# Die Bedeutungsueberschneidung von Beschreibungskategorien als Problem der Unterrichtsforschung - Eine methodenkritische Untersuchung am Beispiel des Ratingverfahrens

Klaus Beck

FORSCHUNGSBERICHTE

AUS DEM

OTTO - SELZ - INSTITUT

FÜR

PSYCHOLOGIE UND ERZIEHUNGSWISSENSCHAFT

DER

UNIVERSITÄT MANNHEIM (WH)

# Die Bedeutungsueberschneidung von Beschreibungskategorien als Problem der Unterrichtsforschung - Eine methodenkritische Untersuchung am Beispiel des Ratingverfahrens

Klaus Beck

Forschungsbericht Nr. 6

Mai 1980

Otto-Selz-Institut
fuer Psychologie und
Erziehungswissenschaft
Universitaet Mannheim

Diese Untersuchung wurde aus Mitteln des Otto-Selz-Instituts fuer Psychologie und Erziehungswissenschaft der Universitaet Mannheim gefoerdert.

## Ueberblick

In der Literatur wird haeufig beklagt, dass sich vor den Bemuehungen um eine gesicherte empirische Basis der Unterrichtsforschung - ueber die allgemeinen erkenntnistheoretischen Probleme hinaus - besondere Schwierigkeiten auftuermten. Ihre Umgehung sei nur auf zwei Wegen moeglich, die aber beide den Verzicht auf wesentliche Guetestandards erforderlich machten; der eine verlaufe zwar auf dem zuverlaessigen Fundament praeziser Verhaltensbeschreibungen, ende jedoch bei Informationen ohne theoretische Relevanz (z.B. Lidschlagfrequenz), der andere fuehre zu adaequaten theoretischen Aussagen, entbehre aber der soliden methodischen Grundlage. Angesichts dieses Dilemmas, naemlich hier unzulaengliche Ergebnisse, dort unzuverlaessige Mittel, waehlen viele Forscher die zweite Alternative in der Hoffnung, die mit ihr verbundenen Nachteile durch ein besonderes Verfahren in ausreichendem Masse beseitigen zu koennen; das Rating. Tatsaechlich liegt eine grosse Zahl von Arbeiten vor, in denen die methodischen Aspekte der Schaetz- oder Beurteilungsverfahren untersucht werden mit dem Ziel, solche Maengel zu minimieren oder doch wenigstens berechenbar zu machen.

Ordnet man die einzelnen Fehlerquellen in der Ratingprozedur nach der Chronologie ihrer Entstehung im Prozess der Datenerhebung, so laesst sich unsere Fragestellung am Beginn der ersten Phase lokalisieren, naemlich bei der Datenaufnahme durch den Beobachter; es wird untersucht, mit welchen qualitativen und quantitativen Konsequenzen gerechnet werden muss, wenn Unterrichtsbeobachtung unter mehreren Aspekten erfolgt und die Vorgabe dieser Aspekte in einer theoretischen bzw. abstrakten Sprache abgefasst ist (z.B. "entspannt", "einfuehlsam"). Man muss damit rechnen, dass aufgrund von Ueberschneidungen im Bedeutungsfeld solcher Kategorien Uebereinstimmungen zwischen den dazu erhobenen Daten auftreten. Gemeinsamkeiten von dieser Art sind allein im Sprachgebrauch begruendet und hinsichtlich des Ziels der Realitaetserfassung als artifiziell zu bezeichnen.

In der vorliegenden Untersuchung wird die Bedeutungsueberschneidung von Beobachtungskategorien zunaechst unter methodischen und theoretischen Gesichtspunkten eroertert und systematisiert. Der "Bedeutungsfehler" erweist sich dabei als Validitaetsproblem mit erheblichen inhaltlichen und anwendungsrelevanten Implikationen.

Eine empirische Studie, die sowohl konventionelle Ratings als auch Sprachaehnlichkeiten erfasst, dient der Beschaffung von Informationen, mit deren Hilfe Ausmass und Folgen

von Bedeutungsueberschneidungen abgeschaetzt werden koennen. Es wird gezeigt, dass diese Prozedur unter verschiedenen sprachtheoretischen Voraussetzungen durchzufuehren ist und die Ergebnisse auf diese Varaussetzungen zu relativieren sind. Die dazu in der Literatur vertretenen Auffassungen werden idealtypisch auf zwei polare Positionen reduziert: "Privatsprache" und "Einheitssprache".

Die Analysen fuer beide Positionen liefern Resultate, die eine Vernachlaessigung des Bedeutungsfehlers nicht als vertretbar erscheinen lassen: seine Beseitigung fuehrt zu teilweise voellig veraenderten Datenstrukturen und damit zu Aussagen ueber Unterricht, die mit denjenigen, die auf der Basis unbereinigter Daten formuliert werden, logisch und inhaltlich nicht kompatibel sind. Dieser Befund stabilisiert sich in der anschliessend vorgenommenen methodischen "Gegenprobe"; allein aus der Kenntnis der semantischen Struktur der verwendeten Beobachtungswoerter lassen sich im nachhinein Ergebnisse, die bei Untersuchungen auf Ratingbasis gefunden wurden, mit betraechtlicher Genauigkeit "prognostizieren".

Die Arbeit schliesst mit der Eroerterung von Konsequenzen, die aus der Beachtung des Bedeutungsfehlers gezogen werden koennten. Zwei Strategien stehen zur Diskussion: die erste reflektiert, welche Massnahmen zu seiner Vermeidung oder Minimierung zu ergreifen waeren, wenn das Ratingverfahren selbst beibehalten werden soll. Dabei ergibt sich auch, dass die bereits vorliegenden quantitativen Analysen weiterer, chronologisch spaeterer Ratingfehler (z.B. Halo-, Milde - Effekt) unter Beachtung des Bedeutungsproblems revidiert werden muessten, bevor ihre abschliessende Beurteilung moeglich ist.

Die zweite Strategie vermeidet angesichts der dargestellten Maengel das Rating als Instrument zur Erfassung der Unterrichtsrealitaet. Einige Ueberlegungen zu alternativen Designs fuer die Unterrichtsforschung werden skizziert und eines davon anhand der vorliegenden, neu interpretierten Daten illustriert.

# Inhaltsverzeichnis

Wegen der maschinellen Textverarbeitung sind die Abbildungen zusammengefasst worden (S. 120-125). Die Seiten koennen herausgeklappt werden.

Anmerkungen sind durch "<>" gekennzeichnet.

## Verzeichnis der Abbildungen, Tabellen und Gleichungen

## 1. Problemstellung

Obgleich im Schrifttum zur empirischen Sozialforschung stets auf die gravierenden Maengel hingewiesen wird, die mit der Anwendung von Schaetzskalen verbunden sind, scheint sich diese Methode der Datenerhebung im Bereich der Unterrichtsforschung nach wie vor hoher Beliebtheit zu erfreuen. So berichten Rosenshine und Furst (1973, 165) aus den USA von ueber zweihundert verfuegbaren und angewendeten Forschungsinstrumenten auf Rating-Basis. Zu aehnlichen Groessenordnungen gelangt man auch im deutschsprachigen Bereich (vgl. Tent 1971, 864), der vor allem durch die Untersuchungen von Tausch und Tausch (z.B. 1973) gepraegt und, was die Itemauswahl betrifft, eng mit der amerikanischen Unterrichtsforschung verbunden ist (man denke auch etwa an die breite Rezeption der Flandersschen Interaktionsanalyse (197o)).

Ohne zu uebertreiben, wird man sagen koennen, dass der ueberwiegende Teil unserer empirischen Kenntnis unterrichtlicher Lehr-Lern-Prozesse auf Informationen beruht, die unter Verwendung von durch Schaetzung ermittelten subjektgebundenen Sekundaerdaten (vgl. Rosenshine 1970, 445) gewonnen wurden (vgl. z.B. Davis 1964, 48-62, Tent 1971, Doering 1973, 47-7o, Keil/Piontowski 1973, 35-36). Mit diesem Begriff werden jene Elemente unseres Wissens von der (Unterrichts-) Realitaet bezeichnet, derer wir wesentlich durch die "Vermittlung" eines wahrnehmenden Subjekts habhaft werden. Waehrend etwa bei der Bestimmung von Laenge, Zeit oder Intelligenz das zu messende Merkmal oder sein (gedachter) Indikator messtheoretisch "unproblematisch" zu erfassen sind (z.B. Strecke auf einer Messlatte, Zeigerstellung, Ankreuzung von Aufgabenloesungen) und insofern "Primaerdaten" erhoben werden, beruhen im hier zu besprechenden Fall die Daten auf Aussagen von Personen, die der Wirkung des zu messenden Merkmals (oder Indikators) ausgesetzt waren und ueber ihre Wahrnehmungen berichten (z.B. Angaben ueber das Einfuehlungsvermoegen eines Lehrers).

Der entscheidende Unterschied zwischen beiden Wissensarten liegt darin, dass ihre "Objektivierung" in unterschiedlicher "kausaler Distanz" zum Untersuchungsgegenstand erfolgt und dass diese "Distanz" beim Rating durch eine Person, ein wahrnehmendes Bewusstsein, ueberbrueckt wird (vgl. Remmers 1971, Sp. 857, Langer/Schulz v. Thun 1974, 15).<1>

Die mit der "Einschaltung" einer "beobachtenden" Person verbundenen Messfehlermoeglichkeiten sind generell und, wie bereits erwaehnt, auch speziell fuer Schaetzskalen in der einschlaegigen Methodenliteratur seit langem

ausfuehrlich diskutiert worden (vgl. vor allem Guilford 1954, 278-296; Weick 1968; v. Cranach/Frenz 1969; Schulz/ Teschner/Voigt 197o; Hasemann 1971; Merkens 1972; Langer/Schulz v. Thun 1974, 16-26; Fassnacht 1979, 52-69, 15o-151). Verfahren zum Versuch ihrer Beseitigung oder Kontrolle werden ebenfalls angegeben; sie beziehen sich im wesentlichen auf das Beobachtertraining und die Konstruktion des Beobachtungsinstruments (vgl. z.B. Guilford 1954; Langer/Schulz v. Thun 1974), seltener auf das Vorgehen bei der Datenanalyse (vgl. z.B. Guilford 1954; Light 1973).

Die Befolgung dieser Verbesserungsvorschlaege ist allerdings mit einem erheblichen Aufwand verbunden (vgl. v. Cranach/Frenz 1969, 323; Walter 1977, 92; Fassnacht 1979, 151), so dass der Anwender von Schaetzskalen bereits unter diesem Gesichtspunkt in ein Dilemma geraet: einerseits wird an ihnen naemlich Handlichkeit, zeit- und kostenmaessige Sparsamkeit geruehmt (vgl. z.B. Guilford 1954, 263; Bachmair 1977, 45; Kerlinger 1979, 8o2), die sie als Forschungsinstrumente empfehlen; andererseits liegt auf der Hand, dass mit der Nutzung dieser "Vorteile" eine unter bestimmten Umstaenden unzureichende Datenqualitaet verbunden ist (vgl. z.B. Jahoda/Deutsch/Cook 1951, 2o7-2o8; Biddle 1964, 27; Bachmair 1977, 45; Kerlinger 1979, 8oo); m.a.W.: wer bei der Verwendung von Schaetzskalen zur Erforschung von Unterricht von den genannten "Vorteilen" Gebrauch macht, verzichtet darauf, "Unterrichtsforschung" zu betreiben.

Dieses Dilemma ist freilich praktischer, nicht prinzipieller Natur. So scheint es auch nicht verwunderlich, dass die meisten Autoren, die sich unter methodischem Aspekt mit Schaetzskalen fuer die Unterrichtsforschung beschaeftigen, doch zu einer guenstigeren Beurteilung gelangen, die auf dem Glauben an die Vermeidbarkeit der bekannten Schwaechen beruht (vgl. z.B. Guilford 1954, 297; Friedrichs/Luedtke 1971, 17; Muench 1971, 37; Fenner 1973, 132; Langer/Schulz v. Thun 1974, 27-28; Ruprecht 1974, 1o2; Walter 1977, 92-93; Sumaski 1977, 77; Kerlinger 1979, 8o2).

Unter Forschungsgesichtspunkten liegt der Hauptvorteil der Schaetzskalen nach einhelliger Auffassung der genannten Autoren darin, dass mit diesem Verfahren die Luecke uebersprungen werden koenne, die zwischen der "paedagogisch-trivialen" Auszaehlung quantifizierbarer isolierter Merkmale und der theoriehaltigen Erfassung komplexer Sachverhalte klafft. Tatsaechlich gelingt - formal gesehen - die Loesung dieses Problems durch die Verwendung des Messinstruments "Mensch", dessen "Konstruktion" eine differenzierte "Anzeige" hochkomplexer

Unterrichtsvariablen ermoeglichen soll. Als "Anzeigemedium" dient die (Theorie-) Sprache, deren Begriffe als Konstruktionselemente fuer die Mess-Skalen verwendet werden.

Der dieser Vorgehensweise zugrundeliegende Gedanke besteht darin, dass die (i.d.R. faktorenanalytische) Auswertung so gewonnener Informationen Aufschluss darueber gibt, in welchen charakteristischen Zuegen (Faktoren) sich das in der Realitaet gezeigte Lehrverhalten im einzelnen aeussert (Faktorengewichte) und variiert (Faktorenscores), und in welcher Weise es z.B. mit Lerneffekten zusammenhaengt (korreliert). Voraussetzung dafuer ist, dass die in die Analyse eingehenden Messdaten sich auf verschiedene (abhaengige oder unabhaengige!) Elemente der Realitaet beziehen, da anderenfalls Uebereinstimmungen (Interkorrelationen) zwischen den Messergebnissen nicht auf real unterscheidbare Ereignisse, sondern - aufgrund der Uebereinstimmung in den Messinstrumenten (hier: Skalen) - auf die Identitaet des Messobjekts zurueckzufuehren sind und insofern weniger ueber die Realitaet als ueber den Sprachgebrauch aussagen. So ist es beispielsweise sicherlich plausibel, dass die Einschaetzung eines Gegenstandes unter den Gesichtspunkten "Umfang", "Ausmass", "Volumen", "Ausdehnung", "Maechtigkeit" etc. wahrscheinlich zu einer hohen Interkorrelation der Daten fuehren muesste. Dieser Effekt wuerde allerdings weder darauf beruhen, dass jene Merkmale dem beobachteten Gegenstand zugleich zukaemen (dass er also hinsichtlich der Art ihrer Kombination prinzipiell auch anders "aussehen" koennte), noch darauf, dass die Mess-Subjekte ueber eine eingeschraenkte Wahrnehmungsfaehigkeit verfuegten (vgl. dazu z.B. Hofer (1969) und die in Abschnitt 2.2.4. angegebene Literatur). Vielmehr liesse sich wohl mit Recht vermuten, dass durch die sprachliche Fassung der Messgesichtspunkte die Aufmerksamkeit der Mess-Subjekte in Wirklichkeit stets weitgehend auf denselben Realitaetsausschnitt gerichtet wurde.<2>

Das hier angedeutete Problem ist nicht allein fuer die Konstruktion von mehrdimensionalen Rating-Instrumenten bedeutsam, sondern es betrifft die Grundlagen von kategoriengeleiteter Unterrichtsbeobachtung ueberhaupt. Fuer den Fall, dass mit ihrer Anwendung zugleich die oben erwaehnte Schwierigkeit der Gewinnung theoretisch relevanter Messdaten geloest werden soll, stellt sich darueber hinaus die Frage nach der Begrifflichkeit der jeweils zu erstellenden oder zu pruefenden (Unterrichts-)Theorie. In diesem Punkt wird die Unterscheidung gegenueber anderen Anwendungen von Schaetzskalen deutlich, mit denen beispielsweise in der psychologi-

schen Diagnostik (vgl. Mittenecker 1971) oder in der sozial-psychologischen Feldforschung (vgl. z.B. Schmidtke/Groffmann/Schaller 1978) Einstellungen erhoben oder Sprachverstaendnis (vgl. Hofstaetter 1973, 258ff) untersucht werden soll. Waehrend dort mit den Skalen gerade an die individuellen Gegebenheiten der Probanden angeknuepft wird, fungieren sie in der Unterrichtsforschung als Instrumente wissenschaftlicher Beschreibung und sind also Teil der "normierten" Forschungssprache. Nur mit dem Skalenproblem der wissenschaftlichen Beschreibung beschaeftigen wir uns im folgenden. Soweit wir sehen, ist seiner rein sprachlichen Seite in der Methodenliteratur zum Ratingproblem bisher wenig Beachtung geschenkt worden (einige Hinweise finden sich bei Friedrichs/Luedtke 1971, 41, 69; Graumann 1974, 35; Rosenshine 1974, 2o3; Six 1975, 28o-281; Mueller-Wolf 1977, 1o4-1o5; Fassnacht 1979, 25, 15o).

Im folgenden Abschnitt soll zunaechst eine genauere Analyse dieses Bedeutungsfehlers, wie wir ihn nennen wollen, vorgenommen und seine Beziehungen zu den weiteren in der Literatur diskutierten Schaetzfehlern herausgestellt werden; ausserdem ist es erforderlich, seinen methodischen und theoretischen Status zu charakterisieren. Danach werden anhand der Ergebnisse einer empirischen Studie zur Beschreibung von Lehrverhalten die Effekte des Bedeutungsfehlers qualitativ und quantitativ untersucht. In einem weiteren Abschnitt eroertern wir unter Bezugnahme auf vorliegende Untersuchungen die Konsequenzen, die sich fuer die dort gefundenen Aussagen bei Beruecksichtigung des Sprachproblems ergeben koennten. Abschliessend wird auf die Moeglichkeiten der Vermeidung des Bedeutungsfehlers eingegangen.

## 2. Das Zuordnungsverhaeltnis von Sprache und Realitaet als Problem der Unterrichtsforschung

### 2.1. Die Rolle des Bedeutungsumfangs von Termini in der Datenerhebung

#### 2.1.1. Die Relation von Wort und Bedeutung als Problem der Wissenschaftssprache

In allen bekannten Sprach- und Sprechtheorien wird - mit unterschiedlicher Terminologie und in verschiedenen konzeptuellen Zugriffen - eine Differenzierung zwischen Wort<3> und Bedeutung vorgenommen. Die Relation zwischen beiden erhaelt dabei unter den verschiedensten Fragestellungen erhebliches Gewicht; allein schon die Darstellung der spezifischen Problemlagen, der philosophischen, psychologischen, biologischen, philologischen, informationstheoretischen, ethnologischen Sichtweisen - um nur die gelaeufigsten zu nennen - , wuerde umfangreiche Ausfuehrungen erforderlich machen (vgl. z.B. die Hinweise bei Miller/Johnson-Laird 1976, 1-8), ganz zu schweigen von den dort vorgelegten Bearbeitungs- und Loesungsversuchen.

Im "einfachen" Falle meint "Bedeutung" irgendeinen realen Gegenstand (Objekt) oder das Merkmal (die Eigenschaft) eines Gegenstandes (Referenzbereich), worauf wir mit dem "Wort" (Praedikator)<4> Bezug nehmen koennen. Dagegen liegen in der Unterrichtsforschung die Verhaeltnisse meist etwas komplizierter. Sieht man naemlich von den streng am sichtbaren Verhalten orientierten Ansaetzen ab (vgl. z.B. Brannigan/ Humphries 1972), zu deren empirischer Fundierung haeufig Datenerfassungsgeraete eingesetzt werden, so richtet sich hier das Interesse hauptsaechlich auf "komplexere" Gegebenheiten, die mit Hilfe von Schaetzskalen erhoben werden sollen. Formulierungen wie z.B. "energisch", "selbstsicher" (Mueller-Wolf 1977, 197) werden den Beobachtern als Aspekte vorgegeben - Woerter also, deren Referenzbereich nicht unmittelbar ueber die fuenf Sinne erreicht werden kann.

Bei unserer Analyse versuchen wir, hinsichtlich der Behandlung des Relationsproblems so formal wie moeglich in der Hinsicht zu verfahren, dass unterschiedliche Deutungen im Kontext von philosophischen und einzelwissenschaftlichen Positionen, wie sie oben genannt sind, nicht praejudiziert werden. So sollen beispielsweise Ueberlegungen zum Problem der Vermittlung zwischen den beiden Endstellen der Relation, zwischen dem Wort und seinem Referenzbereich, unberuehrt bleiben, also weder die philosophische Bewusstseins- und die Leib-Seele-Pro-

blematik noch psychologische oder neuro-physiologische Wahrnehmungstheorien einbezogen werden. Auch sehen wir von der Frage nach der Semantik-Konzeption ab, in der psycholinguistische (vgl. z.B. Engelkamp 1974, 87ff.) und biologische (vgl. z.B. Lenneberg 1977, 403 ff.) Positionen unvereinbar nebeneinander stehen.

Voraussetzungslos laesst sich freilich nicht argumentieren. Wir gehen von der prinzipellen Unterscheidbarkeit von Wort und Bedeutung aus und unterstellen, dass die Relation zwischen beiden - unbeschadet entwicklungspsychologischer Befunde (vgl. z.B. Oerter 1973, 511 ff.) - grundsaetzlich der konventionellen Festlegung verfuegbar ist. Wuerde diese Bedingung fallengelassen, so geriete man nicht allein im Hinblick auf die Erklaerung des Funktionierens verbaler Kommunikation in eine Reihe von Schwierigkeiten; auch die auf der Idee intersubjektiver Pruefung entworfenen Wissenschaftsprogramme, in deren Geltungsbereich der hier vorliegende Text selbst einzuordnen ist, waeren von vornherein zum Scheitern verurteilt.

Unter Bezugnahme auf Wittgenstein hat Stegmueller (1977, 288 ff.) darauf hingewiesen, dass die "Darstellungsfunktion" der Sprache in der Wissenschaft von anderen Aufgabenstellungen scharf zu trennen sei. Die vorliegende Arbeit beschaeftigt sich mit Sprachproblemen von (Erziehungs-)Wissenschaft, naeherhin mit der Frage nach Praezisierungsmoeglichkeiten in ihrem deskriptiven Bereich. Untersuchungen zur Umgangssprache sind dabei zwar insofern bedeutsam, als sie wissenschaftlich relevante Hinweise auf kontroll- und explikationsbeduerftige immanente Regeln des alltagspraktischen "Sprachspiels" (Wittgenstein) geben koennen. In ihrem Analysecharakter unterscheiden sie sich aber von den Intentionen wissenschaftsmethodischer Ueberlegungen, die wesentlich kritisch-konstruktiver Art sind. So steht in der hier behandelten Version der Frage nach der Relation zwischen Wort und Bedeutung das Konstruktionsproblem einer (moeglichst exakten) Beschreibung (von Lehrverhalten) im Mittelpunkt; damit weichen unsere Ueberlegungen in systematischer Weise von Analysen des alltagssprachlichen situationsverhafteten "Benennens" ab (vgl. Herrmann/Deutsch 1976, 14-15), deren Ergebnisse jedoch - wie angedeutet - z.B. unter der Frage nach "Verboten" fuer die wissenschaftliche Beschreibung ausgewertet werden koennen.

## 2,1,2, Die Dimensionalitaet der Bedeutungsrelation

Die in dieser Untersuchung eroerterte Form der Datenerhebung erfolgt im Prinzip in der Weise, dass dem Beobachter eine Reihe von Skalen vorgelegt wird, auf denen er seine Urteile ueber Lehrer, Schueler, Atmosphaere usw, quantifizieren soll, Fuer Angaben zur Lehrperson - wir konkretisieren unsere Ueberlegungen an diesem Beispiel - findet man etwa folgende Vorgaben (vgl, Mueller-Wolf/Fittkau 1971, 167):

| | | |
|---|---|---|
| unfreundlich, kuehl | \|---\|---\|---0---\|---\|---\| | freundlich, warmherzig |
| der Dozent respektiert die Studenten als Partner und Persoenlichkeiten | \|---\|---\|---0---\|---\|---\| | er sieht auf die Studenten herab |

2-1 Schaetzskalen aus der Unterrichtsforschung

Durch Ankreuzen auf der Skala gibt der Beobachter seinen Eindruck wieder,<5> Sieht man von der Frage ab, ob mit so gewonnenen Daten Aussagen ueber beobachtetes Verhalten oder erschlossene Verhaltensdispositionen gemacht werden (vgl, dazu Abschn, 2,3,3,), und laesst man ausserdem die uebrigen "Schaetzfehler" ausser Betracht (vgl, Tab, 1, Ziff, 2-2o), so ist zunaechst zu klaeren, worauf sich die Aufmerksamkeit des Beobachters richtet bzw, welche Merkmale er reflektiert, wenn er sich ein Urteil bildet, Der Gesamteindruck "freundlich - unfreundlich" erwaechst - i,d,R, unbewusst - im Laufe der Beobachtung (a) aus der Beachtung bestimmter - wahrscheinlich aber subjektiv nicht differenzierbarer - Aspekte <6>, (b) aus ihrer Gewichtung gemaess dem subjektiven Bedeutungsmuster des vorgegebenen Wortes <7> und (c) aus ihrem - wahrgenommenen - Auspraegungsgrad (Intensitaet bzw, Haeufigkeit) <8>,

Der Bedeutungsbereich eines Wortes dieser Komplexitaet <9> laesst sich - formal gesprochen - als dreidimensionaler Raum mit orthogonalen Koordinatenachsen (unterschiedlichen Skalenniveaus; vgl, Anm, 6-8) rekonstruieren,<1o> Allgemein bestimmt sich die Relation zwischen Woertern (W) und Bedeutung (B) demzufolge als Funktion von {W} in {B},<11> Die graphische Darstellung einer Wortbedeutung ergibt ein Profil <12> (Abb, 2-2; Bedeu-

tungsprofil eines Wortes).

