situationen*. Stuttgart: Metzler.
- Kalleberg, A. L. & Kluegel, J. R. 1975. Analysis of the multitrait-multimethod matrix: Some limitations and an alternative. *Journal of Applied Psychology*, 60, 1-9.
- Kaufman, I. C. & Rosenblum, L. A. 1966. A behavioral taxonomy for *M. nemestrina* and *M. radiata*; based on longitudinal observation of family groups in the laboratory. *Primates*, 7, 205-258.
- Kavanagh, M. J., MacKinney, A. C. & Wolins, L. 1971. Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin*, 75, 34-39.
- Kaye, K. 1980. Estimating false alarms and missed events from interobserver agreement: A rationale. *Psychological Bulletin*, 88, 458-468.
- Kendall, M. G. 1948. *Rank correlation methods*. London: Griffin.
- *Kenny, D. A. 1976. An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology*, 12, 247-252.
- Kent, R. N. & Foster, S. L. 1977. Direct observational procedures: Methodological issues in naturalistic settings. In: Ciminero, A. R., Calhoun, K. S. & Adams, H. E. (eds): *Handbook of behavioral assessment*. New York: Wiley, 279-328.
- Keys, A. & Kihlberg, J. K. 1963. Effect of misclassification on estimated relative prevalence of a characteristic. *American Journal of Public Health*, 53, 1656-1665.
- Kish, L. 1953. Selection of a Sample. In: Festinger, L. & Katz, D. (eds): *Research methods in the behavioral sciences*. London: Staples Press.
- Klauer, K. J. 1978. Kontentvalidität. In: Klauer, K. J. (Hrsg.): *Handbuch der pädagogischen Diagnostik*. Band 1. Düsseldorf: Schwann, 225-255.
- Kluckhohn, F. R. 1956. Die Methode der teilnehmenden Beobachtung. In: König, R. (Hrsg.): *Beobachtung und Experiment in der Sozialforschung*. Köln: Kiepenheuer & Witsch, (3. Aufl. 1965), 97-114.
- Koch, G. G. 1969. The effect of non-sampling errors on measures of association in 2 x 2 contingency tables. *Journal of the American Statistical Association*, 64, 852-863.
- König, R. 1962. Die Beobachtung. In: König, R. (Hrsg.): *Handbuch der empirischen Sozialforschung*, Band I. Stuttgart: Enke, 107-135.

- Kraemer, H. C. & Korner, A. F. 1976. Statistical alternatives in assessing reliability, consistency, and individual differences for quantitative measures: Application to behavior measures of neonates. *Psychological Bulletin*, 83, 914-921.
- Krause, M. S. 1972. The implications of convergent and discriminant validity data for instrument Validation. *Psychometrika*, 37, 179-186.
- Krippendorff, K. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30, 61-70.
- Krippendorff, K. 1970. Bivariate agreement coefficients for reliability of data. In: Borgatta, E. F. & Bohrnstedt, G. W. (eds): *Sociological methodology 1970*. San Francisco: Jossey-Bass (a).
- Kruglanski, A. W. 1975. The human subject in the psychological experiment: Fact and artifact. In: Berkowitz, L. (ed.): *Advances in experimental social psychology*, Vol. 8. New York: Academic Press, Vol. 8, 101-147.
- Kruskal, W. H. 1958. Ordinal measures of association. *Journal of the American Statistical Association*, 53, 814-861.
- Landis, J. R. & Koch, G. G. 1975. A review of statistical methods in the analysis of data arising from observer reliability studies. *Statistica Neerlandica*, 29, Part I: 101-123, Part II: 151-161.
- Landis, J. R. & Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Landis, J. R. & Koch, G. G. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, (a), 33, 363-374.
- Lawlis, G. F. & Lu, E. 1972. Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin*, 78, 17-20.
- Leach, G. M. 1972. A comparison of the social behavior of some normal and problem children. In: Blurton Jones, N. (ed.): *Ethological studies of Child behavior*. Cambridge: Univ. Press, 249-281.
- Levy, P. 1974. Generalizability studies in clinical settings. *British Journal of Social and Clinical Psychology*, 13, 161-172.
- Levy, S. & Guttman, L. 1975. Structure and dynamics of worries. *Sociometry*, 38, 445-473.