Eine Ursache fuer den Praezisionsunterschied zwischen Umgangssprache und exakter Wissenschaftssprache besteht demnach darin, dass auf der Ebene der Wortsemantik die Relationen zwischen Wort und Bedeutungsraum variieren koennen. Die Vagheit von Praedikatoren der Umgangssprache ist als interpersonell und gegebenenfalls intrapersonell verschiedene Fixierung auf den einzelnen Dimensionen rekonstruierbar; in Abb. 2-2 wuerde sie sich durch unterschiedliche Profillinien fuer ein und dasselbe Wort ausdruecken.

Unter einem messpraktischen Aspekt ist die Dimensionalitaet der Wort-Bedeutung-Relation als das Aequivalent zum Konstruktionsprinzip eines Anzeigegeraetes zu deuten. Waehrend solche Vorrichtungen i.d.R. nur fuer eine oder wenige Veraenderliche sensibel sind (z.B. Thermometer, elektrische Mehrfachmessgeraete), eroeffnet der Einsatz eines menschlichen Beobachters als "Datenanzeige" eine Vielzahl von Messdimensionen. Dabei fungiert die Vorgabe eines Wortes (z.B. im Erhebungsbogen) als "Einstellung" des Aspekts (des Sensibilitaetsbereiches), die Abgabe der Beschreibung bzw. des Urteils (z.B. Ankreuzen auf einer Skala) als Messwert ueber die Auspraegung des Merkmals (vgl. auch Merkens 1974, 15-18). Dazwischen liegt der Messvorgang, in dem - nach unterschiedlichen theoretischen Vorstellungen (vgl. 2.1.1.) - Sinnesreize gemaess der dreidimensionalen Bedeutungsstruktur auf die Verbaldimension "abgebildet" werden (vgl. auch Fassnacht 1979, 37-49). Die besonderen Probleme solcher "Messungen" bestehen darin, dass das Zuordnungsverhaeltnis von Wort und Bedeutung, also das "Konstruktionsprinzip" des "Messgeraets", sowohl von Individuum zu Individuum (von "Geraet" zu "Geraet") als auch innerhalb des Individuums - von Situation zu Situation - variieren kann.

### 2.1.3. Interpersonelle Differenzen in der Sprachanwendung

Auf der Suche nach Erklaerungen fuer Unterschiede in den Beurteilungen verschiedener Unterrichtsbeobachter (P) in der gleichen Skala wird man, neben den uebrigen, in der Literatur diskutierten Fehlern (vgl. Tab. 1, Ziff. 2-20) auch die individuelle Dimensionalitaet der Wortbedeutung in Betracht ziehen muessen. So koennen Differenzen daraus erwachsen, dass P(1) und P(2) sich in der Intensitaets-/Haeufigkeitskonzeption eines Wortes unterscheiden, d.h. dass sie fuer die Vergabe desselben Urteils (z.B. Ankreuzen der Extremstelle "unfreundlich") unter-

schiedliche Wahrnehmungen ueber die Intensitaet der zu beachtenden Merkmale erwarten bzw, dass sie bei gleicher Intensitaetswahrnehmung (und sonst gleichen Bedingungen) unterschiedliche Ankreuzungen vornehmen wuerden. Dieser Effekt, das gilt es zu beachten, ist voellig unabhaengig von wahrnehmungspsychologisch beschreibbaren individuellen Unterschieden, die sich ueber Differenzen in der (subjektiven) Bedeutsamkeit der relevanten Merkmale (vgl, "Kontrastfehler", Tab,1, Ziff,14) oder in der Vigilanz (vgl, v,Cranach/Frenz 1969, 276-278) auswirken.

Fuer die Gewichtung der mit ihren Intensitaeten beobachteten Merkmale, die zu einer Wortbedeutung integriert werden, gelten analoge Ueberlegungen <13>: P(1) und P(2) koennen die Anwendung eines Wortes auf eine Menge von Wahrnehmungen (relevante Merkmale samt Intensitaeten) davon abhaengig machen, ob "wichtige" Wahrnehmungselemente eine bestimmte Ausspraegung aufweisen, P(1) mag beispielsweise fuer die Vergabe der Bezeichnung "(extrem) unfreundlich" voraussetzen, dass vor allem eine erhoehte Lautstaerke vorliegt, waehrend P(2) diesem Umstand bei der gleichen Urteilsabgabe weniger Gewicht verleiht, es dafuer aber als wesentlich erachtet, dass bestimmte mimische Merkmale in einem gewissen Ausmass gezeigt werden.

Was schliesslich die einzelnen Merkmale betrifft, so ist insofern mit Unterschieden zu rechnen, als P(1) und P(2) ein und dasselbe Wort unter Beruecksichtigung unterschiedlicher Gesichtspunkte (im Extremfall ohne Gemeinsamkeiten) verwenden koennen (vgl, auch die Darstellung bei Friedrichs/Luedtke 1973, 37) (Abb, 2-3; Verhaeltnis der Merkmalsmengen bei verschiedener Bedeutung desselben Wortes durch zwei Personen).

Fuer die Diagnose interpersoneller Unterschiede zwischen Beobachterdaten wird in der Literatur die Berechnung eines Uebereinstimmungs- oder Aequivalenzkoeffizienten vorgeschlagen (vgl, z,B, v,Cranach/Frenz 1969, 300 ff,; Medley/Mitzel 1971, Sp, 659 ff,; Rosenshine/Furst 1973, 168 f,),<14> Dabei weisen die Autoren stets darauf hin, dass hohe Uebereinstimmungsmasse nicht allein schon die Fehlerfreiheit der Beobachtungen signalisieren. Umgekehrt stellt sich jedoch die Frage, ob durch solche hohen Werte bereits die "Objektivitaet einer Beobachtungstechnik" erwiesen ist (Medley/Mitzel 1971, Sp, 661). Mit Recht hebt Weick (1968, 404) im Anschluss an Dunnette hervor, dass fuer jede einzelne Fehlerquelle bei der Datengewinnung spezifische Reliabilitaetsueberlegungen anzustellen sind.

Unsere Erorterungen zur Wortanwendung legen es nahe,

diesen Hinweis auf die drei bedeutungsrelevanten Dimensionen anzuwenden. Man kann sich leicht klarmachen, dass eine Vernachlaessigung dieses Gesichtspunktes trotz hoher Aequivalenzkoeffizienten zu falschen Forschungsresultaten fuehren kann, wenn man bedenkt, dass durch die Beobachtung unterschiedlicher Merkmalsmengen (vgl. Abb. 2-3) "echte" Korrelationen (Kovariationen) zwischen Merkmalen bzw. Merkmalsgruppen unentdeckt bleiben, also mit Sprache "zugedeckt" werden koennen.

Nach den vorliegenden Untersuchungen zu dieser Frage der interindividuellen Differenzen im Sprachgebrauch ist damit zu rechnen, dass aus dieser Quelle nicht zu vernachlaessigende Fehlervarianzen fliessen. So stellt Natsoulas (1968) die vielfaeltigen Probleme dar, die sich bei der Interpretation von Beobachtungsdaten fuer die Konzeption des Zusammenhangs von Wort und Bedeutung in verschiedenen philosophischen Sichtweisen ergeben. In einer Untersuchung zum amerikanischen Wortschatz ermittelt und beschreibt Britton (1978) die Vieldeutigkeit von Woertern, die z.T. auch in Beobachtungsinstrumenten auftauchen. Schliesslich gibt Scherer (1974) einen Ueberblick ueber empirische Studien zu Fehlerquellen fuer Beobachtungsdaten ("observer bias"), die auf die unterschiedliche Konzeption der Wort-Bedeutung-Relation bei verschiedenen Beurteilern hinweisen. Es zeigt sich beispielsweise, dass Geschlecht (Exline/Winters; Kleck/Nuessle) und Extraversion (v.Cranach; Krueger/Hueckstedt) die Sensibilitaet der Beobachter fuer die Wahrnehmung von Blickkontakt beeinflussen (99-100); klinische Psychologen ermitteln Emotional-Variablen eher ueber die Beobachtung des Gesichtsausdrucks, Taenzer dagegen durch Deutung von Koerperbewegungen (100); Erfahrung und Spezialisierung von Beobachtern (Shapiro) beeinflussen ihre Sensitivitaet fuer Gesichtsausdruck und Sprachverhalten (100). Der Frage, ob solche Unterschiede durch Training beseitigt werden koennen und vor allem auch, ob und welche Unterschiede durch das Training erst freigelegt (aktiviert) werden (vgl. Stanz 1974, 34), soll und kann hier nicht nachgegangen werden. Es ist jedoch offensichtlich, dass Zuverlaessigkeit, Objektivitaet und Validitaetsaspekte von Beobachtungsdaten aufs engste mit ihr verknuepft sind.

### 2.1.4. Intrapersonelle Differenzen in der Sprachanwendung

Ein Mass fuer die Zuverlaessigkeit von Beobachtern ist mit der Konstruktion des Stabilitaetskoeffizient en entwickelt worden (vgl. Guilford 1954, 395 ff.; Med-

ley/Mitzel 1971 Sp. 799 ff.). Er gibt Information darueber, bis zu welchem Grade die einzelnen Angaben einer Beobachterperson voneinander abweichen, wenn zu verschiedenen Zeitpunkten der gleiche Sachverhalt (z.B. Film-/Tonband-/Textinhalte) beurteilt werden soll (vgl. auch v.Cranach/Frenz 1969, 300-301). Strukturell stimmt die Problematik der Ermittlung des Stabilitaetskoeffizienten mit der soeben eroerterten des Aequivalenzkoeffizienten ueberein. Insoweit kann auf die Darstellung der moeglichen Wortanwendungsfehler hier verzichtet werden.

Auf die Instabilitaet der intrapersonellen Wort-Bedeutung-Relation weisen neben entwicklungspsychologischen Befunden (vgl. z.B. Oerter 1973) auch speziell zum Beobachterproblem vorgelegte Untersuchungen hin. Drei Arten von Veraenderungsursachen koennen unterschieden werden: (a) situative Einfluesse (vgl. Webb 1975, 41), (b) die Auswirkungen der Beschaeftigung mit den Beobachtungsinhalten, insbesondere der Identifizierung mit den Problemen der zu beobachtenden Personen, der "going-native-Effekt" (Friedrichs/Luedtke 1973, 189; Gruemer 1974, 64-65) und (c) Einfluesse, die langfristig den Sprachgebrauch von Personen modifizieren (Webb 1975, 177-178).

Diese Schwankungen werden von einem weiteren Effekt ueberlagert, der sich ebenfalls als Sprachgebrauchswechsel manifestiert. Er tritt dann ein, wenn ein Wort, das den Beobachtungsaspekt angibt, beim Beobachter nicht eindeutig im Bedeutungsraum fixiert ist, wenn er also keine klare Vorstellung von dessen Bedeutung hat. Die jeweils aktualisierte Relation schwankt in solchen Faellen zufallsabhaengig, wenn sich der Beobachter in einer forced-choice-Situation befindet (vgl. Tent 197, Sp. 886 ff.). Neben Weick (1968, 406) und Friedrichs/Luedtke (1973, 40) hat vor allem Feger auf dieses Problem hingewiesen (1972, 29, 31, 217).

Schliesslich ist noch zu erwaehnen, dass sich aus den psychologischen Befunden zur Dimensionalitaet der Personwahrnehmung Konsequenzen fuer den Spielraum bei der "Herstellung" der Wort-Bedeutung-Relation ergeben koennen. Wenn es zutrifft, dass wir andere Menschen letztlich nur auf drei bis fuenf Eigenschaftskontinua diskriminieren koennen (vgl. z.B. Hofer 1969, Nickel 1976, Belschner/Spaeth 1977, Schmidtke/Groffmann/Schaller 1978), so waere dies ein Hinweis darauf, dass der Bedeutungsraum, in dem Beobachtungswoerter lokalisiert werden, begrenzt ist. Damit stiesse man auf ein Problem, mit dem wir uns im naechsten Abschnitt aus anderen Ueberlegungen heraus beschaeftigen wollen. Es ist naem-

lich denkbar, dass fuer einen umgrenzten Bedeutungsbereich mehr Woerter vorgegeben werden, als "ueberschneidungsfreie" Profilverlaeufe in ihm konstruierbar sind. Die dadurch entstehenden Interferenzen koennen Beobachtungsdaten in erheblichem Masse beeinflussen.

### 2.1.5. Der Bedeutungsfehler

Bereits aus unseren bisherigen Ueberlegungen geht hervor, dass die Hoffnung vergeblich sein muss, mittels der Berechnung eines aus Stabilitaet und Aequivalenz kombinierten Reliabilitaetskoeffizienten etwas ueber die Zuverlaessigkeit des Messinstruments zu erfahren.<15> Dies waere - in entsprechend modifizierter Form - erst moeglich, wenn die Dimensionen der Wort-Bedeutung-Relation inter- und intrapersonell isoliert und kontrolliert werden koennten. Mit Fassnacht (1979, 22) ist ausserdem auf den moeglichen Kompensationseffekt zwischen den Fehlerquellen hinzuweisen.

Eine zusaetzliche Problemdimension wird sichtbar, wenn wir uns jetzt dem (ueblichen) Fall zuwenden, dass die Datenerfassung sich nicht auf einen einzigen Aspekt beschraenkt, sondern dass der (oder die) Beobachter aufgefordert ist (sind), eine Ereignisabfolge, also etwa das Lehrerverhalten, unter mehreren Gesichtspunkten zu betrachten. Im Prinzip handelt es sich dabei um eine "einfache" und zugleich plausible Aufgabenstellung, deren Struktur man sich leicht am Beispiel eines Tisches verdeutlichen kann, der nach den Merkmalen Farbe, Form, (Ober-)Flaeche, Volumen, Material usf. untersucht werden soll. Jedesmal geht es um denselben Gegenstand (Ereignisabfolge, Lehrerverhalten), der unter verschiedenen Aspekten (Kategorien, Dimensionen) thematisiert wird, beispielsweise, um zu pruefen, ob zwischen den Merkmalen bei verschiedenen Gegenstaenden dieser Art Zusammenhaenge bestehen.<16>

Das Problem bei der Unterrichtsforschung besteht nun darin, dass die ueber bestimmte Woerter vermittelte Vorgabe der verschiedenen Betrachtungsaspekte ihre Funktion erst dann erfuellt, wenn zugleich gewaehrleistet ist, dass damit auch faktisch unterschiedliche Merkmale erfasst werden. Falls die einzelnen Woerter (z.B. Skalenbezeichnungen), mit denen die Aufmerksamkeit des Beobachters -, wie der Lichtstrahl eines Scheinwerfers - auf verschiedene Bereiche des Objekts gelenkt werden soll, nicht so differenzieren, dass wirklich stets neue, noch nicht beachtete Regionen in den Blick genommen werden - falls also die Lichtkegel sich ueberschneiden -, so er-

haelt man bei der "Datenausgabe" (Ankreuzen auf der Skala) mehr oder weniger redundante Informationen ueber dieselben Realitaetsbereiche.

Sieht man einmal von den Dimensionen 'Gewichtung' und 'Intensitaet' ab, so laesst sich dieser Sachverhalt beschreiben als die (teilweise) Ueberschneidung der Merkmalsmengen, ueber denen zwei verschiedene Woerter ihre Bedeutung integrieren (Abb. 2-4; Die Bedeutungsgemeinsamkeit zweier Woerter (W)).

In analoger Weise koennen die Beziehungen zwischen zwei Woertern mit identischem Merkmalsbestand, aber unterschiedlicher Gewichts- bzw. Intensitaetsverteilung rekonstruiert werden; der Extremfall besteht in der totalen Uebereinstimmung der Bedeutungsprofile zweier verschiedener Woerter.

Es ist an dieser Stelle wichtig, sich klar zu machen, dass solche Bedeutungsgemeinsamkeiten ausschliesslich auf der Basis der Merkmalsueberschneidung erwachsen, dass also gleichartige Profilverlaeufe in der Gewichts- und Haeufigkeitsdimension selbstverstaendlich nur dann Bedeutungsueberschneidungen der zugehoerigen Woerter quantitativ beeinflussen, wenn dieses Phaenomen ueber teilweise oder vollstaendig gemeinsamer Merkmalsgrundlage auftritt. Wir werden deshalb auch im weiteren unsere Argumente hauptsaechlich auf den Merkmalskomplex richten.

Der Fehler, der demnach in vielen z.T. aufwendigen Untersuchungen (vgl. z.B. Mueller-Wolf 1977) im Bereich der Unterrichtsforschung auftritt, besteht darin, dass Daten zur Grundlage von Analysen gemacht werden, denen faelschlicherweise unterstellt wird, sie enthielten Informationen ueber die Auspraegung von - im strengen Sinne - verschiedenen Merkmalen bzw. Eigenschaften der Lehrpersonen (oder des Unterrichtsverlaufs usw.). Tatsaechlich gibt es aber einen Zusammenhang zwischen den so ermittelten Informationen, der auf ihrer (teilweisen) Redundanz beruht. Der Umfang der auf diese Weise gewonnenen echten Information haengt ausschliesslich vom individuellen Sprachgebrauch des Beobachters ab, d.h. also von den Relationsstrukturen, die bei ihm zur Zeit der Datenerhebung zwischen den Wortvorgaben und ihren Bedeutungsentsprechungen etabliert sind. Diese koennen interpersonell unterschiedlich aufgebaut sein und von Situation zu Situation variieren (vgl. Abb. 2-3).

Nach der Einteilung der Fehlerarten von Singleton (1972) rechnet der Bedeutungsfehler zu den "input errors". Er entsteht bei der "Eingabe" der Merkmalsauspraegungen

(und ihrer Variationen) in das Messgeraet. Dabei kann die "Eingabe" ueber die sensorische Peripherie oder durch "Abrufen" von bereits frueher abgespeicherten Wahrnehmungen erfolgen. Der letztere Modus bildet den Standardfall der Ratingmethode, bei der ja Aussagen ueber Ereignismengen (z.B. das "Gesamtverhalten" der Lehrperson) gefordert werden. Aus diesem Grunde ist auch mit dem Auftreten von "Interventions-" und "Zeitraumfehlern" (vgl. Tab. 1, Ziff. 3, 4) zu rechnen. Der Zeitraumfehler entsteht, wenn sich verschiedene Beobachter oder Untersuchungsleiter und Beobachter im (latenten) Dissens hinsichtlich der einzubeziehenden Ereignismenge befinden. Ein Interventionsfehler ereignet sich, wenn zwischen "Aufnahme der Daten" und "Abgabe der Messwerte" Veraenderungen der "Daten" im "Speicher" (Gedaechtnis) erfolgen, die z.B. auf dazwischenliegende ueberlagernde Erlebnisse oder Vergessenseffekte zurueckzufuehren sind.

Auch der "Hintergrundeffekt" (vgl. Tab. 1, Ziff. 2) kann unabhaengig vom Bedeutungsfehler definiert werden; er bezieht sich auf Interferenzen zwischen wortzugehoerigen und nicht zugehoerigen Merkmalen aus dem Beobachtungsumfeld. Durch ihn werden also die Auspraegungen der gemaess Wortbedeutung zu erfassenden Aspekte modifiziert<17> (Abb. 2-5; Bedeutungsfehler und Hintergrundeffekt).

Sind die "Daten" einmal aufgenommen, so koennen alle uebrigen bekannten Verfaelschungseffekte bei der Verarbeitung/Entscheidung und beim "output" (Singleton 1972) zusaetzlich auftreten (vgl. Tab. 1, Ziff. 5-17; Ziff. 18-20, jedoch ohne 1-4).<18>

## 2,2, Methodische und methodologishe Dimensionen von sprachlichen Inkonsistenzen

### 2,2,1, Beobachtertraining und Bedeutungsfehler

Haeufig wird in der Literatur darauf hingewiesen, dass sorgfaeltige Beobachterschulung eine weitgehende Minimierung der Fehlerquellen erlaube (vgl, z,B, Guilford 1954, 280; v, Cranach/Frenz 1969, 305; Glueck 1971, 60; Hasemann 1971, 833; Friedrichs/Luedtke 1973, 172-173; Dechmann 1978, 90-91), Langer und Schulz v, Thun sind der Auffassung, man koenne u,a, durch "Vorzeigen" von "praegnanten Beispielen" (1974, 25) im Rahmen des Rater-trainings das Sprachproblem loesen, Tatsaechlich verfehlen aber alle derartigen Ueberlegungen das oben charakterisierte Bedeutungsproblem, Selbst wenn man naemlich einmal von der prinzipiellen Schwierigkeit des kontrollierten Sprachgebrauchsvergleichs zwischen Personen absieht (vgl, dazu Fassnacht 1979, 150), sind die in diesem Zusammenhang gemachten methodischen Vorschlaege in ihrer Leistungsmoeglichkeit auf zwei andere Fehlerquellen beschraenkt; im guenstigsten Fall kann mit ihnen inter- und intrapersonelle Reliabilitaet hergestellt werden (vgl, 2,1,3,/2,1,4,), Davon unberuehrt bleibt jedoch die "Bedeutungsueberlappung" der Beobachtungswoerter (bzw, -kategorien), die dann fuer alle Beobachter in gleicher Weise gelten wuerden,

### 2,2,2, Die Problematik des Skalen-Rater-Mittelwerts

Auch der Gedanke, dass die in der Person des einzelnen Beobachters wirksamen Fehlerquellen dadurch ausgemerzt wuerden, dass man einen Mittelwert aus den Angaben der Beobachter bildet (vgl, Phillips 1970, 101-103; Fenner 1973, 127), fuehrt keineswegs aus dem Bedeutungsproblem heraus, Man muesste zunaechst dabei unterstellen, dass die Schaetzfehler multidimensional (!) einer Zufallsverteilung folgen wuerden und dass der Beobachtermittelwert damit eine gute Schaetzung des wahren Wertes des Beobachtungsmerkmals darstellen wuerde (vgl, z,B, Smits 1978), Diese Annahme kann zwar prinzipiell zutreffen - auch systematische Fehler koennen im mehrdimensionalen Raum zufaellig verteilt sein - , aber sie muesste fuer diesen Fall empirisch fundiert sein, Mit dieser Konstruktion waere zugleich ein Modell unterlegt, das fuer die Datenerhebung den Einsatz einer repraesentativen Beobachterstichprobe erfordern wuerde, Das setzte wie-

derum die (theoretische) Kenntnis der zugehoerigen Population voraus, deren Bestimmung gerade wegen des Einflusses der Beobachterschulung, also der "systematischen" Beeinflussung des systematischen Fehlers, erhebliche Probleme mit sich bringen muesste. Unabhaengig davon waere noch die Frage nach der Beruecksichtigung des Zufallsfehlers (im "klassischen" Sinne) beim Messvollzug zu beantworten, die theoretisch und experimentell mit dem Konzept der Messwiederholung angegangen wird.

## 2.2.3. Die Reliabilitaetsfrage bei Ratingdaten

Wir halten die einer so motivierten Mittelwertbildung zugrundezulegenden Annahmen - jenseits ihrer experimentellen Pruefung - nicht fuer plausibel und ihre Implikationen fuer problematisch. Abgesehen davon erhebt sich jedoch umgekehrt die Frage, mit welcher Begruendung in der - uns bekannten - Literatur durchweg vom Einsatz mehreren Beobachter ausgegangen wird (vgl. die in diesem Abschnitt eingangs angegebenen Quellen und z.B. Medley/Mitzel und Schulz/Teschner/Voigt 1971). Wuerde der Aequivalenz- oder Uebereinstimmungskoeffizient tatsaechlich als Mass fuer die "Objektivitaet einer Beobachtungstechnik" (Medley/Mitzel 1971, Sp. 661) interpretiert, so muesste seine Bestimmung und Beurteilung grundsaetzlich im Rahmen eines Instrumentenpretests erfolgen, der eine Entscheidung ueber Abweisung oder Zulassung der Erhebungstechnik herbeizufuehren haette. Nach dem erfolgten Zuschlag wuerde der Einsatz eines einzigen Beobachters genuegen (also keine "Instrumentenstichprobe"), der Messwiederholungen im "klassischen" testtheoretischen Sinne auszufuehren haette. Auch das Stabilitaetsproblem gehoert modellgemaess zur Pruefung des Instruments, da es ebenfalls von einer "Instrumentenstichprobe" ausgeht. Zieht man zum Vergleich z.B. die physikalische Temperaturmessung heran, so entspricht der Einsatz mehrerer Beobachter hier dem Einsatz mehrerer Thermometer (bzw. des mehrmaligen Verwendens eines Thermometers bei "fixierter" Temperatur). Es wird deutlich, dass das Problem der Ermittlung des "wahren" Wertes so nicht angegangen werden koennte. Darueberhinaus hilft auch dieser Ansatz im Bedeutungsproblem nicht weiter, da mehrere "wahre" Werte immer noch konfundiert sein koennen. Immerhin ergibt sich, dass wir den Bedeutungsfehler auf der Ebene des Instruments (also des einzelnen Beobachters) diskutieren und die Frage der "Instrumentenstichprobe" (mehrere Beobachter) als irrelevant abweisen duerfen.

## 2.2.4. Der Bedeutungsfehler als Validitaetsproblem

Die Ueberlegungen des vorigen Abschnitts verweisen deutlich darauf, dass der Bedeutungsfehler zur Kategorie der Validitaetsprobleme zu rechnen ist. Es handelt sich bei ihm allerdings um einen Sonderfall der Inhaltsvaliditaet (content validity) insofern, als nicht "direkt" ueberlegt werden muss, "was das ist, was das Instrument misst", Jede einzelne vorgegebene Skala provoziert ja das "Messgeraet" (den Beobachter) zur Abgabe eines "Messwertes", der mit den uebrigen Werten nach Erwartung nicht uebereinzustimmen braucht, weil sich in ihm etwas jeweils anderes widerspiegelt. Man kann dies auch so ausdruecken, dass jeder Wert auf einer Skala das Ergebnis einer Messung mit einem jeweils anderen Messinstrument sei (analog dem Beispiel des elektrischen "Vielfachmessgeraets"; vgl. 2.1.2.). Damit lautet aber die Validitaetsfrage im Hinblick auf den Bedeutungsfehler so; Ist das, was mit Instrument A gemessen (mittels Skala a ausgedrueckt) wird, verschieden von dem, was mit Instrument B gemessen wird? Oder; In welchem "Umfang" messen Instrument A und B das gleiche? Prinzipiell ist es fuer die Beantwortung erforderlich, das Inhalts-Validitaetsproblem fuer jedes einzelne "Instrument" zu loesen und danach die Ueberschneidungsfrage zu pruefen. Die damit zusammenhaengenden Schwierigkeiten sind z.T. mit den in der "klassischen" Testtheorie diskutierten identisch (vgl. z.B. Lienert 1967). Dabei denken wir weniger an die quantitative Ermittlung eines Koeffizienten als an das Problem der inhaltlichen Bestimmung desjenigen Merkmalsuniversums, aus dem die "gemessenen" Objekte eine Stichprobe darstellen sollen. Der entscheidende Unterschied zur Testtheorie besteht naemlich darin, dass der Repraesentationsschluss nicht eigentlich auf eine (Lehr-) Person (vgl. Michel 1971, 49), sondern auf (Lehr-) Personen in/und (Lehr-) Situationen erfolgt.

Dies alles - und darin liegt ein weiteres Grundproblem - gilt nur fuer den Fall, dass sich die Schaetzaufgabe auf theoretische (psychologische) Konstrukte oder wenigstens auf eine Vorform von ihnen ("freundlich") bezieht; Haeufigkeitsratings ueber beobachtbare Verhaltensweisen (behavior) weisen im Prinzip ueberhaupt keine testtheoretische, sondern "lediglich" eine erkenntnistheoretische Problemdimension (Basissaetze (Popper)) auf. In der Literatur fehlt die Diskussion darueber, ob die Abgabe von Schaetzurteilen nicht im Kontext des Definitionsproblems rekonstruiert werden koenne, und zwar in dem Sin-

ne, wie wir oben (2,1,) die Zuordnung von Wort und Bedeutung dargestellt haben, Wir sind der Auffassung, dass diese Interpretation (Wort als Definiendum und Bedeutung (verbal oder nonverbal) als Definiens) einen plausiblen und zugleich viele Schwierigkeiten vermeidenden Loesungsansatz abgeben koennte, Jedenfalls liessen sich das Validitaetsproblem (vgl, z,B, auch v,Cranach/Frenz 1969, 283-285, 305-307) und, wie weiter unten (2,3,) noch gezeigt werden soll, auch einige theoretische und metatheoretische Fragen umgehen bzw, praezisieren, Mit der Abtretung des Validitaetsurteils an "Fachleute der empirischen Sozialforschung und der Didaktik" (Schulz/Teschner/Voigt 1971, Sp, 657; Langer/Schulz v,Thun 1974, 120) scheint uns hingegen eher die Verschleierung des Problems (eine der "Expertenbefragung" generell innewohnende Tendenz!) als seine Klaerung erreicht zu werden,

## 2,2,5, Der Korrelationseffekt des Bedeutungsfehlers

Fuer die Auswertung von Ratinguntersuchungen werden meist parametrische Verfahren vorgeschlagen (vgl, Schulz/Teschner/ Voigt 1971, Tent 1971) und angewandt (vgl, z,B, Ryans 1961, Mueller-Wolf 1977, die Zusammenstellungen bei v,Cranach/Frenz 1969, 314-315 und 320-321),<19> Solche Prozeduren bauen i,d,R, auf der Analyse der Dispersionsmatrizen (Varianz-/Kovarianz-Matrizen) im Rahmen von P-, Q- oder R-"technischen" Ansaetzen auf, Gelegentlich werden auch Aehnlichkeits- oder Distanzmasse zugrundegelegt (z,B, fuer Cluster-Analysen), Das in diesem Zusammenhang wohl am weitesten verbreitete multivariate Modell ist die Faktorenanalyse mit ihren verschiedenen Abwandlungen, Dort wird versucht, die Korrelations- bzw, Kovarianzmatrix durch eine (mathematisch) einfachere Matrix (factor pattern) in Verbindung mit Rechenregeln zu reproduzieren, Die (bei bestimmten Annahmen) "beste" Loesung wird zur Grundlage der (realwissenschaftlichen) Interpretation gemacht, Es ist dabei wichtig, wie sich weiter unten noch zeigen wird (vgl, 2,2,6,), den "Einsprung" in die (formalwissenschaftliche) Modelldarstellung (Quantifizierung und statistische Rekonstruktion der Daten) und den "Ruecksprung" in die inhaltliche Deutung im Auge zu behalten,

Fragt man nach den Auswirkungen des Bedeutungsfehlers im statistischen Modell, so stoesst man schnell auf den Sachverhalt, dass die Struktur der Korrelations- (Kovarianz-)matrix wesentlich durch ihn gepraegt sein kann, In dem Masse naemlich, in dem die Bedeutungsueberschnei-

dung von zwei Beobachtungswoertern zunimmt, steigt auch die Korrelation zwischen den auf ihrer Grundlage ermittelten Daten. Diese Korrelation beruht aber einzig und allein darauf, dass sich die Werte der beiden Variablen (z.T.) auf ein und dieselben Merkmale beziehen. Mit anderen Worten: die errechnete Ko-Variation zwischen den Variablen erwaechst nicht aus der "Aehnlichkeit" der Veraenderungen verschiedener Merkmale; vielmehr drueckt sich in ihr (z.T.) die "Aehnlichkeit" der Veraenderungen gleicher Merkmale aus (als Bedeutungsgemeinsamkeiten verschiedener Beobachtungswoerter; vgl. Abb. 4). In Wirklichkeit handelt es sich dann nicht um eine "Aehnlichkeit", sondern um (teilweise) Uebereinstimmung der in Daten ausgedrueckten Merkmale und damit also nicht um Ko-Variation (oder Ko-Relation), sondern um Identitaet - ein Artefakt.

Interpretationen, die auf solchen Analysen beruhen, gehen in die Irre, wenn sie von vornherein die Ueberschneidungsfreiheit der Merkmale und damit der Daten unterstellen und als Kovariation deuten. Da die Daten bei teilweiser Bedeutungsueberschneidung einen interpersonell (situativ und langfristig), intrapersonell und "intersemantisch" schwankenden und zugleich unbekannten Anteil an Scheinvarianz spiegeln, ist weder eine Deutung im Hinblick auf den beobachteten Sachverhalt noch auf den Sprachgebrauch der Beobachter moeglich. Auch der Versuch, "faktorenreine" Skaleninstrumente zu entwickeln (Skalen spannen als Koordinaten einen orthogonalen Raum auf), kann mit einem solchen Verfahren nicht gelingen (vgl. Ryans 1960, Becker 1964, Tausch 1970, 162 ff. und die Zusammenstellung von Schmitz 1975), da hier der Anteil der "echten" Merkmalsvarianz als Stoergroesse bekannt sein muesste. Als Konsequenz daraus ergibt sich, dass fuer beide Untersuchungszwecke, Merkmals- und Sprachgebrauchsanalyse, jeweils spezifische Vorkehrungen zu treffen sind, die den Einfluss der wechselseitigen "Stoervarianz" auszuschalten vermoegen. Wir werden fuer den ersten Fall weiter unten (3.) einige Vorschlaege diskutieren. Ob es allerdings sinnvoll ist, den Versuch der Konstruktion eines faktorenreinen mehrdimensionalen Ratinginstruments zu unternehmen, halten wir angesichts unserer obigen Ueberlegungen (2.1.) fuer aeusserst fraglich. Nach den Ergebnissen von Mueller-Wolfs Hochschuluntersuchung (1977, 98-101) waeren z.B. folgende drei unabhaengigen "Dimensionen des Lehrverhaltens" in der Beobachtung/Beurteilung zu unterscheiden:

I; "Emotional positives >demokratisches< vs. emotional negatives >autoritaeres< Lehrverhalten"

II; "Verstaendlicher, didaktisch anregender vs. unverstaendlicher, didaktisch ineffektiver Vorlesungsstil"

III; "Studentenorientiert-strukturierendes vs. monokratisch-dirigistisches Lehrverhalten".

Es ist u.E. unabweislich, dass die Vorgabe solcher "Beobachtungswoerter" als Beschriftungen von Skalenextremen erneut all jene Bedeutungsprobleme aufwerfen wuerde, die wir bisher eroertert haben. Sie wuerden manifest, wenn man den formal zulaessigen und inhaltlich in gleicher Weise zu rechtfertigenden zweiten Schritt einer analogen Untersuchung auf der Basis dieser drei Dimensionen machen wuerde; man kaeme zu einer weiteren Reduzierung (der bestmoeglichen Zwei-Faktoren-Loesung) und endete schliesslich - das gilt generell - beim eindimensionalen Design. Dass dies weder beabsichtigt noch theoretisch befriedigend waere, braucht kaum betont zu werden (wir werden weiter unten (2.3.) einige Gruende dafuer diskutieren).<20> Die Crux fuer die Unterrichtsforschung besteht aber darin, dass sich damit auch die dem ersten Untersuchungsschritt innewohnende Forschungsstrategie als fragwuerdig erweist, dass mithin eine beachtliche Zahl von Untersuchungen (vgl. 4.) unter diesem Aspekt als weniger effizient bezeichnet werden muss.

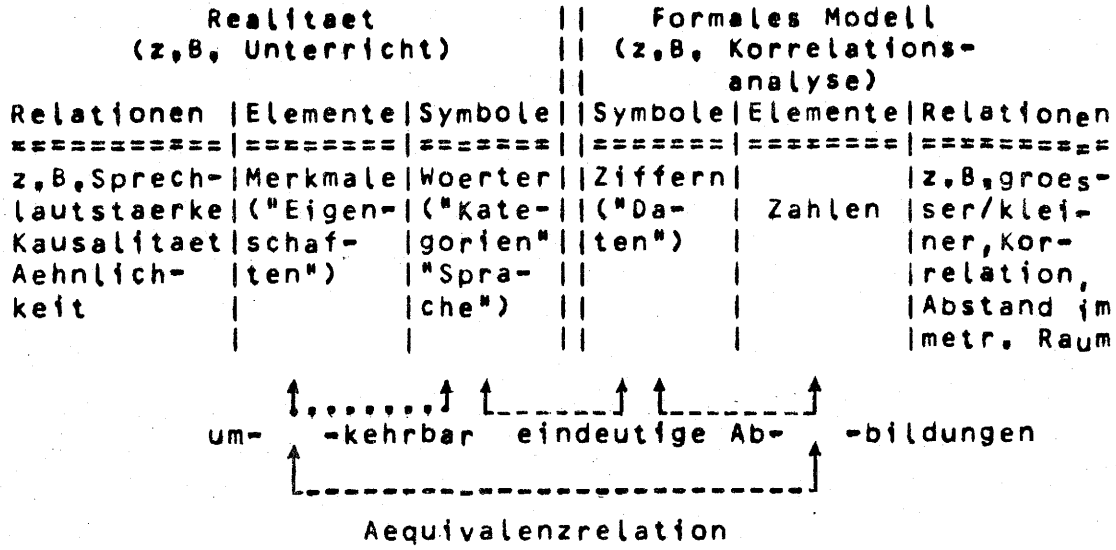
### 2.2.6. Messtheoretische Aspekte der Wort-Bedeutung-Relation

Obgleich wir im Rahmen dieser Untersuchung keinesfalls die Grundsatzdebatte ueber den methodologischen Status sozialwissenschaftlicher Messungen aufnehmen koennen, halten wir einige Bemerkungen zur Problematik der Schaetzskalen unter diesem Gesichtspunkt fuer erforderlich. Dies nicht zuletzt deshalb, weil unsere eigene Untersuchung, in der Daten dieses Genres eine Rolle spielen (vgl. 3.), davon nicht unberuehrt bleibt.

Der Haupteinwand gegenueber einer verbreiteten Praxis erfahrungswissenschaftlicher Individual- und Sozialforschung besteht darin, dass sie die erforderliche scharfe Trennung zwischen dem materiellen Gehalt der Empirie (Theorie) und dem formalen Gehalt des Modells verwische und damit den unterschiedlichsten Irrtuemern verfalle (vgl. z.B. Cicourel 1970, 28 ff.). Die Verwechslung von

Kausalitaet und Korrelation oder die Rede von 'signifikanten Merkmalen' von Personen, Gruppen etc. sind dafuer nur die augenfaelligsten Beispiele. Das eigentliche Problem liegt in der fehlenden Axiomatisierung der (materiellen, erfahrungswissenschaftlichen) Theorien, die fuer die Anwendung (isomorpher) mathematisch-statistischer Modelle zur Voraussetzung gemacht werden muss (vgl. Cicourel 1970, 20-22 und passim). Soll naemlich die formalwissenschaftliche Analyse von Daten nicht eine sinnentleerte Spielerei sein, so muss gezeigt werden, dass Aequivalenz (vgl. Meschkowski 1972, 28-29) besteht zwischen den realen (Theorie-)Elementen, ihren Beziehungen sowie Veraenderungsmoeglichkeiten einerseits und den formalen (Modell-)Elementen, ihren Relationen und Variationsmoeglichkeiten andererseits. Erst dann ist es ueberhaupt nuetzlich, reale Ereignisse bzw. Wahrnehmungen mittels Ziffern (vgl. Stevens 1951, 1) auf formale Modelle abzubilden, die Abbildung unter bestimmten Aspekten zu untersuchen und den Befund zurueckzuuebertragen.

Die generelle Problematik besteht nun darin, dass quantitative Analysen eine realwissenschaftliche Theoriebildung voraussetzen, die dem Aequivalenzpostulat - wenigstens prinzipiell - genuegen. In unserem speziellen Fall erweist sich der Bedeutungsfehler auch unter diesem messtheoretischen Aspekt als fundamental(2-6): die erforderliche umkehrbar eindeutige Beziehung zwischen den (beobachteten) Merkmalen und Zahlen ist solange nicht erfuellt, als unklar bleibt, nach welchen Regeln (s. 2-6 gepunkteter Pfeil) die Beobachtungswoerter auf die Realitaet angewendet werden (vgl. Cicourel 1970, 29).




## 2-6 Der messtheoretische Ort des Bedeutungsfehlers

Folgt man dem Vorschlag, den unter vielen anderen Langer und Schulz v.Thun (1974, 127 ff.) machen, und beschraenkt sich nicht auf die Bezeichnung eines Skalenendes, so begibt man sich in die zusaetzliche Gefahr, dass durch die sprachlichen Unterschiede in den Beschreibungen der Skalenstufen bereits innerhalb ein und derselben Messdimension Inkonsistenzen auftauchen, die sonst "lediglich" zwischen ihnen zu konstatieren waeren: der Beobachter kann sich veranlasst sehen, die Merkmals- und Gewichtskonzepte gemaess den variierenden Wortvorgaben von Stufe zu Stufe zu wechseln; er wuerde dann nicht Intensitaets-/Haeufigkeitsvariation innerhalb eines Wortkonzeptes beachten, wie die Skalenkonstruktion es (in diesem Fall: dem Forscher) suggeriert, sondern "kategorisieren"; seine Ankreuzungen repraesentierten also nominelle oder qualitative Daten und nicht (wenigstens) ordinale (vgl. Rosenshine 1974, 203 und zum damit zusammenhaengenden Problem der "Aequidistanz" auf Skalen: Rohrmann 1978).

Die Moeglichkeitsbedingung mathematisch-statistischer Auswertung von Daten besteht, nach dem bisher Gesagten, im vorausgesetzten Entwurf eines realwissenschaftlichen Aussagesystems. Seine Elemente muessen so konstruiert sein, dass die zwischen ihnen gedachten Relationen zu denen im Modell isomorph sind. Die Bedenken, die gegenueber Ratingskalen in dieser Hinsicht vorgetragen wurden, muessen weiter unten (3.) auch gegenueber den Daten geprueft werden, die wir im Zusammenhang mit einer eigenen Untersuchung erhoben haben. Wir werden dort zu zei-

gen versuchen, dass derjenige Teil unserer Analyse, der zu methodologischen Aussagen herangezogen wird, im Prinzip den genannten Anforderungen genuegt,

## 2,3, Theoretische und metatheoretische Implikationen des Bedeutungsfehlers

### 2,3,1, Das Verhaeltnis von Schaetzung und Messung als Problem

Es mag verwundern, dass in der Fuelle der Literatur zur Datenerhebung mittels Skalen oder skalenaehnlichen Instrumenten eine praezise Kennzeichnung der zu erbringenden Messleistung nirgends zu finden ist. I.d.R. beschraenken sich die Autoren auf Umschreibungen, in denen die Begriffe Schaetzung (rating), Urteil (judgement) und Beobachtung (observation) in den unterschiedlichsten Kombinationen vorkommen (vgl. die Aufzaehlung bei Remmers (1971, Sp. 858)). Es scheint, als werde die naehere Bestimmung des Verfahrens zugunsten der Beschreibung seiner Ergebnisse vernachlaessigt; deren Vielfalt kann dagegen den Eindruck vermitteln, dass ihre Entstehung gar nicht mit einer irgendwie einheitlichen Prozedur herbeigefuehrt werde.

So spricht z.B. Fassnacht (1979, 148) eher vorsichtig davon, "der Beobachter (schaetze) das Ausmass eines bestimmten Verhaltens oder bestimmter Verhaltensaspekte (ab)"; v.Cranach und Frenz (1969, 272) und Hasemann (1971, 820) heben demgegenueber hervor, dass aufgrund von Beobachtungen Urteile abgegeben werden, die sich auf Eigenschaften von Objekten beziehen. Langer und Schulz v.Thun (1974, 102-103) differenzieren nach der Art der "Informationsverarbeitung" in den Kategorien "einfach (a), komplex (b), rational (c) und emotional (d)" vier verschiedene Messleistungen, die sich als Schaetzung (a, c), Empfindung (a, d), Eindruck (b, c) und Erlebnis (b, d) manifestieren. Sie betonen aber, dass die Grenzen fliessend seien und dass man den Anspruch auf eine scharfe Trennung der "Merkmale eines Ratings" (103) gar nicht erheben duerfe. Solche eher verschleiernden als klaerenden Aussagen gipfeln in der Erklaerung Remmers' (1971, Sp. 857), dass das, was "wir ueber das Verhalten und die Eigenschaften anderer Menschen oder Gegenstaende wissen wollen, uns mitunter nur von einem Registrierinstrument mit der Empfindlichkeit, Komplexitaet und Wachsamkeit des menschlichen Beobachters vermittelt werden (koenne)."

Beinahe unausweichlich bestehen angesichts der Unklarheit des Messvorgangs auch tiefgreifende Meinungsverschiedenheiten ueber die Relevanz des Verfahrens in der empirischen Forschung. Sie reichen von der Forderung, es "selbstverstaendlich" in Konkurrenz mit psychometrischen Tests treten zu lassen (v.Cranach/Frenz 1969, 285,

323), bis zu der Behauptung, man koenne beide nicht "sinnvoll miteinander vergleichen, da es sich um zwei voellig verschiedene Dinge (handele)" (Fassnacht 1979, 52).<21> Guilford wiederum (1954, 296) raet im Anschluss an Paterson, man solle Skalen nicht verwenden, wenn reliablere und objektivere Daten zur Verfuegung gestellt werden koennten.

Wir koennen diese Diskussion hier weitgehend vermeiden, weil unsere Fragestellung wesentlich enger gefasst ist und daher fuer eine globale Argumentation keine Basis abgibt. Dennoch laesst sich zeigen, dass der Bedeutungsfehler in gewissem Umfang fuer die geschilderte Situation mitverantwortlich ist. Hinsichtlich der Unklarheiten des Messvorgangs tritt er naemlich im Gewande der Werturteilsproblematik auf, hinsichtlich der wissenschaftlichen Relevanz verbirgt er sich hinter der Unterscheidung von Umgangs-, Beobachtungs- und theoretischer Sprache. In beiden Faellen schliesslich sorgt er fuer zusaetzliche Schwierigkeiten insofern, als Unsicherheit darueber bestehe, ob aus den Messdaten etwas ueber das Messinstrument oder das Messobjekt erfahren werden kann. Diesen Problemen wenden wir uns jetzt unter unserem eingegrenzten Aspekt zu.

### 2.3.2. Schaetzurteile im Lichte der Forderung nach Werturteilsfreiheit

Die wissenschaftstheoretische Debatte in der ersten Haelfte unseres Jahrhunderts ("Werturteilsstreit"), aber auch die Auseinandersetzung der sechziger Jahre ("Positivismusstreit") beschaeftigte sich hauptsaechlich mit der Frage, ob Wissenschaft frei von Werturteilen sein koenne und solle (vgl. Albert/Topitsch 1971; Adorno u.a. 1972). Weniger Aufmerksamkeit erfuhr die damit implizierte Ueberlegung, woran man Werturteile erkennen koenne (vgl. zu diesem Problem Zecha 1976, insbes. 617 ff.). Von diesem Defizit scheinen die Schaetzverfahren in einer besonderen Weise betroffen zu sein: einerseits fuehren sie haeufig den Urteilsbegriff selbst "im Schilde", andererseits werden in ihrem Kontext z.T. recht "sophistizierte" mathematisch-statistische Prozeduren beschrieben, mit deren Hilfe u.a. der Nachweis ihrer "Objektivitaet" erbracht werden soll (vgl. z.B. Guilford 1954; Schulz/Teschner/Voigt 1971, Sp. 799 ff.; Langer/Schulz v.Thun 1974, 62 ff.). Auch wenn man dem Namen kein allzu grosses Gewicht beimisst, mag er doch Anlass zu der Frage geben, ob die Anwendung von Schaetzskalen prinzipiell mit einem Wissenschaftsbegriff verbunden

sei, der mit dem Gebot der Werturteilsfreiheit gebrochen habe. Dass dies nicht der Fall ist, kann - ohne eine (wuenschenswerte) detaillierte Untersuchung - durch den Hinweis auf einige "unverdaechtige" Autoren plausibel gemacht werden (wie z.B. Guilford, Hofer, Ryans), die das Ratingverfahren in ihren eigenen Arbeiten einsetzen.

Mit dieser - systematisch selbstverstaendlich unzulaenglichen - Feststellung soll zunaechst lediglich darauf hingewiesen werden, dass Schaetzurteile nicht von vornherein in den Zusammenhang einer bestimmten wissenschaftstheoretischen Position gerueckt werden duerften, wenn man von der publizierten Forschungspraxis ausgeht.

Die eigentliche Schwierigkeit besteht aber darin, dass die Identifikation von "Werturteilen" eine Reihe von Problemen mit sich bringt (vgl. z.B. Zecha 1976, 627 f.). Das haengt u.a. damit zusammen, dass in Wissenschaftssprachen Woerter aufgenommen werden, die nicht explizit definiert sind, und dass Deskriptionen und Wertungen in gleichem grammatischen Gewande erscheinen koennen ("Das ist gruen," - "Das ist gut."). Immer, wenn ein undefiniertes Wort an der Praedikatstelle einer - grammatisch gesehen - deskriptiven Aussage steht, koennen daher Zweifel darueber aufkommen, ob wirklich eine Deskription vorliegt oder ob etwa eine Wertung ausgedrueckt worden ist.