- *Lienert, G. A. 1961. *Testaufbau und Testanalyse*. Weinheim: Beltz, 2. Aufl. 1967.
- Lienert, G. A. 1973. *Verteilungsfreie Methoden in der Biostatistik*. Band I. Meisenheim: Hain.
- Light, R. J. 1971. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, 76, 365-377.
- Lisch, R. & Kriz, J. 1978. *Grundlagen und Modelle der Inhaltsanalyse*. Reinbek bei Hamburg: Rowohlt.

- Longabaugh, R. 1980. The systematic observation of behavior in naturalistic settings. In: Triandis, H. C. & Berry, J. W. (eds): *Handbook of Cross-cultural psychology*. Vol. 2: Methodology. Boston: Allyn & Baton, pp. 57-126.
- Lord, F. M. & Novick, M. R. 1968. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.
- Lu, K. H. 1971. Statistical control of „impurity“ in the estimation of test reliability. *Educational and Psychological Measurement*, 31, 641-655.
- Luce, R. D. 1963. Detection and recognition. In: Luce, R. D., Bush, R. R. & Galanter, E. (eds): *Handbook of mathematical psychology*. Vol. I. New York: Wiley, 103-189.
- Luce, R. D. & Galanter, E. 1963. Discrimination. In: Luce, R. D., Bush, R. R. & Galanter, E. (eds): *Handbook of mathematical psychology*. Vol. I. New York: Wiley, 191-243 (a).
- Luce, R. D. & Galanter, E. 1963. Psychophysical scaling. In: Luce, R. D., Bush, R. R. & Galanter, E. (eds): *Handbook of mathematical psychology*. Vol. I. New York: Wiley, 245-307 (b).
- Lyerly, S. B. 1952. The average Spearman rank correlation coefficient. *Psychometrika*, 17, 421-428.
- Manz, W. 1974. Beobachtung verbaler Kommunikation im Laboratorium. In: van Koolwijk, J. & Wieken-Mayser, M. (Hrsg.): *Techniken der empirischen Sozialforschung. Band III, Erhebungsmethoden: Beobachtung und Analyse von Kommunikation*. München: Oldenbourg, 27-65.
- Maxwell, A. E. & Pilliner, A. E. G. 1968. Deriving coefficients of reliability and agreement for ratings. *British Journal of Mathematical and Statistical Psychology*, 21, 105-116.
- Mayer, L. S. & Younger, M. S. 1974. Multiple indicators and the relationship between abstract variables. In: Heise, D. P. (ed.): *Sociological methodology 1975*. San Francisco: Jossey-Bass, 191-211.
- McCall, G. J. & Simons, J. L. 1969. *Issues in participant Observation: A text and reader*. Reading, Mass.: Addison-Wesley.
- McDermott, P. A. & Watkins, M. W. 1979. A Fortran program for testing agreement of multiple observers with a categorical standard on nominal scales. *Educational and Psychological Measurement*, 39, 669-672.
- McGrew, W. C. 1972. *An ethological study of children's behavior*. New York: Academic Press.
- McKinlay, S. M. 1975. The design and analysis of the observational study - A review. *Journal of the American Statistical Association*, 70, 503-520.
- McNicol, D. 1972. *A primer of signal detection theory*. London: Allen & Unwin.
- Medley, D. M. & Mitzel, H. E. 1970. Measuring classroom behavior by systematic Observation. In: Gage, N. L. (ed.): *Handbook of research on teaching*. Chicago: Rand McNally, 1963, 247-328. Deutsche Bearbeitung: Schulz, W., Teschner, W. P. & Voigt, J.: *Verhalten im Unterricht. Seine Erfassung durch Beobachtungs-*

- verfahren. In: Ingenkamp, K. (Hrsg.): Handbuch der Unterrichtsforschung, Teil I. Weinheim: Beltz, 632-851.
- Mees, U. & Selg, H. 1977. Verhaltensbeobachtung und Verhaltensmodifikation. Stuttgart: Klett.
- Mote, V. L. & Anderson, R. L. 1965. An investigation of the effect of misclassification on the properties of chi square tests in the analysis of categorical data. *Biometrika*, 52, 95-109.
- Munroe, R. L. 1973. Differential predictions based on time sampling versus event sampling behavior observations. Paper presented at Mathematical Social Science Board Conference on Human Behavior Observations. Monroeville, Pa. October.