Es ist offensichtlich, dass je nach der Bedeutungsart der in ein Schaetzverfahren eingefuehrten Skalenbezeichnungen die mit seiner Hilfe gewonnenen Informationen variieren: Wertpraedikate provozieren Aeusserungen des Raters ueber "Einstellungen" gegenueber dem, was er wahrgenommen oder sich vorgestellt hat, sie informieren ueber eine interne "Disposition" des Raters; deskriptive Praedikate rufen Aeusserungen des Raters ueber Wahrnehmungen von etwas hervor, sie informieren ueber rater-externe Sachverhalte. Auf diese Unterscheidung und ihre Konsequenzen ist schon haeufig hingewiesen worden (vgl. z.B. Hofer 1969, 11 und die dort angegebene Literatur); behaelt man sie im Auge, so kann die Anwendung von Schaetzskalen im Rahmen "wertfreier Wissenschaft" insoweit auch als unproblematisch gelten <22>, weil jeweils klar ist, ob Theorien ueber Einstellungen oder "Objekte" generiert, geprueft oder zur Erklaerung herangezogen werden.

Schwierigkeiten ergeben sich, wenn die Trennung nicht vorgenommen (vgl. z.B. bei Diehl/Kohr 1977) oder die Auffassung vertreten wird, "Urteile" informierten in Wirklichkeit doch ueber Objekte, also beispielsweise ueber Lehrverhalten (vgl. z.B. bei Mueller-Wolf 1977,

42, 93). Im ersten Falle bleibt es dem Beurteiler ueberlassen, welche Deutung der einzelnen Items er vornimmt und ob er sich ueberhaupt Klarheit in dieser Hinsicht verschafft.<23> Ergebnisse aus solchen Erhebungen, in denen ja ausserdem die Verschiedenheit der Beurteilerdeutungen fuer ein und dasselbe Item zur Unvergleichbarkeit der Daten fuehrt, koennen als theoretisch irrelevant bezeichnet werden. Dass sie nicht replizierbar sind, wie es sich fuer den Fragebogen von Diehl und Kohr zur Beurteilung von Hochschulveranstaltungen erwies (vgl. Kleine/Merkens 1979), ist allein schon aus diesen Gruenden zu erwarten.

Aber auch im zweiten Falle laesst sich ein theoretischer Bezug nicht praezise konstruieren. Weil die Varianzquelle "Haltungs-(Einstellungs-)verschiedenheit" in den Daten nicht von der Quelle "Objektverschiedenheit" zu trennen ist, bleibt offen, ob diese fuer Entitaeten aus einer Wenn- oder einer Dann-Komponente stehen.<24> Besonders "verdunkelnd" wirkt in diesem Zusammenhang das Ratertraining. Ginge es naemlich um die Erfassung von "subjektiven Daten"(vgl. Langer/Schulz v.Thun 1974, 103; Mueller-Wolf 1977, 35), so waere gar nicht zu verstehen, aus welchem Grunde die Beurteiler darin geschult werden sollten, einheitliche ("reliable") Aussagen zu machen. Auf diese Weise wuerden ja gerade jene Unterschiede verwischt, die zu erfassen Aufgabe der Untersuchung ist. Darueberhinaus bliebe unklar, ob der Vereinheitlichungseffekt auf der Beseitigung von Sprachdifferenzen oder von Haltungsunterschieden eingetreten ist.<25> Ein Reliabilitaetsmass liegt in Wahrheit dann gar nicht vor; der errechnete Koeffizient drueckt hier vielmahr eine Merkmalskorrelation im Sinne einer R-Analyse aus.<26>

Unterstellt man andererseits, dass es fuer ein zu beobachtendes Ereignis einen "richtigen ... Skalenwert" (Langer/Schulz v.Thun 1974, 137; ebenso Friedrich/Luedtke 1973, 69) gibt, so drueckt sich darin die Absicht aus, theoretisch relevante Angaben ueber ein Beobachtungsobjekt, nicht ueber seine Wirkung zu beschaffen. Die methodischen Konsequenzen dieses Ansatzes haben wir oben (2.2.2.) bereits diskutiert.<27> Sie fuehren grundsaetzlich weg vom Rating, genauer: sie implizieren den Versuch, sich instrumentell moeglichst wenig von "idealen" Messbedingungen zu entfernen. Beobachtungsschaetzungen liegen aber nicht zuletzt wegen des Bedeutungsfehlers am entgegengesetzten Ende der Praezisionsskala; sie fuehren ueber Alltagspraxis prinzipiell nicht hinaus (vgl. Langer/Schulz v.Thun 1974, 22). Die mangelnde intersubjektive Kontrollierbarkeit ihrer Genese erweist sich als Defizit des Informationsgehalts der auf ihnen begruendeten Theorien.<28>

Eine ganz andere Qualitaet kommt dem Ratingverfahren im Kontext der Handlungsforschung zu (vgl. z.B. Moser 1975).<29> Dort kann es zur Initiierung angestrebter gruppendynamischer Prozesse, zur Bewusstmachung individueller und sozialer Beduerfnisse (Selbst-/Fremdrating; vgl. Anm. 24), zur diskursiven Verstaendigung etc. hilfreich sein. Diese Verwendung erwaechst aber aus einem voellig anderen Wissenschaftsverstaendnis als dem oben zugrundegelegten. In ihm ist die Trennung von "objektiver Deskription" und Urteil, von "Erkenntnis und Interesse" (Habermas) "aufgehoben". Das bedeutet zugleich, dass dort das Anlegen testtheoretischer Guetemassstaebe als gaenzlich unangemessen zu betrachten ist (vgl. Moser 1975, 117-127). Tatsaechlich unterscheidet sich diese metawissenschaftliche Position so entschieden von einer analytischen (etwa i.S. Esslers 1972), dass die vielerorts versuchte wissenschaftstheoretische (vgl. z.B. Klafki 1976) wie forschungspraktische (vgl. z.B. Mueller-Wolf 1977) Verbindung beider als logisch von vornherein unmoeglich und hinsichtlich des Gehalts der formulierbaren Aussagen als willkuerlich bezeichnet werden muss.

### 2.3.3. Das terminologische Niveau von Schaetzurteilen in der Praxis der Unterrichtsforschung

Mit der Entscheidung fuer den Ausschluss von Werturteilen aus dem (objektsprachlichen) Korpus der als wissenschaftlich geltenden Aussagen ist der systematische Status von Schaetzinformationen noch nicht eindeutig geklaert. Seit den Diskussionen des Wiener Kreises hat sich, insbesondere im Anschluss an Carnap, bei den analytisch-empirisch orientierten Wissenschaftlern der Gedanke durchgesetzt, dass eine Unterscheidung zwischen theoretischen und Beobachtungsbegriffen forschungslogisch unentbehrlich sei (vgl. z.B. die Darstellung bei Stegmueller 1974). Zwar wirft die Konstruktion der Relation zwischen beiden erhebliche Probleme auf, die z.T. bis heute nicht voellig geloest sind (vgl. ebd.); dennoch wird von der Grundidee gerade auch in den Sozialwissenschaften durchweg selbstverstaendlicher Gebrauch gemacht (vgl. z.B. Falter 1977): es wird davon gesprochen, dass Hypothesen auf der Ebene der theoretischen Sprache entworfen werden koennen, und dass man, um sie zu pruefen, die in ihnen enthaltenen ausserlogischen Begriffe operationalisieren, d.h. in die Beobachtungssprache "uebersetzen" muesse (vgl. z.B. Opp 1976).

Gerade in der Umgehung dieser Prozedur wird ein wichtiger, oft entscheidender Vorteil des Ratingverfahrens gesehen. Langer und Schulz v.Thun versuchen sogar unter Verwendung eines Beispiels von Wolf zu zeigen, dass Operationalisierungen zu Konsequenzen fuehren, "die den Empiriker sowohl beim Laien als auch beim Geisteswissenschaftler mit Recht suspekt (machen)" (1974, 19). Sie fuehren dies auf den Umstand zurueck, dass man beim "Herunterkonkretisieren" den "Bezug zum theoretischen Ausgangsbegriff" verliere und bei Trivialitaeten" oder immer noch ">ratingbeduerftigen<" Begriffen lande (20, 21). Gegenueber einer solchen "elementaren Indikatorloesung" favorisieren sie die dem Rating innewohnende "Indikatorenverschmelzung", die vom Beobachter "im Wege einer gestalthaften Wahrnehmung der Gesamtsituation " (21) durch die dem menschlichen Gehirn eigene Faehigkeit "zur automatischen Integration einer Vielzahl von Einzelindikatoren" (20) geleistet werde.

Eine aehnliche, aber in wichtigen Punkten doch verschiedene Differenzierung benuetzt Bachmair (1977, 20-21). Nach seiner Ansicht sind Kategorien unterschiedlicher Ebenen voneinander zu unterscheiden, die durch eine 'Enthaltenseinsrelation' miteinander verbunden sind; "uebergeordnete Kategorien setzen sich aus "kleineren" Unterrichtsereignissen zusammmen, die in sich wieder gegliedert sein koennen." Mit dieser Konstruktion wird die Beschreibungssprache nicht wie bei Langer und Schulz v.Thun verlassen; waehrend bei ihnen von Indikatoren auf etwas anderes Unanschauliches zurueckgeschlossen werden muss, ist hier nur von Zusammenfassungen die Rede. Operationalisieren bedeutet also bei Bachmair das Aufloesen einer "Summe" in ihre "Summanden", bei Langer und Schulz v.Thun eher den Uebergang von "theoretischen" Ursachen auf Wirkungen ("Indikatoren"). In diesem Sinne koennen die beiden zuletzt genannten Autoren davon reden, dass man dabei "unterwegs" etwas verliere (20); es werden ja nicht alle Wirkungen der gedachten Ursacne betrachtet.

Wieder eine andere Nuance drueckt sich in der bei Fassnacht (1977, 91-95) eingefuehrten Unterscheidung von konkretem und abstraktem Sprachbezug aus. Danach verfuegen wir ueber Kenntnisse von der "sprachlichen oder nur gedachten abstrakten Bedeutung" eines Begriffs und koennen ihr den "konkreten Bezug" durch "von jedermann nachvollziehbare Operationen" (93) zuordnen. Weil dabei "Unkorrektheiten" unterlaufen koennen (95), empfiehlt der Autor die "ehrlichere" Methode, gar nicht erst von abstrakten Begriffen auszugehen; man vermeidet s.E. damit zugleich den Operationalisierungsfehler und die Probleme zweier "Kode-Systeme" (93).

Nach Ansicht von Mueller-Wolf muessen "Variablen" unterschiedlicher "Komplexitaet" voneinander unterschieden werden, und zwar in Entsprechung zum (Unterrichts-)Geschehen; weil "wesentliche Prozesse" dort "in hoechstem Masse komplex" seien, koenne mit "atomistischen" Variablen nicht angemessen gearbeitet werden, obwohl diese, im Gegensatz zu den komplexeren, leicht messbar seien (36). Hinter dieser Auffassung steht - wenn wir recht sehen - der Gedanke, dass Begriffe nach ihrer Bedeutungsstruktur differenziert werden koennen. Komplexitaetszuwachs wird in dieser Vorstellung insbesondere auch durch die Einbeziehung von Wirkungsaspekten erreicht; die unangemessene Trennung der unterrichtlichen Interaktion in "Bedingung" und "Auswirkung" soll durch den Untersuchungsansatz von Mueller-Wolf aufgehoben, d.h. also auch begrifflich eingeholt werden (vgl. 32).

In der amerikanischen Literatur findet man in diesem Problemzusammenhang haeufig die Unterscheidung von Begriffen, mit deren Anwendung geringere oder hoehere schlussfolgernde Leistungen des Beobachters verbunden seien ("low-/high-inference variables"; z.B. Rosenshine/Furst 1973, 166; vgl. auch Gruemer 1974, 889-93). Diese Differenzierung bezieht sich also nicht, wie die soeben dargestellten, direkt auf die Wortbedeutung, sondern darauf, welche Messleistungen erbracht werden muessen: je groesser die 'kausale Distanz' zwischen Beobachtung und Messobjekt, desto hoeher der Begriffsrang. Unter Bezugnahme auf Davis, Gerard, Jones und Weick heben v.Cranach und Frenz (1969, 283-284) die Unterscheidung von Schluessen auf Handlungsursachen und Handlungsfolgen hervor. Dabei wird allerdings nicht geklaert, ob auch dann noch von 'hoch-inferenten' Begriffen gesprochen werden darf, wenn in der Ursache-Wirkungskette nach einigen Gliedern sichtbare Ereignisse (z.B. Schuelerverhalten, Erziehungsverhalten der Eltern des beobachteten Lehrers) auftreten.

Schliesslich entwickeln Friedrichs und Luedtke die Vorstellung, dass die "zentralen soziologischen Begriffe ... unpraezis definiert" seien; durch ">Operationalisierungen<" muesse ihnen, um diesen Mangel zu beheben, "empirischer Gehalt gegeben werden" (1973, 62).

Eine Zusammenstellung der skizzierten Auffassungen ergibt den folgenden Ueberblick:

| Autoren | Begriffsdifferenzierung |
|---|---|
| Bachmair (1977) | niedrige bis hohe Kategorienebene |
| Fassnacht (1979) | konkret vs. abstrakt |
| Friedrichs/Luedtke (1973) | praezise vs. unpraezise |
| Langer/Schulz v. Thun (1974) | elementare vs. "verschmolzene" Indikatoren |
| Mueller-Wolf (1977) | atomistisch bis komplex |
| Rosenshine/Furst (1973) | low to high inferent |

2-7 Begriffskonzepte in der Unterrichtsforschung

## 2.3.4. Die Praxisrelevanz von Unterrichtsforschung auf der Basis von Ratingdaten

So verschieden - und teilweise unklar, ja inkonsistent - die Vorstellungen der einzelnen Autoren auch sein moegen, einig sind sie sich darin, dass sie durch die jeweils eingefuehrte Unterscheidung zugleich ein Relevanzkriterium gefunden haetten; die schwieriger zu erfassenden Variablen sind danach stets auch die "relevanten", die leichter zugaenglichen die "trivialen" (vgl. die oben angegebenen Literaturstellen) und das Rating als Verfahren der Erhebung daher unentbehrlich. Langer und Schulz v.Thun praezisieren, was damit gemeint ist: "Kein anderes... Messverfahren kommt so nah an die Erlebnisrealitaet von Menschen und deren alltaeglich zur Lebensbewaeltigung und Entscheidungsbildung praktiziertes Einordnen ihrer Umwelt heran" (1974, 22).

Zwar verweisen die Autoren darauf, dass "wissenschaftliche Beobachtung und Alltagsbeobachtung" nicht "gleichgesetzt" werden duerfen (ebd.), Dennoch wird der zugrundeliegende Gedanke deutlich formuliert: als Relevanzkriterium gilt die Erhellung alltaeglicher Erfahrung und das heisst fuer die hier zitierten Autoren zugleich, dass dies nur unter Verwendung der Alltagssprache gelingen koenne. Diese Ueberlegung soll, trotz ihrer grundsaetzlichen Bedeutung und der dahinterliegenden philosophischen wie wissenschaftstheoretischen komplizierten Materie, hier in aller Kuerze aufgegriffen und aus der spezifischen Sicht der Unterrichtsforschung

schlaglichtartig beleuchtet werden. Zunaechst ist bei der vorgetragenen These zur Notwendigkeit umgangssprachlicher Terminologie in der Personwahrnehmung zu unterscheiden zwischen dem Mittel- und dem Zweckargument. In Frage steht hier nicht der Zweck, also die Beschaffung von (unterrichts-)praktisch relevantem Wissen, sondern die Hoffnung, dass dieses Ziel mit einer solchen Sprache besser erreicht werden koenne als mit zwei anderen Moeglichkeiten, naemlich der Verwendung von "echten" Beobachtungswoertern oder der Einfuehrung von unanschaulichen, eventuell intuitiv unzugaenglichen theoretischen Begriffen.<30> Waehrend naemlich im ersten Falle das ins Auge gefasste Problemniveau nicht erreicht werden koenne (vom "Lidschlag" zum "Unterrichtsklima"), fehle im zweiten die Herstellbarkeit eines inhaltlichen Bezugs, eines angemessenen Problemgehalts ("Vigilitaet" vs. "Unterrichtsklima").

Wir muessen hier anfuegen, dass das zweite Argument in den von uns benuetzten Quellen nicht explizit auftaucht und dass - billigerweise - auch der Versuch unternommen werden muesste, die vorgeschlagenen Skalenbezeichnungen als theoretische Terme i.S. der Carnapschen Zweistufentheorie zu rekonstruieren. Zumindest schliessen die Autoren eine solche Moeglichkeit nicht explizit aus, sie lassen mit der Verwendung des Operationalisierungsbegriffs sogar ausdruecklich diese Assoziation zu. Eine genaue Pruefung der in Tab. 2-7 aufgelisteten Begriffskonzepte - wir koennen dies hier nicht im einzelnen ausfuehren - ergibt jedoch, dass keines davon diesem Anspruch genuegt, und zwar schon deshalb nicht, weil in ihm gerade jene forschungslogische Strategie impliziert ist, die von den Verteidigern der Schaetzverfahren als unerfuellbar bezeichnet wird: nach der Zweistufentheorie beduerfen die deskriptiven theoretischen Terme zum Zwecke der - experimentellen oder praktischen - Anwendung der (partiellen) Interpretation durch Zuordnungsregeln, die den Bezug zu Beobachtungstermen herstellen, deren Bedeutung sich auf sinnlich (im engeren Sinne) wahrnehmbare Eigenschaften oder Relationen von Realitaet beschraenkt (vgl. die Darstellung bei Stegmueller 1974, 293 ff.).

Im Rating soll dagegen das Unanschauliche "unmittelbar" erfasst werden, genauer: die Vermittlung wird der Intuition des Beobachters (auch des geschulten!) ueberlassen. Dies ergibt aber eben nicht nur einen defizienten Modus der Verwirklichung des Zweistufenkonzepts, sondern eine Alternative dazu; war fuer diese konstituierend, dass der Uebergang von der theoretischen zur Beobachtungssprache explizit und prinzipiell unabhaengig von der Beobachtung selbst erfolgen muesse, so gilt fuer jene -

ebenso grundsaetzlich - die unaufloesliche Verbindung von beidem, die These von der Untrennbarkeit des Datenerhebens und des Theoretisierens.

Freilich wird dies nicht so in aller Schaerfe formuliert, wie ueberhaupt metatheoretische Ueberlegungen im Kontext der Literatur zum Ratingverfahren eher spaerlich sind. Aber aus der Art der vorgetragenen methodologischen Argumente laesst sich diese Gedankenfuehrung unschwer erkennen. Darueber duerfen auch die "nachgeschobenen" mathematisch-statistischen Auswertungsverfahren nicht hinwegtaeuschen (vgl. z.B. den Hinweis bei Langer/Schulz v.Thun 1974, 20). Sie beruhen auf Daten, fuer die, wie wir bereits dargelegt haben, weder eine ausreichende axiomatische Grundlage (Metrik etc.) gelegt ist noch - wegen des Bedeutungsfehlers - klar gesagt werden kann, auf welchen Gegenstandsbereich sie sich beziehen.

Kehren wir zur genaueren Lokalisierung des Bedeutungsfehlers in diesem Problembereich noch einmal zu der Argumentation gegen die Verwendung "echter" Beobachtungsbegriffe in der Unterrichtsforschung zurueck. Dort wurde gesagt, dass mit ihnen keine "paedagogisch relevante(n) Kategorien" (Bachmair 1977, 20) ins Spiel gebracht seien, Unterrichtsforschung auf sie also nicht bauen koenne. Damit ist vorausgesetzt, dass andere "Kategorien" zur Verfuegung stehen, die mindestens das - methodologische - Merkmal der Unanschaulichkeit (Nicht-Beobachtbarkeit i.e.S.) und das - metatheoretische - Merkmal der inhaltlichen Relevanz ("Praxisnaehe" o.ae.) aufweisen muessen - das erste nach Definition, da sonst das Schaetzverfahren ueberfluessig waere, das zweite entsprechend dem empfundenen Problemniveau, das auf der Ebene einer irgendwie gearteten Verbesserung der Unterrichtspraxis liegt. Woerter, die zur Bezeichnung der fuer relevant gehaltenen Merkmale eingefuehrt werden, sind demzufolge auch weitgehend dem Bereich der umgangssprachlichen Situation- und Person-Wahrnehmung entnommen (vgl. 2-1 und die dort angegebene Literatur).

Nun ist bereits verschiedentlich auf die mangelhafte Praezision der Umgangssprache, vor allem aber ihrer "abstrakten Begriffe" hingewiesen worden (vgl. Cicourel 1970, 250 ff.; Engelkamp 1974, 90 ff.; Lenneberg 1977, 406; Stegmueller 1977). Dies gilt auch dann, wenn man die in ihnen haeufig enthaltenen Bewertungsbedeutungen nicht mit veranschlagt (vgl. Friedrichs/Luedtke 1973, 41).

An der Skalenbezeichnung "der Dozent respektiert die Studenten als Partner und Persoenlichkeiten" vs. "er

sieht auf die Studenten herab" mag man sich dieses Problem verdeutlichen (Mueller-Wolf 1977, 198). Der Schaetzer kann sich bei dieser Formulierung auf Merkmalsbereiche beziehen, die von "Dispositionen", "Absichten", ueber "Verhalten", "Handeln" bis "Wahrnehmen", "Empfinden" reichen und jeweils innerhalb dieser Bereiche noch differenziert werden koennten. Es muss unklar bleiben, ob und wo der einzelne Beobachter Schwerpunkte bildet und auch, ob er die erste Bezeichnung eher auf die zunaechst genannten, die zweite eher auf die danach genannten Bereiche bezieht (vgl. 2.2.). Ueberlegt man nun, worauf der beobachtende Student sich beziehen mag, wenn er urteilen soll, ob er "durch die Art des Dozentenverhaltens... persoenlich irgendwie vorangekommen" vs. "nicht vorangekommen" sei (ebd., 202), so wird nicht nur sichtbar, dass er in beiden Faellen ueber die gleichen "Eindruecke" (Langer/Schulz v.Thun 1974, 102) berichten kann<31>; die auf solchen Grundlagen formulierten "praxisrelevanten" Aussagen sinken auf das Niveau von Tautologien, je umfassender die Bedeutungsueberschneidung zwischen solchen "Ursache-" und "Wirkungsvariablen" wird.<32> Wuerde man jedoch den Grad dieser inhaltlichen Konfundierung kennen (im Unterschied zur methodischen; vgl. Anm. 31), so bliebe immer noch offen, worueber Saetze informieren, in denen Skalenbezeichnungen oder etwa aus ihrer Deutung (!) gewonnene Faktorenbenennungen vorkommen.<33>

Der unklare wissenschaftstheoretische Status der Woerter, der sich u.a. in der Inkaufnahme des Bedeutungsfehlers zeigt, belastet die Unterrichtsforschung in erheblichem Masse, sowohl hinsichtlich ihrer erkenntnis- und damit handlungspraktischen Effizienz als auch im Hinblick auf ihre innerwissenschaftliche Reputation ("Binnenlegitimitaet" (Lepsius 1973)), wie beispielsweise die Kontroverse zwischen Walter und Nicklis verdeutlicht.<34> Wuerde es gelingen, den Bedeutungsfehler zu vermeiden, so waere damit ein Schritt getan, der als konservativ in dem Sinne gelten koennte, dass angesichts der vorgetragenen Kritik der Stab nicht endgueltig ueber dieses Verfahren gebrochen wuerde. Trotzdem blieben auch dann noch viele essentielle Fragen offen (vgl. Tab. 1).

Unsere im weiteren durchgefuehrten Untersuchungen verfolgen nicht das Interesse, eine Entscheidung pro oder contra herbeizufuehren, obgleich wir unsere skeptische Position nicht verhehlen wollen. U.E. kann die wissenschaftstheoretische Qualitaet einer solchen Entscheidung aber gehoben werden, wenn nicht allein die moeglichen, sondern auch die tatsaechlichen qualitativen und quantitativen Konsequenzen des Bedeutungsfehlers sichtbar gemacht worden sind. Wir behalten insoweit die bereits

bisher eingehaltene kritische Perspektive bei, nicht zuletzt, weil auch ueber metatheoretische und methodologische Alternativen nur vergleichend geurteilt werden kann - eine Aufgabe, die ohne volle Kenntnis der Eigenschaften dessen, was zu vergleichen ist, rational nicht loesbar waere.

# 3, Der Einfluss des Bedeutungsfehlers auf die Ergebnisse von Untersuchungen mit Rating-Daten

## 3,1, Fragestellung und Anlage der Studie

### 3,1,1, Die Vorgehensweise und ihre Problematik

Um die qualitativen und quantitativen Auswirkungen des Bedeutungsfehlers untersuchen zu koennen, muessen sowohl Informationen vorliegen, welche die von ihm ausgehenden Effekte enthalten, als auch solche, die - indirekt oder direkt - das Ausmass dieser Effekte erfassen, Dementsprechend bestand die Aufgabe zunaechst darin, ein in der Unterrichtsforschung "uebliches" Design zu ermitteln, Diese Frage laesst sich differenzieren gemaess den drei wichtigsten Komponenten; Bestimmung von Items, Erhebungsmodus und Analyseverfahren,

Fuer die Suche nach "gaengigen" Kategorien- bzw, Ratingaspekten wurde eine Itemkartei angelegt, in die alle verschiedenen, in deutschsprachigen Untersuchungen mit informellen Messinstrumenten aufgefundenen Formulierungen aufgenommen wurden, Die - vorlaeufige - Beschraenkung auf muttersprachliche Terminologie war erforderlich, weil fuer die geplante empirische Erfassung der Bedeutungsbeziehungen nur deutsch sprechende Vpn, zur Verfuegung standen, Mit beruecksichtigt wurden in einem zweiten Schritt auch Uebersetzungen von hauptsaechlich amerikanischen Arbeiten und schliesslich in einem dritten Schritt fremdsprachige Originaltexte,<35> Je nachdem wie streng das Kriterium der Uebereinstimmung gefasst wird, erhaelt man so etwa 150 bis 400 Items, die sich durch zwei weitere Einschraenkungen (Lehrverhalten/Hochschule) auf knapp 100 verwendbare reduzieren lassen,<36>

Von den Erhebungsverfahren konnten die technisch aufwendigen (time-/event-)Samplingmethoden ausgeschlossen werden, da diejenigen Merkmale, in denen sie sich von den einfacher zu handhabenden "Gesamterfassungen" unterscheiden, keinen systematischen Einfluss auf den Bedeutungsfehler ausueben (vgl, 2,1,2,), Ohnehin beruht die Mehrzahl der von uns gepruefften Untersuchungen auf ex-post-ratings, In der Frage der Neutralitaet und Geuebtheit der Beobachter findet man zwar, wie erwaehnt, haeufig den Hinweis auf die Wichtigkeit beider Gesichtspunkte; die Arbeiten zum Lehrverhalten selbst beruhen jedoch oft auf Angaben von ungeschulten "Betroffenen", Mueller-Wolf (1977, 103) findet keine bedeutsamen Reliabilitaetsunterschiede zwischen ihnen und den geschulten "Neutralen",

Als Analyseverfahren schliesslich werden meist multivariate Prozeduren (Korrelations-, Cluster-, Faktorenanalyse, seltener Varianzanalyse) gewaehlt, zur Untersuchung von Kausalrelationen oft auch univariate Methoden.

Der Versuch, mit diesen Ueberlegungen ein "Standard"-Design zu rekonstruieren, beruht auf der Absicht, eine moeglichst breite Uebertragbarkeit der Ergebnisse zu erreichen. Damit ist allerdings zugleich impliziert, dass die eingeschlagene Vorgehensweise von vornherein auch jene Unzulaenglichkeiten umschliesst, die aus methodischen Gruenden vermieden werden muessten. So verzichten praktisch alle in unserer Literaturanalyse enthaltenen Arbeiten auf die Mitteilung und d.h. wahrscheinlich auch auf die Pruefung von Verteilungskennwerten ((mehrdimensionale) Normalverteilung der (gruppierten) Daten, Varianzhomogenitaet), die fuer die Anwendbarkeit bestimmter multivariater Prozeduren und parametrischer Tests der Hypothesen entscheidend sind. Ausserdem werden ein und dieselben Daten oft zur Pruefung verschiedener Hypothesen mehrfach analysiert. Beide Maengel tragen, vor allem wenn sie kombiniert auftreten, zu u.U. betraechtlichen Erhoehungen des Typ-I-Fehlers (alpha-Fehler) bei. Auch das Problem des Skalenniveaus der Daten wird in den vorliegenden empirischen Untersuchungen unter dem Aspekt der fuer die Interpretation zulaessigen statistischen Modelle selten beachtet (vgl. dazu z.B. Allerbeck 1978), obwohl gerade Ratingdaten je nach Konstruktion des Instruments Nominal-, Ordinal- oder Intervallskalenniveau aufweisen koennen.

Die Strategie der vorliegenden Studie besteht mithin darin, zunaechst eine "gaengige" empirische Untersuchung vollstaendig durchzufuehren und dann in einem zweiten Schritt eine Re-Analyse der um den Bedeutungsfehler korrigierten Daten vorzunehmen, um schliesslich im dritten Schritt den Vergleich und die Diskussion der Unterschiede zwischen beiden Vorgehensweisen zu ermoeglichen. Damit endlich auch eine umrisshafte Vorstellung davon entwickelt werden kann, welche Konsequenzen aus der Beachtung des Bedeutungsproblems fuer die Beurteilung bereits vorliegender Untersuchungsergebnisse erwachsen koennen, lehnen wir uns in der Konzeption des "konventionellen" Teils an eine Arbeit von Mueller-Wolf (1977) zum "Lehrverhalten von Hochschullehrern" an, die sich neben den oben diskutierten Merkmalen durch die Wiedergabe vieler Einzelergebnisse auszeichnet und damit Vergleichbarkeit in einigen wesentlichen Punkten sichert.

### 3.1.2. Die Untersuchungsbereiche

Aus dem "Fragebogen fuer Vorlesungen" (Mueller-Wolf 1977, 197-200) entnahmen wir zunaechst die Globalurteile (Item Nr. 5, 10, 16, 21, 22, 24, 26, 30), die in etwa mit der Begrifflichkeit von Stilkonzepten uebereinstimmen (z.B. Nr. 16 "sozialintegrativ"), und reduzierten sie um diejenigen, von denen von vornherein hohe semantische Uebereinstimmung mit einem der beibehaltenen zu erwarten war (z.B. beibehalten: Nr. 10 "autokratisch" - gestrichen; Nr. 21 "autoritaer"; beibehalten; Nr. 16 "sozialintegrativ" - gestrichen; Nr. 5 "demokratisch"); ausserdem waehlten wir bei Mehrfachbezeichnungen der Skalen (z.B. Nr. 22 "optimistisch, emotional warm") jeweils die erste oder, wenn eine der folgenden nach unserem Vorverstaendnis spezieller oder hochschulspezifischer war, die jeweils naechste. Wir ergaenzten diese Gruppe um eine weitere Dimension, die u.a. bei Tausch und Tausch (1970) und in der von ihnen ausgewerteten amerikanischen Literatur benutzt wurde (vgl. Tab. 2, Ziff. I.6.).

Eine zweite Itemgruppe besteht aus den Urteilen zu einzelnen Dimensionen des Lehrverhaltens und der Lehrerpersoenlichkeit. Wir bildeten sie nach den gleichen Regeln wie oben aus dem Fragebogen von Mueller-Wolf. Da wir aus Gruenden, die in der Bedeutungsuntersuchung liegen und weiter unten dargestellt werden, eine zahlenmaessige Beschraenkung der Items anstreben mussten, fuehrten wir als weiteres Ausscheidungskriterium die von Mueller-Wolf und Fittkau (1971, 170) aus einer Voruntersuchung berichteten Uebereinstimmungswerte der Beurteiler ein. Die oberhalb des cut-off-Punktes (.64; naechster Wert: .60) liegenden Items (ohne Globalurteile) uebernahmen wir;<37> sie sind in Tab. 2, Ziff. II aufgelistet.

Die Trennung nach Partial- und Globalurteilen erscheint in vielen Untersuchungen in methodischer Hinsicht als Unterscheidung von Items und Faktoren. Da die letzteren i.d.R. aus orthogonalen Rotationen stammen, muesste erwartet werden, dass Bedeutungsueberschneidungen zwischen ihnen kein ueberzufaelliges Ausmass annehmen, es sei denn, die sprachlichen Rekonstruktionen (Deutungen) der Faktoren als unabhaengige Merkmale waeren im Sprachgebrauch der Rater nicht in dieser Weise repraesentiert. Diese Befuerchtung mag beispielsweise fuer die Konzepte (Faktoren) "emotional warm" und "nachsichtig" zutreffen, die - nach Konstruktion - die Pole unabhaengiger Dimensionen kennzeichnen (vgl. z.B. Tausch/Tausch 1970, 172), aber in der Beobachtungsrealitaet u.U. doch ueber teilweise identischen Merkmalen gebildet werden.<38>

Dieser Problemzusammenhang gab den Anlass, in die Untersuchung eine dritte Gruppe von Items aufzunehmen, die nach Erwartung semantisch staerker getrennt sein muessten. Sie beziehen sich ausschliesslich auf beobachtbares Verhalten (z.B. Schwankungen der Stimmlage), ohne jedoch ueber alltagspraktisch uebliche Elementarisierungen hinauszugehen (z.B. nicht: Lippen-Mundwinkel-Konstellationen). Solche Haeufigkeitsbeobachtungen sollten gemaess ihrer Konstruktion ueber Zaehlungen (bzw. Messungen) prinzipiell intersubjektiv kontrollierbar und "mikro-skopisch" rekonstruierbar sein. Sie sind in Tab. 2 unter Ziff. III zusammengestellt.

### 3.1.3. Methoden und Probleme der empirischen Erfassung des Bedeutungsfehlers

Das Phaenomen der Bedeutungsueberschneidung wurde in Kap. 2 ausfuehrlich eroertert. Hier stellt sich jetzt die Frage, auf welche Weise es empirisch erfasst werden kann. Da es sich aus intra- und interpersonellen Differenzen zusammensetzt, muessen beide gesondert bedacht werden. Wir wenden uns zunaechst den intrapersonellen Differenzen zu.

Jede Anwendung eines Wortpaares auf einen bestimmten Sachverhalt kann - statistisch gesehen - als Element einer abzaehlbar endlich maechtigen Ereignismenge gedeutet werden, die sich aus der Zahl der Anwendungen innerhalb einer begrenzten Zeit (maximal: individuelle Lebensdauer; hier: Dauer der intrapersonellen Bedeutungskonstanz beider Begriffe) zusammensetzt. Bei vorausgesetzter Dreidimensionalitaet der Wortbedeutung (vgl. Kap. 2.1.2.) enthaelt der Stichprobenraum Elemente geordneter Tripel, deren Glieder die drei Relationen zwischen den beiden individuellen Wortkonzepten sind. Die Aufgabe bestuende dann darin, bei systematisch variiertem Reizmaterial (Situationen) Wortanwendungsstichproben zu ziehen und ueber die Parameterverteilungen den gemeinsamen Bedeutungsraum der beiden bzw. der in Frage kommenden Termini zu ermitteln.

Mit dieser Konstruktion ist das Problem der Sprachanwendung als synthetischer Zusammenhang gedeutet und entspricht insofern einer sprachpsychologischen Aufgabenstellung. Es besteht allerdings auch die Moeglichkeit der Annahme einer analytischen Beziehung zwischen Sprache und Situation; sie wird den Verhaeltnissen in der Wissenschaft besser gerecht, indem sie die Relation zwi-

schen Wort und Realitaetsmerkmalen und damit auch zwischen Wort und Wort als explizite definiert rekonstruiert. Fuer die empirische Erfassung des individuellen Sprachgebrauchs ergibt sich in dieser Sicht eine wesentliche Vereinfachung: es genuegt, die vollstaendigen Definitionen abzufragen und sie untereinander zu vergleichen. Obwohl in der Durchfuehrung von Unterrichtsbeobachtung solche vergleichbaren Definitionen selbst bei Beobachterschulung praktisch nie eingefuehrt werden, waere dies doch prinzipiell moeglich - auch dann, wenn in die Bedeutung Wertungen eingehen sollen.

Im Rahmen unserer methodenkritischen Fragestellung geht es zunaechst nicht darum, Loesungen fuer das gekennzeichnete Problem zu entwerfen (vgl. dazu Kap. 4). Vielmehr bemuehen wir uns herauszufinden, in welchem Umfang Bedeutungsueberschneidungen auftreten und wie sie sich auf die Untersuchungsergebnisse auswirken. Fuer beide Komplexe uebernimmt unsere Arbeit aber lediglich eine hypothesengenerierende Funktion. Wir deuten den Fall des Einsatzes von ungeschulten Beobachtern als den der Anwendung individueller impliziter Definitionen, obgleich gerade hier auch ein sprachpsychologischer Ansatz begruendbar waere. Wir nuetzen damit zugleich den Vorteil, auf eine einzige Befragung ueber die individuelle Relation zwischen je einem Wortpaar als zulaengliche Basisinformation zurueckgreifen zu duerfen und vermeiden so das Problem der personenbezogenen Sprachanwendungsstichprobe.

Die untersuchungspraktische Loesung dieser Aufgabe besteht darin, nicht etwa die impliziten Definitionen von den Beobachtern explizit machen zu lassen - was ebenfalls eine moegliche und zugleich ergaenzende Strategie waere - , sondern die Aehnlichkeit zweier Wortbedeutungen quantitativ zu erheben. Jedem Beobachter wurden alle moeglichen Wortpaare vorgelegt, mit der Bitte anzugeben, in welchem Masse ihre Bedeutungen miteinander uebereinstimmen (vgl. Skalenbeispiel in Tab. 2).<39>

Dass mit dieser Versuchsanordnung die gestellte Frage nicht praezise beantwortet werden kann, ist offensichtlich. Es sind hauptsaechlich drei Quellen fuer die Ungenauigkeit der Ergebnisse zu benennen: zum einen bleibt verborgen, in welchen und wievielen Dimensionen sowie mit welcher "internen Metrik" (Was heisst: "sehr aehnlich"?) der Vergleich erfolgt; zum zweiten koennen von ungeschulten Beobachtern in der forced-choice-Situation fuer ungelaeufige Ausdruecke zufaellige Angaben gemacht werden, was allerdings nicht notwendig so sein muss, da auch einiges fuer eine semantische Zuordnung des fremden Wortes zu einem bereits bekannten spricht; zum dritten

ist nicht gewaehrleistet, dass der eigene Sprachgebrauch - auch auf Befragen hin - kognitiv vollstaendig kontrolliert wird. Man kann nicht ausschliessen, dass ein Sprecher trotz faktischer Bedeutungsueberschneidung glaubt, er gebrauche zwei Ausdruecke unabhaengig.

Auch eine vorsichtige Beurteilung wird zu dem Ergebnis gelangen, dass solche Fehlerquellen im Rahmen der Ueberpruefung von Verdachtsmomenten in Kauf genommen werden koennen. Wir tragen im Rahmen der Datenauswertung diesem Problem Rechnung, indem wir - wie unten gezeigt wird (3.4.2.) - unterschiedlich rigorose Deutungen der Befragungsergebnisse durchspielen.

Die empirische Bestimmung der interpersonellen Bedeutungsueberschneidungen, also des Ausmasses, in dem zwei Beobachter sich auf Verschiedenes beziehen, wenn sie den gleichen Ausdruck verwenden (et vice versa) wirft erheblich schwierigere praktische Probleme auf. Sie resultieren aus der Kluft, die das Bewusstsein eines Individuums von allen anderen trennt. Sieht man von der epistemologischen Dimension dieses Problems ab, so koennen zunaechst auf der sprachphilosophisch-konzeptuellen Ebene zwei Hauptklassen von Auffassungen ueber diese "Kluft" unterschieden werden, die fuer die Deutung der empirisch-praktischen Schwierigkeiten und die Begruendung ihrer Loesungen wichtig sind. Wir kennzeichnen diese beiden Klassen hier nur schematisch und weder mit dem Anspruch auf Vollstaendigkeit hinsichtlich der sprachphilosophisch moeglichen Konzeptionen noch auf ausreichende Differenziertheit hinsichtlich sprachphilosophisch deskriptiver Kategorien.

Unter dem fuer unsere Fragestellung konstitutiven Aspekt laesst sich die erste Klasse dadurch kennzeichnen, dass sie die erwaehnte Kluft zwischen "Bewusstseinen" als prinzipiell unueberbrueckbar betrachtet in dem Sinne, dass jedes Individuum letztlich seine eigene, von "aussen" nicht zugaengliche "Privatsprache" hat (vgl. dazu und zum folgenden z.B. die Darstellung bei Schnelle 1973). Mit diesen im Extrem solipsistischen Positionen kontrastiert die zweite Klasse von Auffassungen, wonach wir alle an der ueberindividuell existierenden Sprache teilhaben, Glieder einer transzendentalen Sprachgemeinschaft sind und insofern zwar akzidentiell und faktisch fehlerhaften Gebrauch von der Sprache machen koennen, aber prinzipiell und kontrafaktisch totale Verstaendigung, und das heisst: semantisch eindeutige Kommunikation moeglich ist.

Wir werden weiter unten (3.3. und 3.4.) auf die theoretischen und methodischen Aspekte noch naeher eingehen

und zeigen, in welcher Weise diese sprachphilosophische Position das empirische Design und die Auswahl statistischer Modelle beeinflusst. An dieser Stelle kommt es nur darauf an, in Abhebung von der soeben gekennzeichneten Unterscheidung auf ein weiteres davon unabhaengig diskutierbares Problem hinzuweisen. Es besteht darin, dass die Genese der Relation zwischen Designatum, Sprachzeichen und Sprecher verschieden gedeutet werden kann. Eine der Idee nach streng rationale Rekonstruktion hebt auf den analytischen Charakter ab und deutet sie als gestaltungs- und entscheidungsbeduerftige Beziehung. Die empirische Untersuchung muesste demzufolge darauf abzielen, die individuumspezifischen Definitionen der Beobachtungswoerter explizit zu machen. Dagegen kann auch eine erfahrungsgebunden-lernpsychologische, also synthetische Genese der Relation betont werden. Fuer diesen Fall waere etwa ueber die Variation von kontrolliertem Reizmaterial der Bedeutungsumfang von Ausdruecken bei verschiedenen Personen zu erschliessen und zu vergleichen, wie oben bereits dargestellt.

Wir haben angesichts der ungeklaerten Konzeptions- und Methodenprobleme auf einen Erfassungsversuch interindividueller Unterschiede gaenzlich verzichtet, obgleich eine vollstaendige Analyse von Beobachtungsdaten auch diese Dimension soweit wie moeglich ausleuchten muesste. Der Fehler, der durch die Gleichsetzung von Daten entsteht, die von verschiedenen Personen bezueglich ein und desselben fuer sie u.U. bedeutungsverschiedenen Wortes "produziert" wird, ist demzufolge auch noch in den bereinigten Ergebnissen enthalten, die unten analysiert werden. In diesem Punkt unterscheidet sich die Datenbasis unserer Untersuchung nicht von derjenigen anderer Ratinguntersuchungen. Wir verfolgen dieses Problem hier allerdings auch deshalb nicht weiter, weil wir glauben, dass die Loesung des Bedeutungsproblems in einer anderen Richtung zu suchen ist.

### 3,1,4, Das Design

Die zu erfassende Information umfasst drei Teilbereiche, die in sich weiter gegliedert sind (vgl, zu (2,) und (3,) auch Tab, 2):

| | Items |
|---|---|
| 1, Merkmale der Vpn, | |
| a) Identifikation und Personaldaten (2 Teilerhebungen mit je 5 Items) | 10 |
| +b) Fragebogen zur Leistungsmotivation (nach Tent 1963) | 22 |
| +c) Fragebogen zur Erfassung von autoritaeren Einstellungen (F-Skala nach Roghmann 1966) | 22 |
| +d) Fragebogen zur Erfassung von Studienerwartungen | 8 |
| 2, Wortaehnlichkeiten | |
| a) Globalurteile untereinander | 15 |
| b) Partialurteile untereinander | 36 |
| c) Globalurteile / Partialurteile | 54 |
| d) Globalurteile / Haeufigkeitsbeobachtungen | 72 |
| e) Partialurteile / Haeufigkeitsbeobachtungen | 108 |
| 3, Unterrichtsbeobachtung (Wortanwendung) | |
| a) Globalurteile | 6 |
| b) Partialurteile | 9 |
| c) Haeufigkeitsbeobachtungen | 12 |
| | --- |
| | 374 |
| | === |

+ Die Daten zu (1b,c,d) werden im Rahmen dieser Arbeit nicht ausgewertet,

### 3-1 Untersuchungsbereiche

Die verhaeltnismaessig grosse Itemzahl machte eine Teilung der Datenerhebung auf zwei Termine erforderlich (Dez, 1977 und Jan, 1978), Einbezogen waren Studenten der Wirtschaftspaedagogik im ersten bis fuenften Semester, die an vier Pflichtveranstaltungen des erziehungswissenschaftlichen Grundstudiums teilnahmen, Von den insgesamt 155 Teilnehmern konnten 125 vollstaendige und verwertbare Datensaetze erhoben werden,

Um Abfolgeeffekte pruefen zu koennen, wurden die Teilbereiche in den vier Probandengruppen in unterschiedlicher Weise angeordnet, Die Beobachtungsaufgabe wurde stets am Schluss gestellt, damit die Aehnlichkeitsangaben nicht durch den Logikfehler beeinflusst werden sollten, Zum

Zweck der Vergleichbarkeit der Beobachtungen zwischen den verschiedenen Untersuchungsgruppen waehlten wir Videoaufzeichnungen von Lehrveranstaltungen, die fuer die Studenten unbekannte Hochschullehrer abgehalten hatten. Je nach Gruppengroesse wurden mehrere Monitore zur Wiedergabe aufgestellt, so dass jede Vpn. den Bildschirm gut einsehen konnte. Ein Einfuehrungstext wurde vom Vl. verlesen und auf Anfragen naeher erlaeutert.<40> Nach dem Eintragen der Angaben zur Person trug der Vl. auch den Anleitungstext vor und beantwortete ggf. Fragen. Danach arbeiteten die Vpn. selbstaendig bis zum Ende des letzten Wortaehnlichkeitsteils. Nachdem alle Teilnehmer bis zu diesem Punkt gelangt waren, wurde der Film eingespielt und im Anschluss daran die Bearbeitung des jeweiligen Beobachtungsteils vorgenommen. Die beiden Abschnitte der Erhebung waren unterschiedlich umfangreich (60 Min./90 Min.), da die Untersuchungsteilbereiche als Ganzheiten erhalten bleiben sollten. Ausserdem wurden, um Interferenzen zwischen den Teilbereichen auszuschalten, Abfolgerestriktionen beachtet:

- Zwischen zwei Wortaehnlichkeitsteile ist moeglichst ein Personenmerkmal-Fragebogen einzufuegen!
- Wortgruppe (i) "innerhalb" und Wortgruppe (i/j) "zwischen" duerfen nicht nacheinander stehen!
- Wortanwendung von Global- und Partialurteilen duerfen nicht aufeinander folgen!

Damit ergab sich der folgende Versuchsablaufplan;

| Gruppe / Termin | I n=25 | II n=21 | III n=42 | IV n=37 |
|---|---|---|---|---|
| Dez, '77 | Pers,daten H - P L,mot, P - P G - G F-Skala Beob,: H,G =226 Items | Pers,daten H - P F-Skala G - G P - P L,mot, Beob,: H,G | wie I Jan, '78 | wie II Jan, '78 |
| Jan,'78 | Pers,daten G - P Stud,erw, G - H Beob,: P =148 Items | Pers,daten G - H Stud,erw, G - P Beob,: P | wie I Dez, '77 | wie II Dez, '77 |

Es bedeuten;
G; Globalurteile
P; Partialurteile
H; Haeufigkeits-beobachtungen

Zu den uebrigen Abkuerzungen vgl, oben (1a - d),

## 3-2 Untersuchungsplan

Die Antworten aus den Fragebogen zur Leistungsmotivation und zur F-Skala wurden nach den Angaben der Herausgeber (Tent 1963, Roghmann 1966) zu einem Score verrechnet; entsprechendes gilt fuer die Studienerwartungen,

Die Antworten zu den Wortaehnlichkeiten lassen sich in einer Dreiecksmatrix, die Beobachtungsangaben in einem Vektor anordnen, Der gesamte Datensatz pro Person erhaelt somit die folgende Struktur;

| Aehnlichkeiten<1> | | Globalurteile 1 2 3 4 5 6 | Partialurteile 7 8 9 10 11 12 13 14 15 | Haeufigkeitsbeobachtgn. 16 17 18 19 20 21 22 23 24 25 26 27 | Beob. daten zu... |
|---|---|---|---|---|---|
| Global-urteile | 1 | 9<2> | | | 1<1> |
| | 2 | 9 G - G | G - P | G - H | 2 |
| | 3 | 9 | | | 3 |
| | 4 | 9 | | | 4 |
| | 5 | 9 | | | 5 |
| | 6 | 9 | | | 6 |
| Partial-urteile | 7 | | 9 | | 7 |
| | 8 | | 9 | | 8 |
| | 9 | | 9 P - P | P - H | 9 |
| | 10 | | 9 | | 10 |
| | 11 | | 9 | | 11 |
| | 12 | | 9 | | 12 |
| | 13 | | 9 | | 13 |
| | 14 | | 9 | | 14 |
| | 15 | | 9 | | 15 |
| Haeufigkeits-beob. | 16 | | | 9 0 0 0 0 0 0 0 0 0 0<2>0 | 16 |
| | 17 | | | 9 0 0 0 0 0 0 0 0 0 0 0 | 17 |
| | 18 | | | 9 0 0 0 0 0 0 0 0 0 0 | 18 |
| | 19 | | | 9 0 0 0 0 0 0 0 0 0 | 19 |
| | 20 | | | 9 0 0 0 0 0 0 0 0 | 20 |
| | 21 | | | 9 0 0 0 0 0 0 0 | 21 |
| | 22 | | | 9 0 0 0 0 0 0 | 22 |
| | 23 | | | 9 0 0 0 0 0 | 23 |
| | 24 | | | 9 0 0 0 0 | 24 |
| | 25 | | | 9 0 0 0 | 25 |
| | 26 | | | 9 0 0 | 26 |
| | 27 | | | 9 0 | 27 |

<1> Zur Zuordnung von Kennziffer und Wort vgl. Tab. 2.
<2> Die Ziffer "9" in der Diagonalen steht gemaess der Skalenmetrik fuer Bedeutungsgleichheit, die "0" fuer die unterstellte, nicht erhobene Ueberschneidungsfreiheit des Wortpaares.