- Nay, W. R. & Kerkhoff, T. 1974. Informational feedback in training behavioral coders. *Psychological Reports*, 35, 1175-1181.
- Naylor, J. C. & Schenck, E. A. 1966. Q_m as an „error free“ index of rater agreement. *Educational and Psychological Measurement*, 26, 815-824.
- Naylor, J. C., Dudycha, A. L. & Schenck, E. A. 1967. An empirical comparison of Q_α and Q_m as indices of rater policy agreement. *Educational and Psychological Measurement*, 27, 7-20.
- Newtonson, D. 1973. Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, 28, 28-38.
- Newtonson, D. 1976. Foundations of attribution: The unit of perception of ongoing behavior. In: Harvey, J., Ickes, W. & Kidd, R. (eds): *New directions in attribution research*. Hillsdale, N. J.: Erlbaum, 223-247.
- Newtonson, D. & Engquist, G. 1976. The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology*, 12, 436-450.
- Newtonson, D., Engquist, G. & Bois, J. 1977. The objective basis of behavior units. *Journal of Personality and Social Psychology*, 35, 847-862.
- Nicewander, W. A. & Price, J. M. 1978. Dependent variable reliability and the power of significance tests. *Psychological Bulletin*, 85, 405-409.
- Nußbaum, A. 1980. Konstruktion, Planung und Analyse lehrzielorientierter Tests auf der Grundlage der Generalisierbarkeitstheorie. Dissertation, RWTH Aachen.
- Overall, J. E. 1968. Estimating individual rater reliabilities from analysis of treatment effects. *Educational and Psychological Measurement*, 28, 255-264.
- Peak, H. 1953. Problems of objective observation. In: Festinger, L. & Katz, D. (eds): *Research methods in the behavioral sciences*. New York: Holt, Rinehart, and Winston, 243-299.
- Pearson, K. 1901. Mathematical contributions to the theory of evolution. IX On the principle of homotyposis and its relation to heredity, to variability of the individual, and to that of rate. Part 1: Homotyposis in the vegetable kingdom *Philosophical Transactions of the Royal Society (London)*, Series A, 197, 385-479.
- Polansky, N., Freeman, W., Horowitz, M., Irwin, L., Paponia, N., Rapaport, D. & Whaley, F. 1949. Problems of interpersonal relations in research on groups. *Human Relation*.

- Rajaratnam, N. 1960. Reliability formtdas for independent decision data when reliability data are matched. *Psychometrika*, 25, 261-271.
- Ray, M. L. & Heeler, R. M. 1975. Analysis techniques for exploratory use of the multitrait-multimethod matrix. *Educational and Psychological Measurement*, 35, 255-265.
- Reid, J. B. 1970. Reliability assessment of observational data: a possible methodological problem. *Child Development*, 41, 1143-1150.
- Rock, D. A., Werts, C. E., Linn, R. L. & Jöreskog, K. G. 1977. A maximum likelihood Solution to the errors in variables and errors in equation model. *Multivariate Behavioral Research*, 12, 187-197.
- Rogot, E. 1961. A note on measurement errors and detecting real differences. *Journal of the American Statistical Association*, 56, 319-319.
- Romanczyk, R. G., Kent, R. N., Diamant, C. & O'Leary, K. D. 1973. Measuring the reliability of observational data: a reactive process. *Journal of Applied Behavior Analysis*, 6, 175-184.
- Rosenthal, R. 1965. The volunteer subject. *Human Relations*, 18, 389-406.
- Rosenthal, R. & Rosnow, R. L. (eds). 1969. *Artifact in behavioral research*. New York: Academic Press.
- Rosenthal, R. & Rosnow, R. L. 1975. *The volunteer subject*. New York: Wiley.
- Rudinger, G. & Feger, H. 1970. Die Beurteilung formaler Verhaltensmerkmale durch Rating-Skalen: Eine Generalisierbarkeitsstudie. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 2, 96-112.
- Sackett, G. P. 1978. Measurement in observational research. In: Sackett, G. P. (ed.): *Observing behavior*. Vol. II. Baltimore: Univ. Park Press, 25-43.