3-3 Datenmatrix

## 3.2. Formale Datenanalyse

### 3.2.1. Ergebnisstabilitaet und Abfolgeeffekte

Die Stabilitaetspruefung erfolgt entsprechend den Analysezielen in differenzierter Weise. In Bezug auf die Beobachtungsdaten (Rating des Lehrverhaltens) lautet die Frage, ob die ihnen inhaerente Faktorenstruktur stabil ist, d.h. ob die Kovarianzmatrizen (fuer die ja das faktorenanalytische Modell angepasst wird) von zufaellig gebildeten Teilgruppen homogen sind. Die Aehnlichkeitsangaben, die fuer die Modifikation der Beobachtungsurteile herangezogen werden sollen, koennen als stabil gelten, wenn ebenfalls zufaellig gebildete Teilgruppen in ihren Angaben nicht systematisch voneinander abweichen. Die Untersuchung der Beobachtungsdaten erfolgt getrennt nach Inhaltsbereichen: Global- und Partialurteile einerseits sowie die Haeufigkeitsbeobachtungen andererseits werden je fuer sich nach Halbierung der Beobachtergruppe faktorenanalysiert und die erhaltenen Ladungsmuster einer Rotation auf maximale Aehnlichkeit unterworfen.

Wie aus den graphischen Darstellungen fuer die formale Bestimmung der zu extrahierenden Faktorenzahl hervorgeht (Abb. 3-4: Bestimmung der Faktorenzahl fuer Urteilsdaten; 3-5: Bestimmung der Faktorenzahl fuer Haeufigkeitsdaten), koennten bei den Urteilen je nach Kriterium 5 (lambda >= 5% der Gesamtvarianz), 4 (lambda > 1.) oder 3 (Scree-Test) Faktoren bestimmt werden (vgl. Ueberla 1971, 126). Angesichts der Ueberschreitung der Gesamtkommunalitaet durch den dritten Faktor erscheint jedoch eine Beschraenkung auf zwei Dimensionen erforderlich.<41> Entsprechendes gilt fuer die 1-Faktor-Loesung der Haeufigkeitsangaben, fuer die aus Gruenden, die wir spaeter diskutieren werden, von vornherein ein hoher Einzelrestanteil an der Gesamtvarianz zu erwarten war.

Die Berechnung der Aehnlichkeitskoeffizienten (Tucker's Phi; vgl. Revenstorf 1976, 251) fuer die nach dem Varimax-Kriterium rotierten Faktorenloesungen (Kommunalitaetenschaetzung: 1.0) fuehrt zu folgenden Ergebnissen:

| Faktor | Aehnlichkeitskoeffizient Phi: Urteilsdaten | Haeufigkeitsdaten |
|---|---|---|
| 1 | .941 | .952 |
| 2 | .917 | -- |

3-6 Anpassung der Ladungsmuster nach split-half-Faktorenanalysen

Die Hoehe der Koeffizienten (0<phi<1) spricht nicht gegen die Annahme, dass die durch die Unterteilung der Gesamtgruppen hervorgerufenen Unterschiede im Zufallsbereich liegen und damit von der Stabilitaet der Beobachtungsdaten ausgegangen werden kann. Der weitgehend parallele Verlauf der Eigenwertprofile stuetzt diese These.

Zur Untersuchung der Aehnlichkeitsangaben dienen Korrelationsrechnungen, die zwischen den vier Quartilen fuer jeden Schaetzbereich getrennt vorgenommen werden (G-G, P-P, G-P, G-H, P-H; vgl. 3-3). In jeder Analyse koennen fuenf kanonische Faktoren ermittelt werden, die zu hoch signifikanten Korrelationen fuehren, deren Werte fast durchweg bei R > .95 liegen. Lediglich die Aehnlichkeitsschaetzungen zwischen den Partialurteilen (P) und den Haeufigkeitsbeobachtungen (H), bei denen die kanonischen Korrelationskoeffizienten in den Bereich .97>R>.80 (p<=.001) fallen, fuegen sich nicht nahtlos in dieses Bild. Einen Anlass zur Verwerfung der Stabilitaetsannahme liefern sie jedoch nicht, da auch schon eine geringere Zahl signifikanter kanonischer Korrelationskoeffizienten von R > .90 ausreichend gewesen waere.

Der Frage nach Effekten, die durch die Reihenfolge der in der Erhebung zu bearbeitenden Teilgebiete hervorgerufen werden koennen, wird mit Hilfe der multivariaten Varianzanalyse nachgegangen. Fuer jeden Inhaltsabschnitt der Datenmatrix (vgl. 3-3) wurde eine Analyse ueber dem Merkmal Versuchsgruppenzugehoerigkeit durchgefuehrt; es ergaben sich nirgends signifikante Unterschiede zwischen den Gruppen.

### 3.2.2. Reliabilitaet der Beobachtungsdaten

Zuerst ist darauf hinzuweisen, dass eine Reliabilitaetsanalyse der Aehnlichkeitsangaben inhaltlich nicht sinnvoll sein kann. Diese Angaben beziehen sich ja nicht auf

ein und denselben Sachverhalt, den es zu erfassen ("messen") gilt, sondern - bei ungeschulten Beobachtern - auf genau so viele Sachverhalte, wie es Beobachter gibt; jeder von ihnen gibt Auskunft ueber seinen (eigenen) Sprachgebrauch. Geschulte Beobachter muessten hier einheitliche Aussagen machen, weil ihre Beobachtungssprache hinsichtlich der vorgegebenen Begriffe standardisiert wurde.

Wir berechnen die Reliabilitaet mittels des Intraklassenkoeffizienten auf varianzanalytischer Basis, weil er es erlaubt, neben dem Zufallsfehler auch den systematischen Fehler als reduzierenden Faktor miteinzubeziehen, und insofern strenger ist (vgl. Hartmann 1977, 106).<42> Es ergibt sich nach der von Horst (vgl. Langer/Schulz v.Thun 1974, 88) vorgeschlagenen Formel r(nn) =.989. Dieser Wert ist hoch signifikant von Null verschieden und bewegt sich in einem Bereich, der in der Literatur allgemein als sehr zufriedenstellend bezeichnet wird. Einschraenkend ist allerdings zu bemerken, dass die verhaeltnismaessig grosse Beobachterzahl zu diesem Ergebnis beitraegt. Die durchschnittliche Reliabilitaet fuer einen Rater betraegt (nach Ebel, vgl. Guilford 1954, 395) r(11) = .4599. Aber auch dieser Wert faellt in den Rahmen dessen, was fuer ausreichend gehalten wird (vgl. Langer/Schulz v.Thun 1974, 93).

Gemaess den ueblichen Interpretationsanleitungen duerfte aufgrund der vorliegenden Ergebnisse gesagt werden, dass die ermittelten Werte " hinsichtlich Niveau, Abstand und Relation zu 99 % unabhaengig von den Spezifitaeten des Messvorganges" sind (Langer/Schulz v.Thun 1974, 89). Im Lichte unserer frueheren Ueberlegungen (vgl. 2.2.) erscheint diese Deutung - bei den gegebenen Skalen - jedoch als problematisch. Zum einen muessen naemlich die von einem einzelnen Rater angegebenen Werte, soweit sie sich auf die gleichen Merkmalsmengen beziehen, selbst als Messwiederholungen betrachtet werden, die, wenn sie nicht als solche behandelt werden, zu einer Ueberschaetzung der "wahren" Reliabilitaet fuehren. Zum anderen ist aus den Berechnungsformeln leicht erkennbar, dass die Hoehe des Gesamt-Reliabilitaetskoeffizienten (r(nn)) von steigenden Beobachterzahlen staerker "profitiert" als seine durchschnittliche Hoehe pro Rater (r(11)) (zur Symbolik vgl. Anm. 42);

$$[1] \quad r(11) = \frac{SAQ(o) - SAQ(e)/(n-1)}{SAQ(o) + SAQ(e)} \quad \text{und}$$

$$[2] \quad r(nn) = 1 - \frac{SAQ(s) + SAQ(e)}{SAQ(o)\ (n-1)} \ ,$$

Fuer groesser werdendes n (im Rahmen empirischer Daten) geht r(nn) schneller gegen 1, als r(11). Gerade bei hohen Beobachterzahlen, wie auch in der vorliegenden Untersuchung, ist es daher erforderlich, den Wert von r(11) zu beobachten. Wir werden weiter unten zeigen, dass bei den Daten, die um die Bedeutungsueberschneidungen bereinigt wurden, die durchschnittliche Reliabilitaet pro Rater "unter dem Deckmantel" grosser Raterzahlen, also hoher Gesamt-Reliabilitaetskoeffizienten, z.T. erheblichen Schwankungen unterworfen ist.<43>

## 3,3, Das Bedeutungsproblem im Sprachkollektiv

### 3,3,1, Der Ansatz einer korrelationsstatistischen Untersuchung

In einem ersten Zugang untersuchen wir den Effekt der Bedeutungsueberschneidung auf der Basis der aggregierten Daten, d,h, wir unterstellen die "Existenz" einer - gewissermassen - ueberindividuellen Semantik, die sich in der Wortverwendung des einzelnen Sprechers realisiert, Diese Vorstellung laesst sich beispielsweise aus der (rein formalen!) Deutung der Idee einer "Sprachgemeinschaft" gewinnen, wie sie gegenwaertig etwa von Apel (z,B, 1972) vertreten wird, Damit soll der eingangs geaeusserten Absicht gefolgt werden, hinsichtlich der Auffassungen zu dieser Frage im Rahmen der Untersuchung moeglichst offen zu bleiben (vgl, auch oben 3,1,3,),

In Entsprechung zu der gekennzeichneten Klasse von Sprachkonzeptionen koennen die Mittelwerte aus den Aehnlichkeitsangaben, die von den Befragten zu jedem Wortpaar gemacht wurden, als gute Schaetzung fuer die zwischen den einzelnen Ausdruecken "tatsaechlich" bestehenden semantischen Relationen betrachtet werden, Es ist allerdings fraglich, ob das arithmetische Mittel die einzige adaequate Abbildung fuer den skizzierten inhaltlichen Zusammenhang darstellt (dieses Modell wird von Hofer (1969) gewaehlt, aber nicht weiter diskutiert), Seine Einbeziehung setzt voraus, dass die semantische Relation prinzipiell als kontinuierlich variierbar vorgestellt wird, Dies waere zwar gemaess der in (2,1,2,) entwickelten Konstruktion moeglich (Variation von Gewicht und Haeufigkeit in den gemeinsamen Merkmalen zweier Woerter), aber - bei der hier geuebten sprachtheoretischen Zurueckhaltung - nicht notwendig,

Plausible Argumente liessen sich auch dafuer vorbringen, dass der wahre Wert sich in der Sprechrealitaet bzw, in der Reflexion darueber "zeige", dass der Konsens in einer Sprechermehrheit zugleich die semantische Struktur zu "offenbaren" imstande sei, Fuer diese Vorstellung bildet der Modus das adaequate Modell, Und schliesslich waere begruendbar, dass - auch bei Kontinuierlichkeit der in den Variablen abgebildeten Eigenschaften - die extremen Abweichungen vom mittleren Wert bei der Bestimmung des wahren keinen Einfluss nehmen duerften ("Aussenseiter der Sprachgemeinschaft"), was fuer die Wahl des Medians spraeche, Wir werden die Datenpruefung daher fuer alle drei Faelle vornehmen,

Die Angaben liegen als Aehnlichkeitsschaetzungen vor. Wie bei allen derartigen Informationen ist die der individuellen Angabe zugrundeliegende Dimensionalitaet weder qualitativ noch quantitativ bekannt. Fuer die hier skizzierten sprachphilosophischen Konzeptionen hat dieser Einwand allerdings keine Bedeutung, da ja die gemeinsame Teilhabe an der ueberindividuell existierenden Sprache unterstellt wird. Erst wenn - weiter unten (3,4,1,) - diese Annahme aufgegeben wird, muss dieser Gesichtspunkt beachtet werden.

Fuer die Erhebung wurde eine neunstufige Skala verwendet, deren aufsteigende Werte fuer wachsende Aehnlichkeit zwischen Woertern eines Paares (W(i), W(k)) stehen. Aus der Sicht des statistischen Modells lassen sich die verschiedenen Mittelwerte als Korrelationen zwischen den Wortbedeutungen interpretieren, die von den Befragten ueber eine Vielzahl von "Messungen" ermittelt wurden.<44> Neben der Information ueber diesen ausschliesslich in der Sprache angelegten (rein analytischen), also vor aller Sprachanwendung bestehenden Zusammenhang (r(a)) verfuegen wir ueber Daten, die in der Beschreibungssituation erhoben wurden. Dieselben Woerter (W(j)) wurden dort benutzt, um auszudruecken, in welchem Masse das, was sie bedeuten, bei der zu beobachtenden Lehrperson realisiert ist. Die dabei ermittelten Werte koennen ebenfalls zueinander in Beziehung gesetzt werden unter der Frage, in welchem Umfang die mit den woertern bezeichneten Merkmale (bei eben dieser Lehrperson) zugleich auftreten. Auch hierfuer bietet sich im statistischen Modell die Korrelation als Mass an; sie wird aus den Angaben der Beobachter (3-3; Spalte "Beobachtungsdaten") als experimentell fundierter Koeffizient (r(ex)) ermittelt.

Soweit die zur Beschreibung der Beobachtungssituation verwendeten Woerter Bedeutungsueberschneidungen aufweisen, muessen die ihnen zugeordneten numerischen Werte (Z(j)) auch korrelative Beziehungen aufweisen, die mithin relativ zur empirischen Fragestellung als Artefakte zu bezeichnen sind. Um in Erfahrung zu bringen, in welchem Umfang "wahre" Korrelationen (r(w)), also empirische Beziehungen zwischen verschiedenen Merkmalen bestehen, muss der experimentell ermittelte Koeffizient (r(ex)) um den in ihm enthaltenen analytischen (scheinbaren, artifiziellen) Anteil (r(a)) bereinigt werden.

In der Sprache des mathematischen bzw. statistischen Modells ist damit die folgende Vorgehensweise impliziert. Zunaechst werden die Aehnlichkeitsmasse aus den 105 Paarvergleichen (W(i)-W(k)) der 15 Urteilsbegriffe (3-3; "Aehnlichkeiten") zu Korrelationskoeffizienten transfor-

miert. Ihre Werte laufen von 0 (keine Bedeutungsgemeinsamkeit) bis 9 (Bedeutungsgleichheit) und lassen sich - nach Multiplikation mit dem Faktor 10 - geometrisch als Groesse des Winkels alpha (i,k) zwischen zwei Einheitsvektoren v(i), v(k) auffassen (Abb. 3-7: Geometrische Darstellung der Aehnlichkeitsmasse): je groesser der Winkel, desto hoeher die Uebereinstimmung zwischen W(i) und W(k) (also z.B. zwischen "liberal" (Ziff. 3 in Tab. 2) und "entspannt" (12)). Bekanntlich kann der Korrelationskoeffizient geometrisch als Kosinus des Winkels zwischen den als Vektoren dargestellten Variablen gedeutet werden. Da sich in unserem Fall die Daten gegenlaeufig entwickeln und da

[3] sin alpha = cos(90 - alpha),

gilt fuer den vorliegenden Zusammenhang:

[4] r(a) = sin alpha.

In Analogie zum "klassischen" Modell der Varianzanalyse laesst sich die experimentell ermittelte Korrelation (r(ex)) zwischen den zu den Beobachtungswoertern gehoerenden Merkmalsbereichen als Funktion der wahren, der artifiziellen und der auf Zufallsfehler (error: "e") beruhenden Korrelation rekonstruieren:

[5] r(ex) = f(r(w), r(a), r(e)).

Entsprechend unserer Fragestellung kommt es lediglich darauf an, aus der experimentell gefundenen Korrelation die in ihr enthaltene, auf den Bedeutungsfehler zurueckzufuehrende Komponente zu eliminieren, um anschliessend zu fragen, ob die Restkorrelationen dieselbe Interpretation ermoeglichen wie das urspruengliche Muster. Wir koennen daher in Anpassung an unsere Fragestellung vereinfachend schreiben:

[6] r(w) = r(r(ex), r(a)),

Die Koeffizienten r(w) und r(a) enthalten Informationen ueber den Zusammenhang von ein und denselben Variablen Z(w/a,i) und Z(w/a,j). Dieser Zusammenhang wird als auf verschiedenen Ursachen beruhend betrachtet: r(w) und r(a) sind die Masse, die unter Bezugnahme auf zwei unabhaengige Ursachen bestimmt werden; r(ex) enthaelt die Zusammenfassung dieser Ursachen. Varianzanalytisch gewendet beruht r(ex) auf der gesamten gemeinsamen Varianz, r(a) auf derjenigen, die durch die Bedeutungsgemeinsamkeit zustandekommt. In diesem Fall verhalten sich die Quadrate der varianzanalytisch interpretierten Korrelationskoeffizienten ("rr") additiv (die Gleichheit

der Freiheitsgrade und Varianzhomogenitaet werden also unterstellt):

[7a] rr(w) = rr(ex) - rr(a).

Die um den Bedeutungsfehler bereinigte, also "echt empirische" Korrelation betraegt demnach

[7b] r(w) = +-{[rr(ex) - rr(a)]"hoch"1/2}.

Entsprechend dem varianzanalytischen Modell kann die Summe des Radikanden nicht kleiner als Null werden. Fuer den Fall, dass ein solches Ergebnis in den Daten auftritt, waere dies definitionsgemaess der in rr(a) enthaltenen Fehlervarianz zuzuschreiben. Ausserdem soll das Vorzeichen von r(ex) erhalten bleiben, da r(a) lediglich zur Korrektur von r(ex) herangezogen wird.

### 3.3.2. Der "Sprachschleier" ueber der Realitaet

In Tab. 3 sind die Koeffizienten, die sich aus unserer Datenerhebung ergeben, zusammengestellt. Verrechnet man sie in der angegebenen Weise miteinander und bestimmt die "wahren" Korrelationen, so zeigt sich, dass keine interpretierbaren Beziehungen mehr uebrig bleiben. Die auf der Basis der unterschiedlichen Mittelwerte gewonnenen Korrelationsmatrizen enthalten keine signifikant von Null verschiedenen Werte. Der nach der Naeherungsformel von Lawley (vgl. Ueberla 1971, 131) errechnete Chi-Quadrat-Wert muesste bei 95% Wahrscheinlichkeit (df=105) fuer die Geltung der Null-Verschiedenheit-Hypothese die kritische Schwelle von 129,63 uebersteigen. Tatsaechlich ergeben sich folgende Werte:

| Basis | Chi-Quadrat | Irrtumswahrsch. fuer H 1<+> |
|---|---|---|
| Arith. Mittel (MIT) | 81,82 | >,957 |
| Median (MED) | 6,36 | >>,999 |
| Modus (MOD) | 5,08 | >>,999 |

<+> Nach Naeherungsformel bei Graf, Henning, Stange (1966,292)

## 3-8 Signifikanzpruefung fuer Restkorrelationsmatrizen

Auf der Modellebene bedeutet dies, dass die berechnung einer Faktorenanalyse nicht mehr sinnvoll ist, da zwischen den Daten praktisch keine Kovarianz vorliegt.

Inhaltlich ergibt sich vorlaeufig, dass die in der Sprache angelegten Beziehungen sich gleichsam wie ein Schleier ueber die tatsaechlichen Gegebenheiten legen koennen und uns "Bilder" von ihr suggerieren, die faktisch nicht begruendet sind. Haette man die unkorrigierten Daten einer (Faktoren-)Analyse unterzogen, so waere man zu Interpretationen der (Unterrichts-)Realitaet gelangt, die in wesentlichen Teilen aus der Kenntnis der in den Beobachtungswoertern "angelegten" semantischen Struktur haette vorhergesagt werden koennen. Dies ergibt sich formal aus der betraechtlichen Hoehe der Aehnlichkeitskoeffizienten, die bei der Zielrotation der relevanten Faktorenmatrizen erreicht werden: fuer die 2-Faktoren-Loesung der unkorrigierten Beobachtungswerte und der Aehnlichkeitsschaetzungen (Basis: arithmetisches Mittel, Median, Modus) errechnet man Phi-Koeffizienten von .73 bis .75.

Die nach Abzug der analytisch verursachten Korrelationen verbliebenen geringen Zusammenhangswerte weisen darauf hin, dass die ihnen zugeordneten Merkmalsbereiche unkoordiniert variieren, also keinem ursaechlichen Zusammenhang unterliegen. Das gilt allerdings nur fuer diejenigen Merkmale, die nach Ausschluss der Ueberschneidungsbereiche zwischen den Woertern uebrig bleiben. Die Variation der einem Wortpaar gemeinsamen Merkmale wurde durch das angewandte Berechnungsverfahren aus der Analyse ausgeschlossen, d.h. es wurde mehr Information unterdrueckt, als unter der Fragestellung erforderlich gewesen waere.

Unter diesem Gesichtspunkt und unter Beruecksichtigung der Datenaggregation im Sprachbereich muss die durchgefuehrte Korrelationsanalyse als rigide bezeichnet werden. Eine Verfeinerung wuerde voraussetzen, dass nicht nur das Ausmass der Bedeutungsueberschneidung W(i)-W(k) bekannt waere, sondern auch Informationen darueber vorlaegen, bei welchen einzelnen Merkmalen sie auftritt. Es waere dann statistisch moeglich, aus jeder Beobachtungsvariablen Z(j) nur diejenige Variation herauszupartialisieren, die sie mit irgendeiner anderen gemeinsam hat. Damit bliebe auch die Aussage ueber gemeinsame Merkmale erhalten, die ja bei unserer Vorgehensweise ausgeschaltet wurde.

Ein solcher Ansatz wuerde einen erheblichen experimentellen Aufwand erforderlich machen. Doch spricht - unter Zugrundelegung der oben gekennzeichneten Klasse von Sprachtheorien - nichts gegen seine prinzipielle Realisierbarkeit; aus der Sicht dieser Auffassungen koennen Untersuchungen dazu nur als dringlich und wuenschenswert

bezeichnet werden. Immerhin deutet sich in den hier gewonnenen Ergebnissen bereits an, dass unser Verstaendnis von einer mit dem geschilderten Verfahren beschriebenen Realitaet moeglicherweise erheblichen Korrekturen unterworfen werden muss.

Wir verfolgen diese Frage im naechsten Abschnitt unter Voraussetzung einer anderen Klasse von Sprachtheorien, die - statistisch gesprochen - die Eigenschaften des Datenkoerpers besser "zur Geltung kommen lassen". Freilich waere dies kein zulaessiges Argument fuer die Anwendung anderer Modelle. Unter unserer methodenkritischen Frage ist die Aufgabenstellung jedoch eine andere; wir haben zu pruefen, welche Effekte von theoretisch begruendeten Korrekturen der Rohdaten ausgehen koennen und in welcher Weise sie die - inhaltliche - Interpretation, also Beschreibungen und Hypothesen bzw. Theorien beeinflussen.

## 3.4. Die Bedeutungsueberschneidung im individuellen Sprachraum

### 3.4.1. Probleme der Modellkonstruktion

Die Vorgehensweise in diesem zweiten Analyseschritt erfolgt unter Bezugnahme auf jene - weiter oben skizzierte (3.1.1) - Klasse von sprachphilosophischen Positionen, deren Gemeinsamkeit darin besteht, dass sie - formal gesprochen - nicht das Bestehen einer ausserindividuellen strukturierten Sprache postulieren oder - positiv gewendet - dass sie die Moeglichkeit einer individuumspezifischen semantischen Sprachstruktur zulassen (wie etwa bei Mach 1968 (zuerst 1905), 126 ff.; vgl. auch Putnam 1975, 139 ff.). Fuer sie erwaechst die Notwendigkeit, dem Problem der interpersonellen "Verstaendigungsmoeglichkeit" besondere Aufmerksamkeit angedeihen zu lassen. Unsere Fragestellung gehoert in diesem Zusammenhang zum Teilbereich der "Verstaendigung" unter Wissenschaftlern oder - aus einer anderen Perspektive - der Ermoeglichung eines intersubjektiven Zugangs zu den von einem Wissenschaftssubjekt sprachlich dokumentierten Erkenntnissen (i.w.S.), also deren 'Ver-Oeffentlichung'. "Speziell" ist die Fragestellung aus der Sicht dieser "Theorien"-Klasse deshalb, weil hier "Verstaendigung" ohne Rekurs auf den situativen Kontext der Sprechsituation erreicht werden soll und insofern die Bedingungen gegenueber der - beispielsweise von Searle (1960) untersuchten - alltaeglichen Kommunikationssituation verschaerft sind.