- Sackett, G. P., Ruppenthal, G. C. & Gluck, J. 1978. Introduction: An overview of methodological and statistical problems in observational research. In: Sackett, G. P. (ed.): *Observing behavior*. Vol. II. Baltimore: Univ. Park Press, 1-14.
- Scherer, K. R. 1974. Beobachtungsverfahren zur Mikroanalyse nonverbaler Verhaltensweisen. In: van Koolwijk, J. & Wieken-Mayser, M. (Hrsg.): *Techniken der empirischen Sozialforschung*. München: Oldenbourg.
- Schmitt, N. 1978. Path analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, 2, 157-173.
- Schmitt, N., Coyle, B. W. & Saari, B. B. 1977. A review and critique of analyses of multitrait-multimethod matrices. *Multivariate Behavioral Research*, 12, 447-478.
- Schoggen, P. 1976. Ecological psychology and mental retardation. Paper presented at Conference on the Application of Observation/Ethological Methods in the Study of Mental Retardation. Washington, D. C., June.
- Schucany, W. R. & Frawley, W. H. 1973. A rank test for two group concordance. *Psychometrika*, 38, 249-258.
- Schutz, W. C. 1952. Reliability, ambiguity, and content analysis. *Psychological Review*, 59, 119-129.

- Schwartz, M. S. & Schwartz, C. G. 1955. Problems in participant observation. *American Journal of Sociology*, 66.
- Scott, W. A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19, 321-325.
- Scott, W. A. & Wertheimer, M. 1962. Introduction to psychological research. New York: Wiley, (4. Aufl. 1967).
- Selvage, R. 1976. Comments on the analysis of variance strategy for the computation of intraclass reliability. *Educational and Psychological Measurement*, 36, 605-609.
- Simon, A. & Boyer, E. G. 1974. Mirrors for behavior III. An anthology of Observation instruments. Communication Materials Center, Wyncote, Pennsylvania.
- Simons, G. & Papousek, H. 1978. Methoden der Kleinkindforschung: Beobachtung und Experiment. In: Dollase, R. (Hrsg.): *Handbuch der Früh- und Vorschulpädagogik*. Band 2. Düsseldorf: Schwann, 93-110.
- *Singleton, W. T. 1972. Techniques for determining the causes of error. *Applied Ergonomics*, 3, 126-131.
- *Somers, R. H. 1962. A new asymmetric measure of association. *American Sociological Review*, 1962, 27, 799-811.
- Sorembe, V.: Computer program for and some comments on K. Krippendorff's 'Estimating the reliability, systematic error and random error of interval data'. Institut für Psychologie der RWTH Aachen, Arbeitsbericht, o.J.
- Spearman, C. 1906. „Footrule“ for measuring correlation. *British Journal of Psychology*, 2, 89-108.
- Stanley, J. C. 1961. Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. *Psychometrika*, 26, 205-219.
- Stewart, R. A., Powell, G. E. & Chetwynd, S. J. 1979. Person perception and stereotyping. Westmead: Saxon House.
- Stroud, T. W. F. 1974. Comparing regressions when measurement error variances are known. *Psychometrika*, 39, 53-68.
- Susman, E. J., Peters, D. J. & Steward, R. B. 1976. Naturalistic observational child study, a review. Paper presented at the 4th Biennial Southeastern Conference on Human Development at Nashville, Tennessee.
- Sutcliffe, J. P. 1958. Error of measurement and the sensitivity of a test of significance. *Psychometrika*, 28, 9-17.
- Sutcliffe, J. P. 1980. On the relationship of reliability to statistical power. *Psychological Bulletin*, 88, 509-515.
- Taylor, W. L. 1964. Correcting the average rank correlation coefficient for ties in the rankings. *Journal of the American Statistical Association*, 59, 872-876.
- Taylor, W. L. & Fong, C. 1963. Some contributions to average rank correlation methods and to the distribution of the average rank correlation coefficient. *Journal of the American Statistical Association*, 58, 756-769.

- Tesser, A. & Krauss, H. 1976. On validating a relationship between constructs. *Educational and Psychological Measurement*, 36, 111-121.
- Thomae, H. 1959. Forschungsmethoden der Entwicklungspsychologie. In: Thomae, H. (Hrsg.): *Handbuch der Psychologie*, Band 3, Entwicklungspsychologie. Göttingen: Hogrefe, 46-75.
- Thomae, H. 1968. *Das Individuum und seine Welt*. Göttingen: Hogrefe.
- Tinsley, H. E. A. & Weiss, D. J. 1975. Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22, 358-376.
- Trippi, R. R. & Settle, R. B. 1976. A nonparametric coefficient of internal consistency. *Multivariate Behavioral Research*, 11, 419-424.
- Tucker, L. R. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31, 279-311.
- Vegelius, J. 1977. An ordinal scale generalization of the G index invariant over item reflection. *Educational and Psychological Measurement*, 37, 31-35.
- Vegelius, J. 1977. On the weighted G index. *Educational and Psychological Measurement*, (a), 37, 839-842.
- Vegelius, J. 1979. A G index generalization for trichotomized data. *Educational and Psychological Measurement*, 39, 23-27.
- Wackerly, D. D., McClave, J. T. & Rao, P. V. 1978. Measuring nominal scale agreement between a judge and a known standard. *Psychometrika*, 43, 213-223.
- Watkins, M. W. & McDermott, P. A. 1979. A Computer program for measuring levels of Overall and partial congruence among multiple observers on nominal scales. *Educational and Psychological Measurement*, 39, 235-239.
- *Webb, E. J., Campbell, D. T., Schwartz, R. D. & Sechrest, L. 1966. *Unobtrusive measures: nonreactive research in the social sciences*. Chicago: Rand McNally.
- Weick, K. E. 1968. Systematic observational methods. In: Lindzey, G. & Aronson, E. (eds): *The handbook of social psychology*. Vol. 2. Reading, Mass.: Addison-Wesley, 2. Aufl., 357-451.
- Werner, J. 1976. Varianzanalytische Maße zur Reliabilitätsbestimmung von ratings. *Zeitschrift für experimentelle und angewandte Psychologie*, 23, 489-500.
- Werts, C. L. & Linn, R. L. 1970. Path analysis: Psychological examples. *Psychological Bulletin*, 74, 193-212.
- Werts, C. E. & Linn, R. L. 1971. Analyzing school effects: Ancova with a fallible covariate. *Educational and Psychological Measurement*, 31, 95-105.
- Werts, C. E., Jöreskog, K. G. & Linn, R. L. 1971. Comment on 'The Estimation of Measurement Error in Panel Data'. *American Sociological Review*, 36, 110-113.
- *Werts, C. E., Jöreskog, K. G. & Linn, R. L. 1973. Identification and estimation in path analysis with unmeasured variables. *American Journal of Sociology*, 78, 1469-1484.

- Werts, C. E., Linn, R. L. & Jöreskog, K. G. 1974. Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement*, 34, 25-33.
- Wheaton, B., Mutherr, B., Alwin, D. F. & Summers, G. F. 1977. Assessing reliability and stability in panel models. In: Heise, D. R. (ed.): *Sociological Methodology* 1977. San Francisco: Jossey-Bass, 84-136.
- *Whelan, P. 1974. Reliability of human observers. Doctoral Diss., Univ. of Utah at Salt Lake.
- Whiting, J. W. M., Child, I. L. & Lambert, W. W. 1966. Field guide for a study of socialization: six cultures series, Vol. I. New York: Wiley.
- Whyte, W. F. 1953. Observational field work methods. In: Jahoda, M., Deutsch, M. & Cook, S. W. (eds): *Research methods in the social sciences*. New York: Dryden, 493-513.
- Wiley, D. E. & Wiley, J. A. 1970. The estimation of measurement error in panel data. *American Sociological Review*, 35, 112-117. (Nachdruck in: Blalock, H. M. (ed.): *Causal models in the social sciences*. London: Macmillan, 1971).
- Wilson, T. P. 1974. Measures of association for bivariate ordinal hypothesis. In: Blalock, H. M. (ed.): *Measurement in the social sciences*. Chicago: Aldine-Atherton, 327-342.
- Winer, B. J.: *Statistical principles in experimental design*. New York: McGraw-Hill, 1. Ausg. 1962; 2. Ausg. 1971.
- Wright, H. F. 1960. Observational child study. In: Mussen, P. (ed.): *Handbook of research methods in child development*. New York: Wiley, 71-139.
- Yates, F. 1953. *Sampling methods for censuses and surveys*. London: Griffirr, 2. Aufl.