Aus unseren fruecheren Ueberlegungen zur Struktur des Beobachtungsproblems geht hervor, dass schon der Versuch einer praezisen Bestimmung der fuer ein einzelnes Individuum spezifischen semantischen Sprachstruktur und erst recht der interindividuelle Vergleich auf erhebliche Schwierigkeiten stoesst. Wir muessen uns im vorgegebenen Rahmen mit Schaetzungen fuer den jeweils beobachterspezifischen intraindividuellen Bedeutungszusammenhang der einbezogenen Woerter begnuegen.

Im Unterschied zu der Vorgehensweise im vorigen Abschnitt werden jetzt die Aehnlichkeitsangaben jedes einzelnen Beobachters benuetzt, um (nur) seine eigenen Beobachtungsangaben unter dem Bedeutungsaspekt zu korrigieren. Die bereinigten Daten finden danach Eingang in die ueblichen Analyseverfahren (z.B. Faktorenanalyse), und es wird festzustellen sein, ob und in welcher Weise sich die Ergebnisse von denen der Rohdatenanalyse unterscheiden. Fuer die Loesung dieses Problems sind - bei den

gegebenen Daten - verschiedene Ansaetze denkbar, die hier nicht im Detail diskutiert werden koennen. Wir beschraenken uns auf einige Bemerkungen dazu und stellen danach den von uns gewaehlten Weg im einzelnen dar.

Die von den Beobachtern erfragten Aussagen ueber die Aehnlichkeit aller vorkommenden Wortpaare koennen bei der Abbildung auf ein Modell als Distanzen zwischen Punkten oder als Winkel zwischen Koordinaten im mehrdimensionalen euklidischen (Sprach-)Raum dargestellt werden. Im ersten Fall waeren die Winkel zwischen den Radialvektoren fuer die formale Bestimmung der Bedeutungsueberschneidungen heranzuziehen, im zweiten die Projektionen von Strecken auf Koordinaten (zur inhaltlichen Bedeutung dieser formalsprachlichen Begriffe vgl. weiter unten). Beide Konzeptionen haben auf den ersten Blick den Vorzug, die in den Daten enthaltene Information ohne Verlust abzubilden. Dabei spannt sich stets ein Raum auf, dessen Dimensionalitaet direkt von der Zahl m der Elemente (Woerter) abhaengt (im ersten Fall (m-1), im zweiten m Dimensionen), soweit keine Bedeutungsgleichheiten auftreten; die Lage jedes neu hinzugefuegten Punktes bzw. jeder naechsten Koordinate j wird durch die (j(j-1)/2) Aehnlichkeitsangaben mit den bereits "eingezeichneten" (j-1) Elementen bestimmt. In objektsprachlicher Redeweise muesste dies bedeuten, dass jedes Wort, dessen Bedeutung mit keinem der bereits eingefuehrten fuer identisch gehalten wird, einen Bedeutungsueberschuss gegenueber diesen aufwiese in dem Sinne, dass es Bedeutungsteile enthielte, die in jenen nicht vorkommen.

Zweifel an der Angemessenheit solcher Modelle erwachsen im Zusammenhang mit wahrnehmungs- und sprachpsychologischen Untersuchungen (vgl. 2.1.4. und z.B. Hofstaetter 1973, 258 ff.), die auf die Begrenztheit des individuellen Differenzierungsvermoegens hinweisen und die bei der Modellkonstruktion als Begrenzung der Dimensionalitaet des Sprachraums eine Entsprechung finden muessten.

Fuer eine in dieser Hinsicht adaequate Modellkonstruktion waere die Kenntnis der Dimensionen des individuellen Wahrnehmungsdiskriminierungsvermoegens, also der Faehigkeit, diskrete Merkmale (Aspekte) (in) der Wahrnehmung zu differenzieren, und der Zuordnung der einzelnen Woerter zu diesen Dimensionen erforderlich (vgl. dazu z.B. Ahrens 1966). Diese beiden Informationsgruppen haben wir oben (2.1.5.) auch benutzt, um die Struktur des Bedeutungsproblems zu kennzeichnen. Die vorliegenden Aehnlichkeitsangaben sagen aber nichts ueber die "Wahrnehmungseinheiten", auf denen sie sich konstituieren. Es muss mit Blick auf die Begrenztheit der individuellen

Wahrnehmungsfaehigkeit (im o.a.Sinne) jedoch angenommen werden, dass die Dimensionalitaet des Bedeutungs- und damit des Sprachraums ebenfalls begrenzt ist.

Daraus folgt nicht notwendig, dass auch die Zahl der Woerter mit unterschiedlicher Bedeutung endlich ist, da sowohl ihre Gewichts- als auch die Intensitaetsdimension so konstruiert werden kann, dass auf ihnen kontinuierliche Variation moeglich ist. Damit ergibt sich, dass in ein und demselben Raum mehr bedeutungsverschiedene Woerter abgebildet werden koennen als disjunkte Subraeume in ihm konstruierbar sind.

Diese letzte Aussage erscheint trivial hinsichtlich der Eigenschaften des Modells, da ja - nach Konstruktion - bereits der eindimensionale Raum unendlich viele Punkte bzw. der zweidimensionale unendlich viele Geraden enthaelt. Sie ist aber nicht trivial im Hinblick darauf, dass mit ihr die Frage nach dem adaequaten Modell anders beantwortet wird, als es in der am Anfang dieses Abschnitts geschilderten Konstruktion der Fall war.

Die Schwierigkeit besteht nun darin, dass die (m(m-1)/2) empirischen Aehnlichkeitsangaben zu den m Woertern i.d.R. nicht in einem Raum abgebildet werden koennen, der weniger als (m-1) bzw. m Dimensionen aufweist, ohne gegen die Axiome fuer den euklidischen Raum zu verstossen. Fuer den Fall der Abbildung von Woertern als Punkte macht man sich dies leicht klar, wenn man die als Distanz d zwischen den Punkten A und B interpretierte Aehnlichkeit zwischen zwei Woertern als Strecke zeichnet und nun die Aufgabe loesen sollte, auf dieser Strecke einen weiteren Punkt (Wort) C anzugeben, der von A den Abstand e und von B den Abstand f aufweist, fuer e+f ungleich a. Entsprechendes gilt fuer den Versuch, zwischen zwei Geraden, die den Winkel alpha einschliessen, in derselben Ebene eine dritte Gerade (Strahl) einzuzeichnen, die mit der ersten den Winkel (die Aehnlichkeit) beta und mit der zweiten den Winkel gamma bildet, fuer beta + gamma ungleich alpha.

In diesem Sinne sind alle Modelle mit einer geringeren Dimensionalitaet als (m-1) bzw. m durch unsere Daten ueberdeterminiert <45>, und zwar i.d.R. in einer inkonsistenten, d.h. die Axiome des euklidischen Raums verletzenden Weise (Dreiecksgleichung!). Da nach dem bisher Gesagten solche Modelle zwar angemessen sein koennen, andererseits aber weder ihr innerer Aufbau noch die fuer die einzelnen Elemente erforderlichen topologischen Bestimmungsstuecke bekannt sind, ist eine direkte Abbildung der Daten nicht moeglich.

In dieser Situation bieten sich (wenigstens) drei Auswege an. Der im Hinblick auf die vorliegende Methodenliteratur naheliegendste fuehrt zum Verfahren der multidimensionalen Skalierung, die - grob gesprochen - eine Korrektur der erwaehnten Inkonsistenzen fuer einen euklidischen Raum moeglichst geringer Dimensionalitaet vornimmt (vgl. z.B. Ahrens 1974, 32; Shepard 1972, 9-19). Da sie jedoch prinzipiell auf der Idee der Mittelwertsbildung aufbaut - auch bei den fortgeschritteneren Verfahren zur Skalierung interindividueller Differenzen (vgl. Kuehn 1976, 100 ff.) -, scheidet sie aus theoretischen Erwaegungen hier aus.<46>

Der zweite Weg wird mit dem Verzicht auf Voraussetzungen, die in den Axiomen des euklidischen Raumes enthalten sind, beschritten; er fuehrt zu topologischen (allgemeinen) Raeumen, in denen beispielsweise mit dem Konstrukt der "Umgebung" bessere Anpassungen auf individuelle Besonderheiten des Sprach- und Bedeutungsraumes gelingen koennten. Wir unterlassen es, diesen Weg zu gehen, weil er - um im Bild zu bleiben - fuer unsere Bedurfnisse nicht ausreichend geebnet und hinsichtlich des ueber ihn erreichbaren Ziels zu unsicher ist. Hier muesste insbesondere auch ein staerker ausgearbeiteter Theoriestand erreicht werden, um die im Modell aufgehobenen Differenzierungen ausnuetzen zu koennen.

Der dritte Weg schliesslich, den wir einschlagen und im folgenden Abschnitt skizzieren wollen, ist in mathematisch-statistischer Hinsicht viel weniger anspruchsvoll als die beiden anderen; er passt sich aber, soweit wir sehen, nicht nur bezueglich der Voraussetzungen den in den Daten ruhenden empirischen Gehalten besser an, sondern er orientiert sich auch strikt am Prinzip der Individualitaet des Sprechers und erlaubt darueberhinaus einige "Verzweigungen", deren Verfolgung deshalb angezeigt scheint, weil beim gegenwaertigen Kenntnisstand von den "Enden" nicht sicher gesagt werden kann, ob sie das angestrebte Ziel (wahrer Wert der Bedeutungsueberschneidung) darstellen. Da ohne eine Loesung des Dimensionalitaetsproblems die Bestimmung der Bedeutungsueberschneidung (als Winkel zwischen den Radialvektoren) im Wort-Punkt-Modell nicht moeglich ist, waehlen wir als Abbildung fuer das Wort W(j) die Raumkoordinate w(j), auf der die in der Beobachtung wahrgenommene Ausspraegung des Merkmalsfeldes als Strecke (OZ(j)) abgetragen werden kann. Die Aehnlichkeitsrelation wird als Winkel alpha zwischen zwei Koordinaten wiedergegeben, die einen obliquen zweidimensionalen Raum aufspannen. Durch Projektion der (Beobachtungs-)Strecke z(j) auf die jeweils andere Koordinate laesst sich der Anteil OZ'(i) (bzw. OZ'(k)) an der dort aufgetragenen (Beobachtungs-)

Strecke OZ(k) (bzw. OZ(i)) bestimmen, der bei Kenntnis eines der beiden Beobachtungswerte "vorausgesagt" werden kann (Abb. 3-9: Zweidimensionales Modell der Aehnlichkeitsrelation).

In diesem Modell wird also der Anteil der Bedeutungsgemeinsamkeit zwischen zwei von einem Individuum gebrauchten Beobachtungswoertern als Strecke OZ' dargestellt. Die Strecke Z'(k)Z(k) symbolisiert beispielsweise den fuer W(k) eigenstaendigen Bedeutungsanteil nach Projektion von OZ(i) auf OZ(k), so dass fuer diesen Fall - nach der Bereinigungsprozedur - die beiden Daten fuer W(i) und W(k) die Werte der Streckenlaengen OZ(i) und Z'(k)Z(k) annehmen (entsprechend waere bei Schaetzung von w(k) aus zu verfahren).

Die Generalisierung auf den Fall W(j) fuer j>2 erfolgt nun nicht durch Einfuehrung von Restriktionen, die einen konsistenten Sprachraum unterstellen, sondern durch Anpassung an die in der Datengenese eingesetzte Strategie des konsekutiven Paarvergleichs (fuer die Einzelheiten vgl. Abschn. 3.4.2.). Eine solche parameterfreie Schaetzung von individuell korrigierten Werten aus den individuellen Aehnlichkeitsvorstellung en vermeidet eine Manipulation der Daten, die nur aus der vorausgesetzten, aber nicht gepruefteN Gueltigkeit eines handlicheren Modells begruendet werden koennte. Dieses Vorgehen ist hier angezeigt, weil wir nicht auf der Suche nach einem solchen Modell sind, das im uebrigen hinsichtlich seiner Dimensionalitaet persoenlichkeitspsychologisch zu verankern waere; vielmehr versuchen wir ja abzuschaetzen, welches Ausmass an Konsequenzen ins Auge gefasst werden muss, wenn der Gesichtspunkt der Bedeutungsueberschneidung in Unterrichtsanalysen eingebracht wird, die auf Ratingbasis erstellt sind.

Da alle Daten die Verhaltensebene betreffen, ist nicht zu befuerchten, dass mit der im Modell vorgenommenen Projektion auch Kausalrelationen ("Logikfehler") erfasst wuerden. Allerdings kann auch nicht gesagt werden, in welchem Umfang diese und die uebrigen Fehlerquellen in den korrigierten Daten noch enthalten sind bzw. inwieweit sie wegen Ueberlagerung mit der Bedeutungsueberschneidung gleichzeitig eliminiert werden.

## 3.4.2. Das Analyseverfahren

### 3.4.2.1. Das Abfolgeproblem

Mit der Wahl einer flexiblen Schaetzstrategie zur Bestimmung des Bedeutungsfehlers, die auf die jeweils spezifische Konstellation der Daten beim einzelnen Rater angepasst ist, eroeffnen sich drei Entscheidungsfelder, in denen fuer die Durchfuehrung Festlegungen vorzunehmen sind. Zunaechst ist zu bestimmen, nach welchem Konzept eine Auswahl aus den hinsichtlich des Modells redundanten (geometrisch; ueberdeterminierten und inkommensurablen) Aehnlichkeitsangaben vorgenommen werden soll (vgl. 3.4.1.). Dazu muss ueberlegt werden, mit welchem Wort (Beobachtungswert; "Schaetzer"; z(1)) beim Rater (l) die Schaetzung ausgefuehrt und auf welches andere Wort (geschaetzter Wert; z'(2)) sie bezogen werden soll. Die Moeglichkeiten sind formal gesehen sehr vielfaeltig, reduzieren sich aber angesichts der Fragestellung auf drei Gruppen. Da der logische Zirkel zu vermeiden ist, muessen von vornherein alle reziproken Relationen vermieden werden, d.h. dass eine Schaetzung von z(1) auf z'(2) weder direkt noch indirekt zugleich mit einer Schaetzung von z(2) auf z'(1) vorkommen darf. Die erste verbleibende Moeglichkeitsgruppe beruht danach auf dem Konzept der Reihenfolge, das die Schaetzung ("&") so organisieren koennte:

[8a] z(1) & z'(2); z(3) & z'(4);...;z(m-1) & z'(m)oder

[8b] z(1) & z'(2); z(2) & z'(3);...; z(m-2) & z'(m-1); z(m-1)& z'(m).

In beiden Faellen bliebe w(1) als Originalwert erhalten, in [8a] ausserdem ((m/2)-1) weitere Werte. Fuer beide Faelle wuerden inhaltlich begruendbare Abfolgekonzepte vorausgesetzt.

Die zweite Gruppe benuetzt das Konzept der Selektion, mit dem ein Wort ausgewaehlt wird, von dem her die Schaetzung der Werte aller anderen vorgenommen wird:

[9] z(i) & z'(1),...,z'(j),...,z'(m)fuer i ungleich j.

Fuer diese Variante muesste eine begruendete Wahl von z(i) erfolgen.

Die dritte Gruppe stellt eine Kombination der ersten beiden dar, Sie besteht in der sukzessiven Schaetzung aller (m-i) Werte aus z(i) fuer i=1,..,(m-1), Die dabei auftretenden Mehrfachschaetzungen koennen gemaess zusaetzlich einzufuehrender Kriterien (vgl, 3,4,2,2,) behandelt werden;

[10] z(i) & z'(j) fuer i < j; i=1,..,(m-1); j=2,..,(m),

Wir verfahren nach dem in [10] skizzierten Algorithmus, weil er die Moeglichkeit eroeffnet, mit Hilfe von Zusatzkriterien im Hinblick auf die Rigiditaet der Schaetzung gezielt die Extremkonstellationen konservativer und exhaustiver Datenbehandlung anzusteuern,

Fuer die Bestimmung der Abfolge (i) in [10] kann unter diesem Aspekt zwischen aufsteigender und fallender Beobachtungswert-Anordnung gewaehlt werden, Es ist klar, dass die Strecke z(kk) (vgl, 3-9) bei gegebenem alpha (ik) um so kleiner wird, je groesser z(i) ist, so dass bei Schaetzung mit fallenden z(i)-Werten zunaechst hohe z'(j)-Werte und damit geringe Differenzen zu den urspruenglichen Werten z(j) entstehen, vice versa,

Die Strecke OZ'(k) steht im Modell also fuer den Anteil an der von einem Rater konstatierten Merkmalsauspraegung z(k) (genauer; Auspraegung in der Gruppe von Einzelmerkmalen, auf die er bei der Vorgabe des Beobachtungswortes W(k) achtet), der bei Kenntnis seiner Aussage z(i) unter Verwendung des Beobachtungswortes W(i) und seines Sprachgebrauchs vorhergesagt werden koennte; m,a,W,: hat ein Beobachter eine Angabe darueber gemacht, welchen Auspraegungsgrad in dem mit W(i) angesprochenen Merkmalsbereich ein real agierender Lehrer aufweist, so kann eine Vorhersage darueber gemacht werden, welchen Mindestwert er fuer W(k) nennen wird, wenn die Bedeutungsueberschneidung zwischen W(i) und W(k) bekannt ist, Dieser Mindestwert wird im Modell durch z'(k) abgebildet, Nur der ueber Z'(k) hinausgehende Teil, naemlich die Strecke z(kk), symbolisiert die von W(i) unabhaengige Information, die in die Datenanalyse eingehen darf, Sie wird rechnerisch mit

[11] z(kk) = z(k) - z'(k) und

[12] z'(k) = z(i) cos alpha(ik)

ermittelt,

Versieht man beim Rater (l) gemaess [8] - [10] die Beobachtungswoerter mit einer fortlaufenden Rangziffer in der Reihenfolge der fuer sie angegebenen absteigenden oder aufsteigenden Beobachtungswerte, so ergeben sich die folgenden Schaetzreihen;

[10'] z(1) & z'(2,1), z'(3,1),..., z'(m,1)
z(2) & z'(3,2),..., z'(m,2)
.
.
.
z(m-1) & z'(m,m-1).

Man sieht in dieser Schreibweise von [10] deutlich, dass z.B. fuer z(3) zwei, allgemein fuer z(j) also (j-1) moeglicherweise verschiedene Schaetzwerte z'(j) ermittelt werden. Neben diesem Problem (vgl. unten 3.4.2.2.) faellt aber in [10'] auch der damit zusammenhaengen de Umstand ins Auge, dass der als Schaetzer benuetzte Wert z(j) nicht fuer die Bestimmung eines Schaetzwertes z'(j-i) herangezogen werden darf und damit die in der Menge der geschaetzten Werte tatsaechlich vorkommenden Betraege von der gewaehlten Reihenfolge abhaengen. Daher ist es erforderlich, neben dem Kriterium der Werthoehe fuer den Fall gleich grosser Beobachtungsangaben zusaetzliche Entscheidungsgesichtspunkte einzufuehren, die eine Rangfolge erzeugen und inhaltlich begruendet sind. Als erstes Eventualkriterium fuehren wir das Ausmass der Bedeutungsaehnlichkeit mit den noch zur Schaetzung anstehenden Woertern ein, subsidiaer als zweites die Streuung dieser Aehnlichkeitswerte und als letztes - falls vorher keine Entscheidung moeglich war - die Abfolge im Beobachtungsbogen.

Die Rangfolge kann am jeweils kleinsten oder groessten Wert des Eventualkriteriums orientiert werden, d.h. inhaltlich nach dem Ausmass der Bedeutungsaehnlichkeit mit den uebrigen Woertern bzw. nach dem Grad der Affinitaet zu diesen. Je nach der gewaehlten Strategie wird, wie das folgende fiktive Beispiel zeigt, bei gleich grossen Beobachtungswerten der kleinste und groesste Kriteriumswert benuetzt (vgl. zur Struktur der Tabelle; 3-3);

| Wort Nr. | Aehnlichkeit (Groesse von alpha(i,j) in Grad 1 | 2 | 3 | 4 | 5 | 6 | Beob.-Daten |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 10 | 10 | 80 | 50 | 20 | 0 |
| 2 | | 0 | 80 | 70 | 30 | 40 | 5 |
| 3 | | | 0 | 60 | 70 | 50 | 6 |
| 4 | | | | 0 | 10 | 60 | 5 |
| 5 | | | | | 0 | 60 | 5 |
| 6 | | | | | | 0 | 4 |

3-10 Zahlenbeispiel zur Bestimmung der Schaetzrangfolge

Die Woerter Nr. 2,4,5 weisen gleich grosse Beobachtungswerte auf, fuer sie sind die Eventualkriterien zu bestimmen:

| Wort Nr. "Schaetzer" | zu schaetzende Woerter Nr. / Aehnl.werte mit Schaetzer | | | | Werte der Evtl.krit. durchschn. Aehnlichk. | Std.-abw.-ng. |
|---|---|---|---|---|---|---|
| | 1 | 4 | 5 | 6 | -- | -- |
| 2 | 10 | 70 | 30 | 40 | 37,5 | 25,00 |
| | 1 | 2 | 5 | 6 | -- | -- |
| 4 | 80 | 70 | 10 | 60 | 55,0 | entf. |
| | 1 | 2 | 4 | 6 | -- | -- |
| 5 | 50 | 30 | 10 | 60 | 37,5 | 22,17 |

3-11 Zahlenbeispiel zur Bestimmung der Werte in den Eventualkriterien

Damit ergeben sich vier verschiedene Schaetzrangfolgen, die zu unterschiedlichen Ergebnissen fuehren koennen:

| Hauptkriterium | Strategie<+> | Abfolge | | | | | |
|---|---|---|---|---|---|---|---|
| Fallende Beobachtungswerte | konservativ | 3 | 4 | 2 | 5 | 6 | 1 |
| | exhaustiv | 3 | 5 | 2 | 4 | 6 | 1 |
| Steigende beobachtungswerte | konservativ | 1 | 6 | 4 | 2 | 5 | 3 |
| | exhaustiv | 1 | 6 | 5 | 2 | 4 | 3 |

<+> Vgl. zu den beiden Kategorien unten 3,4,2,2.

3-12 Beispiel fuer kriteriumsabhaengige Schaetzrangfolgen

Ein besonderes Problem entsteht bei den Abfolgen der beiden unteren Zeilen von 3-12, wenn, wie in unserem Beispiel, der zum ersten wort gehoerende Beobachtungswert z(1)=0 ist, da dann zunaechst gemaess [12] alle Schaetzwerte z'(j) ebenfalls Null werden. Man kann diese Konsequenz mit der Begruendung vermeiden wollen, dass der Beobachtungswert Null von der graphischen Darstellung der Skala "erzwungen" werde, weil einerseits unterhalb von "1" keine weitere Ankreuzmoeglichkeit vorgesehen sei, andererseits aber stets ein Wert groesser Null erwartet werden muesse (vor allem im Bereich der "Urteile"; vgl. Tab. 2). Fuer diesen Fall haben wir in dem zur Berechnung erstellten Computerprogramm eine Variante vorgesehen, die mit dem kleinsten Wert beginnt, der nicht Null ist.

## 3,4,2,2. Die Schaetzstrategie

Wie bereits besprochen, besteht unsere Aufgabe im Rahmen dieser methodisch orientierten Studie darin, den Wirkungsbereich des Bedeutungsfehlers auszuloten. Das bedeutet, dass wir versuchen muessen, die Extreme der geringsten und der weitestgehenden Folgen zu bestimmen. Der erste Fall liegt dann vor, wenn die urspruenglichen Beobachtungswerte bei gegebenen Bedeutungsaehnlichkeiten nur um den kleinstmoeglichen Betrag veraendert werden. Wir nennen die Verfolgung dieses Ziels eine konservative Strategie, weil sie diejenigen Ergebnisse so weit wie moeglich zu erhalten trachtet, die mit dem konventionellen Ansatz der Unterrichtsforschung ermittelt worden waeren. Im Modell lautet ihre Fassung gemaess [10] bzw.

[10'] und [12];

[13a] $z'(k) = \min|z'(k,i)|$ fuer $i=1...(m-1)$,

Setzt man den mit [13a] gefundenen Wert in [11] ein, so erhaelt man den Anteil z(kk), der spezifisch fuer W(k) ist und der im Extremfall, naemlich wenn $z'(k)=0$, nicht von z(k) abweicht.

Eine analoge Verfahrensweise gilt fuer die exhaustive Strategie, also dasjenige Verfahren, mit dem die hoechstzulaessige Veraenderungsmoeglichkeit der Originaldaten bei gegebenen Bedeutungsaehnlichkeiten ausgeschoepft wird:

[13b] $z'(k) = \max|z'(k,i)|$ fuer $i=1...(m-1)$,

Zwei Sonderfaelle muessen hier kurz diskutiert werden. Rein formal ist es moeglich, dass

[14a] $z(k) < z'(k)$ oder

[14b] $z'(k) < 0$,

Aus inhaltlichen Gruenden muessen aber beide Faelle vermieden werden. Aus [14a] wuerde sich ein negativer wahrer Beobachtungswert ergeben, der nach unserer Skalendefinition nicht denkbar ist. Daher ist [14a] als das Ergebnis eines kumulativen Fehlers zu deuten und fuer [13a,b] nach [11] die zusaetzliche Beschraenkung

[15a] $z(kk) >= 0$

zu formulieren.

Da der Bedeutungsfehler als Gemeinsamkeit zwischen Wortbedeutungen dargestellt wurde, kann die Bereinigung sich stets nur als Verringerung des Originalwertes um den gemeinsamen Teil auswirken. [14b] beschreibt aber den Fall einer Erhoehung des zu schaetzenden Wertes gegenueber dem urspruenglichen. Zwar waere auch im Kontext der hier entworfenen Konzeption des Bedeutungsfehlers ein solcher Fall denkbar, naemlich dann, wenn die Bedeutungen von W(i) und W(k) voellig identisch, die Beobachtungswerte aber unterschiedlich waeren. Eine Korrektur

von z(k) nach z'(k)=z(i) koennte dann u.U. erforderlich sein, Voraussetzung dafuer waere aber die Kenntnis des wahren Wertes, der hier ja gerade geschaetzt werden soll, Aus diesem Grunde fuehren wir generell in [1*a,b] die Vorsichtsmassnahme

[15b] z(kk) <= z(k)

ein, Dem weitergehenden Problem von inhaerenten Strukturen der Wortbedeutungen gehen wir im naechsten Abschnitt nach,

### 3,4,2,3, Das Problem der Sprachebenen

In den Untersuchungen von Mueller-Wolf/Fittkau (1971) und Mueller-Wolf (1977) werden zwei Itemgruppen voneinander unterschieden; eine davon erhaelt die Bezeichnung "Globalkonzepte", "globale Kriterien des Lehrverhaltens", "Globalskalen" (1977, 99, 102), Mit diesen Begriffen soll zum Ausdruck gebracht werden, dass sie in irgendeinem Sinne umfassender sind als die uebrigen (z,B, "sozialintegrativ" gegenueber "entspannt"), Wir haben bereits weiter oben (2,3,[illegible],1,2,) auf metatheoretische und methodologische Probleme hingewiesen, die mit der undifferenzierten Mischung solcher Items im Erhebungsinstrument verbunden sind, Sie liegen hauptsaechlich in der unkontrollierten Vermischung von semantischen Stufen einerseits und Faktoren unterschiedlicher Ordnung andererseits,

Nun ist es in einer konventionalistischen Konzeption selbstverstaendlich prinzipiell moeglich, unter Absehung von umgangssprachlichen Usancen Wort-Bedeutung-Relationen einzufuehren, Gleich wie man sich entscheidet, muss jedoch darauf geachtet werden, dass Erhebungs- und Auswertungsstruktur uebereinstimmen, Es ist zumindest zweifelhaft, ob diese Aufgabe in den beiden angefuehrten Untersuchungen geloest wurde, da im Erhebungsbogen Global- und uebrige Urteile vermischt und erst in der Interpretation der Analyse getrennt werden (vgl, 1977, 91 ff,),

Fuer unsere Untersuchung fuehrt dieser Umstand, wie oben dargelegt wurde, zu der Konsequenz, in der Datenerfassung eine implizite Trennung vorzunehmen (abschnittsweise Praesentation der Wortgruppen (vgl, Tab, 2)), so dass individuelle Bedeutungskonzepte von Ebenendifferenzierung oder Ebenennivellierung benutzt werden konnten, Es

laesst sich aber aus den Aehnlichkeitsangaben eines einzelnen Pbn, nicht zuverlaessig ermitteln, von welcher Sprachsystematik er Gebrauch macht.

Fuer eine heuristische Klaerung dieser Frage unterwarfen wir die Originalwerte aus der Beobachtung einer hierarchischen Clusteranalyse, deren Ergebnisse sich bei unterschiedlichen Aehnlichkeitskoeffizienten (Korrelation, quadratische euklidische Distanz, Fehlerquadratsumme, Mittelwert-, Varianzunterschied), verschiedenen Verbindungskriterien (Centroid-, Median-, Ward's Kriterium) und variierenden Startkonfigurationen (nach Zufall, willkuerlich, Idealkonstellation) einigermassen stabilisieren. Das Dendrogramm zeigt die Ergebnisse einer Centroidloesung auf der Basis von Korrelationskoeffizienten mit standardisierten Rohwerten (Abb, 3-13 Dendrogramm der Beobachtungswoerter).

Die Darstellung 3-13 legt die Vermutung nahe, dass viele Beobachter eine Sprachbereichstrennung zumindest naeherungsweise einhalten, wobei allerdings nicht klar zu erkennen ist, ob eine Niveaustruktur zugrunde liegt. Bei einem Minimum von drei Clustern (Sprachbereichen) laesst sich bis in den vorletzten Zyklus die "theoretische" Struktur weitgehend rekonstruieren. Erst im letzten Durchgang (Strich-Punkt-Linie) erfolgt eine erwartungsfremde Fusion, die auch im Modell durch einen Vorzeichenwechsel des Koeffizienten auffaellig wird.

Ohne hier in die Diskussion weiterer Einzelheiten einzutreten, kann dieser Befund als zureichende Begruendung dafuer genommen werden, dass die Bedeutungsanalyse nicht nur das Ein-, sondern auch das Mehrbereichsmodell von Sprachanwendung beruecksichtigen muss. Waehrend im ersten Falle ("sprachuebergreifend") die vom Pbn. angegebenen Aehnlichkeiten zwischen allen Woertern in Betracht kommen (nach Massgabe der in den beiden vorigen Abschnitten dargestellten Regeln), duerfen im zweiten Falle ("sprachbereichsintern") nur die Aehnlichkeitsangaben fuer Woerter innerhalb jedes der drei Bereiche beachtet werden, d.h. dass fuer jeden Bereich ein gesonderter Schaetzalgorithmus zu durchlaufen ist; dabei werden dann die urspruenglichen Haeufigkeitsbeobachtungswerte unveraendert erhalten, da wir fuer sie, wie dargelegt (vgl. 3.4.1.) keine Aehnlichkeitsangaben erfasst haben.

In der folgenden Tabelle sind die Varianten fuer die Bedeutungsanalyse auf Beobachterbasis systematisch zusammengefasst. Sie ergeben sich aus den in diesem Kapitel eroerterten Gruenden.

| | | Schaetzstrategie | | |
|---|---|---|---|---|
| | | konservativ: | | |
| Abfolgeregel (i<j) | | z(i)>z(j) | z(i)<z(j) | 0<z(i)<z(j) |
| Sprachkonzept | sprachbereichuebergr. | A | B | C |
| | sprachbereichintern | G | H | I |
| | | exhaustiv: | | |
| | sprachbereichuebergr. | D | E | F |
| | sprachbereichintern | J | K | L |

3-14 Varianten der Bedeutungsfehlerschaetzung

Es muessen zwoelf verschiedene Auswertungsmodi verfolgt werden, die den Rahmen fuer die Einordnung der Konsequenzen des Bedeutungsfehlers bilden. Wir werden hier nicht alle zwoelf Varianten im Detail auswerten und diskutieren. Die Kriterien fuer eine Auswahl werden angesichts der Ergebnisse im naechsten Schritt im einzelnen besprochen.

Unter dem Aspekt der Rigorositaet, mit der die Informationen ueber die Bedeutungsbeziehungen im Modell zur Geltung gebracht werden, bildet die Anordnung in 3-14 eine angenaeherte aufsteigende Rangordnung, die bei der Betrachtung der Resultate eine Interpretationsorientierung leisten kann.<47> Neben den anderen Gesichtspunkten werden wir unser Augenmerk stets auch auf die beiden Exponentenpaare (A vs. F bzw. G vs. L) richten, um zu beobachten, wie sich die Schaetzstrategien bei ihnen auswirken.

### 3.4.2.4. Ein Berechnungsbeispiel

Um den Effekt der Auswertungsmodi zu verdeutlichen, stellen wir in der folgenden Tabelle die Berechnungsabfolge gemaess [10] bis [13] und [15] fuer Variante G mit den Zahlen des oben eingefuehrten Beispiels 3-10 und 3-12 dar.

| Abfolge Wort Nr. | W(i) 3 | 4 | 2 | 5 | 6 | ausgew. Sch.-werte | Org.-wert | richtiger Wert |
|---|---|---|---|---|---|---|---|---|
| W'(k);1 | 5,908 | &0,868 | 4,924 | 3,214 | 3,758 | 0,868 | 0,0 | 0,000 |
| 2 | &1,041 | 1,710 | - | - | - | 1,041 | 5,0 | 3,959 |
| 3 | - | - | - | - | - | 0,000 | 6,0 | 6,000 |
| 4 | 3,856 | - | - | - | - | 3,856 | 5,0 | 1,144 |
| 5 | &2,052 | 4,929 | 3,214 | - | - | 2,052 | 5,0 | 2,948 |
| 6 | 3,852 | &2,500 | 4,698 | 2,500 | - | 2,500 | 4,0 | 1,500 |

&: Ausgewaehlter Wert gemaess Formel 13a.

3-15 Beispiel fuer die Bestimmung von z'(k) gemaess Variante G aus (3-12).

Gemaess [10] bleibt der erste Schaetzer (hier: Wort Nr. 3) erhalten und wird daher in der drittletzten Spalte mit dem Subtrahenden Null eingesetzt. Die uebrigen Werte z'(k,i) (vgl. [13a]) werden solange von Spalte zu Spalte berechnet als nicht die zu ihnen gehoerenden Originalwerte z(i) des Wortes W(i) selbst als Schaetzer herangezogen werden. So ergibt sich z.B. der erste Wert in Zeile 2, Spalte 2 gemaess [12]:

$$z'(2) = z(4)\ (\cos \text{alpha}(4,2))$$
$$= 5 \cos 70 \text{ Grad}$$
$$= 5(0,3420) = 1,710.$$

Nach [13a] ist fuer z'(2) jedoch der Wert als Schaetzwert fuer die Bedeutungsuebereinstimmung zu waehlen, der sich aus der Aehnlichkeit mit Wort Nr. 3 ergibt: 0,868. Damit errechnet sich der berichtigte Wert fuer die Beobachtung des (fiktiven) Probanden unter den mit Wort Nr. 2 von ihm beachteten Aspekten nach [11] (vgl. auch Abb. 3-9):

$$z(2,2) = z(2) - z'(2)$$
$$= 5,000 - 1,041 = 3,959.$$

Zur Illustration der Unterschiedlichkeit der verschiedenen Auswertungsmodi sind in der folgenden Aufstellung die Ergebniswerte z(kk) fuer die sprachstufeninternen Varianten zusammengestellt:

| Variante | Org. werte | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|
| Wort Nr. 1 | 0 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| 2 | 5 | 3,959 | 5,000 | 3,290 | 0,670 | 0,670 | 0,670 |
| 3 | 6 | 6,000 | 6,000 | 5,132 | 6,000 | 2,786 | 2,786 |
| 4 | 5 | 1,144 | 5,000 | 3,000 | 0,076 | 0,076 | 0,076 |
| 5 | 5 | 2,948 | 5,000 | 3,000 | 2,948 | 3,000 | 3,000 |
| 6 | 4 | 1,500 | 4,000 | 4,000 | 0,144 | 4,000 | 4,000 |

3-16 Ergebnisse der Eliminierung des Bedeutungsfehlers nach unterschiedlichen Schaetzverfahren

Man sieht, dass die Resulte recht unterschiedlich ausfallen, lediglich Variante J und L fuehren hier zu den gleichen Werten. Dies liegt aber an der Wahl des Beispiels, das einigermassen uebersichtlich bleiben sollte. Man kann sich leicht klarmachen, dass - etwa, wenn mehr als eine Beobachtungsangabe Null ist - auch diese beiden Varianten zu verschiedenen Schaetzungen fuehren.

Fuer Variante H, in der sich hier die Originalwerte erhalten, gelten aehnliche Ueberlegungen: kommt in den Beobachtungsangaben eines Probanden der Wert Null nicht vor, so aendern sich auch hier die urspruenglichen Daten. Insgesamt wird deutlich, dass die konservativen Verfahren (G, H, I) zu geringeren Abweichungen von den urspruenglichen Werten fuehren als die exhaustiven. Das muss allerdings nicht bedeuten, dass mit ihnen auch staerkere Aehnlichkeiten mit den Analyseergebnissen der Originalwerte verbunden sind. Diese bauen ja auf der Kovarianzmatrix der Merkmale, nicht der Personen auf, waehrend die Bedeutungskorrektur personenbezogen erfolgt und insofern Effekte erzeugen kann, die nicht absehbar sind.

### 3.4.3. Die Mehrdeutigkeit der Ratingdaten im Lichte der Bedeutungsanalyse

#### 3.4.3.1. Der Schwankungsbereich des Reliabilitaetsmasses

Fuer die korrigierten Rohdaten koennen zu jeder Schaetzvariante unabhaengige Reliabilitaetskoeffizienten berechnet werden, da die bedeutungsbezogene Rekonstruktion an individuellen Datensaetzen auf der Basis datenunabhaengiger Information erfolgt. Damit wird eine jeweils neue Beurteilung der Datenqualitaet ermoeglicht, wie sie im - von uns kopierten - "klassischen" Design zum Standardinstrumentarium gehoert. Waehrend man sich jedoch dort haeufig darauf beschraenkt, die mittlere Reliabilitaet fuer alle Schaetzungen (z.B. gemaess [2]) anzugeben, ist es in einer methodenkritischen Untersuchung erforderlich, die durchschnittliche Reliabilitaet fuer einen Rater zu untersuchen, weil in ihr "die Einschaetzungsguete der Messprozedur abhaengig von der Zahl der Rater" (Langer/Schulz v.Thun 1974, 92) zum Ausdruck kommt.

Formal gesehen bemisst sich der Unterschied zwischen der Gesamtreliabilitaet r(nn) und der Reliabilitaet fuer einen einzelnen Rater r(11) nach dem Verminderungsfaktor

$$[16] \quad d(min) = \frac{Var(o)}{Var(o)+(n-1)\,Var(e)} \quad \text{(vgl. Anm. 42)}$$

fuer $r(11) = r(nn)\,d(min)$.

Man sieht, dass die Veraenderung der Fehlervarianz fuer den Unterschied zwischen r(11) und r(nn) von entscheidender Bedeutung ist. Insbesondere dann, wenn ein umfassender Begriff der Fehlervarianz benuetzt wird, in den sowohl die Rater- als auch die Restvarianz eingehen, ist - vor allem, wenn eine grosse Zahl von Ratern eingesetzt war - bereits bei geringen Schwankungen mit starken Auswirkungen auf die durchschnittliche Reliabilitaet zu rechnen.

Dieser Begriff der Fehlervarianz ist fuer die folgende Analyse der r(11)-Werte jedoch nicht anzuwenden, da in unserer Fragestellung (R-Technik!) Niveauunterschiede zwischen Ratern nicht untersucht werden (vgl. ausserdem auch Ebel 1951, 411-412; Hartmann 1977, 106-107); d.h. dass eine Wechselwirkung zwischen den Haupteffekten nicht interpretiert, sondern als Fehlervarianz betrachtet wird. Ihre Schaetzung erfolgt daher als Bestimmung

der Restvarianz;

$$[17] \quad Var(e) = \frac{SAQ(tot)-SAQ(o)-SAQ(s)}{(n-1)(k-1)} ,$$

Die Schwankungen, denen die durchschnittliche Reliabilitaet pro Rater unterliegt, ergeben sich somit daraus, dass nach der - jeweils gemaess einer Schaetzvariante vorgenommenen - Beseitigung des Bedeutungsfehlers die Objektvarianz (wegen der "Berichtigung" der "Merkmalsbeschreibungen") und die systematische (Rater-)Varianz (wegen der interpersonellen "Sprachunterschiede") unabhaengigen Veraenderungen unterliegen, die sich gegenseitig kompensieren oder gleichgerichtet sein koennen,

Die im ersten Teil von Tab, 5 zusammengestellten Werte r(11) fuer das Einbereichsmodell von Sprachanwendung (Einbeziehung saemtlicher Variablen/Merkmale) weisen tatsaechlich erhebliche Unterschiede auf (,4942 (Variante A) >= r(11) >= ,2277 (Variante F)), die sich mit der Rangordnung der Schaetzrigorositaet im Einklang befinden, Berechnet man nach dem von Ebel (1951, 413-414) vorgeschlagenen Verfahren die Vertrauensintervalle fuer die einzelnen Koeffizienten, so zeigt sich, dass D und F ausserhalb des Schwankungsbereiches der Varianten A - C und E liegen,

Das bedeutet zunaechst, dass die verschieden strengen Auswertungen der Bedeutungsinformationen auch zu hoechst unterschiedlichen Beurteilungen der Guete des "Messinstruments" Anlass geben, Waehrend die mit A vorgenommene eher konservatie Korrektur zu einer geringen Steigerung des Zuverlaessigkeitsmasses fuehrt, erzeugen rigorosere Eingriffe Werte von r(11), die - falls sie zutraefen - den Verzicht auf das untersuchte Messverfahren nahelegen koennten,

Die durchweg hohen Koeffizienten fuer die Gesamtreliabilitaet (r(nn, A,,,F) > ,95) duerfen nicht darueber hinwegtaeuschen, dass sie - aus Gruenden unserer methodisch orientierten Fragestellung - unter extremen Bedingungen, naemlich hoher Raterzahl, zustande gekommen sind, Gerade in den Faellen, in denen zur Vermeidung der anderen Ratingfehler eine intensive Beobachterschulung betrieben wird, verzichtet man - unter oekonomischen Gesichtspunkten - i,d,R, auf den Einsatz vieler Beobachter, Die so motivierte methodische Sorgfalt waere aber dann, wenn sich unsere Schaetzvarianten D oder F als begruendet erwiesen, bei weitem nicht ausreichend, Lord und Novick haben gezeigt, dass die Reliabilitaet sich nicht proportional zur Raterzahl ("Testlaenge") entwickelt (1968,

118). Nach ihren Berechnungen liegt bei einem r(11) von ,2277 (fuer Variante F) r(nn) beispielsweise fuer 5 Beobachter noch unterhalb von ,60, einem Wert also, der sowohl fuer theoretische Untersuchungen als auch - mit Blick auf das Validitaetsproblem - fuer praktische Entscheidungen (z.B. Zuordnung von Lehrern zu Studentengruppen gemaess einem treat-treatment-Konzept) als unzureichend gelten wuerde (vgl. z.B. Lienert 1969).

Wir betrachten, bevor wir diesen Punkt vertiefen, die Ergebnisse fuer das Mehrbereichsmodell der Sprachanwendung (vgl. Tab. 5, zweiter Teil). Dabei beziehen wir uns zunaechst auf den einfacheren Fall von zwei Wortgruppen, naemlich "Urteilen" und "Haeufigkeitsbeobachtungen" (Variable 1-15 und 16-27). Fuer die letzteren liegen nur die auf die urspruenglichen Daten zurueckgehenden Werte fuer r(11) und r(nn) vor, da - aus den oben dargelegten Gruenden - hier nach Bedeutungsaehnlichkeiten nicht gefragt und deshalb auch eine Korrektur durch Schaetzung nicht vorgenommen wurde.

Mit ,5841 uebersteigt die durchschnittliche Reliabilitaet pro Rater alle uebrigen Koeffizienten betraechtlich. Diese liegen fuer den Urteilsbereich zwischen ,2988 (Variante I) und ,0078 (Variante J). Sie erfuellen damit zwar innerhalb der beiden Schaetzstrategien (konservativ und exhaustiv) nicht die intuitiven Rigorositaetserwartungen. Ein Blick auf die Vertrauensintervalle (5%-Niveau) zeigt jedoch, dass die konservativ ermittelten Werte, die sich nicht wesentlich voneinander unterscheiden, ueber den exhaustiv bestimmten liegen (mit Ausnahme von Variante K).

Wir brauchen uns in der Interpretation wiederum nicht mit der Einzelanalyse der voneinander abweichenden Schaetzergebnisse zu befassen, deren Gruende ja in der jeweils angewandten Strategie liegen. Vielmehr richten wir das Augenmerk auf die Spannweite der Moeglichkeiten, die in Deutungen von Ratingdaten beachtet werden muessten,<48> Fuer den Bereich der Urteile zeigt sich dabei zweierlei: zum einen erstrecken sich, wie beim Einbereichsmodell, die in Erwaegung zu ziehenden Masse ueber ein verhaeltnismaessig breites Band, zum anderen aber liegt dieses Band ueber einem so niedrigen Niveauabschnitt, dass sich hier nicht nur fuer den Fall einer bestimmten, sondern praktisch aller zu erwartenden Datenberichtigungen die Frage stellt, ob das Instrument nicht zu verwerfen sei.

Noch deutlicher wird dieses Problem, wenn man den Urteilsbereich weiter unterteilt. Die Items fuer Globalurteile (Variable 1-6) weisen fuer sich genommen - weit-

gehend erwartungsgemaess - Koeffizienten im Bereich von ,0826 (Variante G) >= r(11) >= ,0108 (Variante L) auf, die der Partialurteile (Variable 7-15) liegen bei ,3746 (Variante H) >= r(11) >= ,0876 (Variante L). Zwar bringt die "Entlastung" der Partialurteile durch die besonders unzuverlaessigen Globalurteile eine gewisse methodische Qualitaetssteigerung, aber es bleibt fraglich, ob Forschungsziele genannt werden koennen, fuer die solche Werte ausreichend sind.

An dieser Stelle ist es nun nicht unwichtig, darauf hinzuweisen, dass die Ergebnisse fuer die Mehrbereichsauffassung von Forschungssprache mit dem uebereinstimmen, was man auch ohne ihre Kenntnis scheinbar intuitiv haette erwarten koennen; je unklarer die vorgegebenen Woerter, desto unzuverlaessiger die Auskunft, die sie erlauben. Der entscheidende Punkt jedoch, der die gefundenen Resultate aus den Gefilden der Trivialitaet heraushebt, besteht darin, dass der Anspruch, unter dem die Begrifflichkeit derartiger Ratinguntersuchungen auftritt, eben nicht auf ihre umgangsprachliche Unklarheit gegruendet wird, fuer die eine durchaus akzeptable Reliabilitaetsprognose mit Hilfe von Plausibilitaetsueberlegungen haette erstellt werden koennen. Waere dies so, dann muesste unverstaendlich bleiben, warum dabei gewonnene "Daten" weitergehenden statistischen Modellbetrachtungen - wie etwa der Faktorenanalyse - unterworfen werden, in denen die Varianzen zur Grundlage theoretischer Interpretationen gemacht werden.

Die Beseitigung des Bedeutungsfehlers legt einen Teil dieses Problemfeldes frei, indem sie abgegebene einzelne "Bruttoinformationen" um denjenigen Teil reduziert, der in - scheinbar - unabhaengigen weiteren Informationen bereits enthalten ist. Dadurch wird der nivellierende Effekt der Redundanz in den Merkmalsbezeichnungen, der zugleich zu einer Ueberschaetzung der Datenzuverlaessigkeit fuehren kann, aufgehoben. Es zeigt sich dabei besonders fuer die umgangssprachlich vagen Wortbedeutungen, dass die erforderliche und vorausgesetzte Praezisierung fuer den deskriptiven Forschungszweck nicht vollzogen ist.

Wie wir bereits oben gezeigt haben, ist dies aber ein Punkt, in dem sich geschulte Beobachter von ungeschulten prinzipiell nicht unterscheiden. Durch eine sorgfaeltige "Operationalisierung" der Urteile koennte man dieses Problem systematisch vermeiden. Damit entfiele aber zugleich die praktische Notwendigkeit und vor allem die systematische Moeglichkeit, sie als Beobachtungskategorien einzusetzen. Sie wuerden vielmehr durch Woerter ersetzt, die auf dem Praezisionsniveau der hier verwende-

ten "Haeufigkeitsbeobachtungen" - oder hoeher - liegen und tatsaechlich auch die Voraussetzung fuer hoehere Reliabilitaet erfuellen.

Wir wollen die damit angesprochenen wissenschaftstheoretischen Ueberlegungen in diesem Abschnitt nicht weiter vertiefen. Dagegen waere hier die Frage aufzuwerfen, wie hoch die durchschnittliche Reliabilitaet pro Rater denn sein muesste, damit ausreichende Datenqualitaet unterstellt werden darf. Selbstverstaendlich handelt es sich hier um ein Entscheidungsproblem, zu dem vom Forschungszweck her argumentiert werden muss. Rein formal kann bei gegebener durchschnittlicher Zuverlaessigkeit durch (endliche) Erhoehung der Raterzahl jeder Reliabilitaetskoeffizient r(nn) < (1,-c) fuer beliebig kleines c > 0, erreicht werden. Wie aus der abgewandelten Spearman-Brown-Formel hervorgeht, besteht die folgende Beziehung zwischen Raterzahl (n) und Gesamtreliabilitaet (r(nn)) bei gegebener Durchschnittsreliabilitaet (r(11)):

$$[18]\quad n = \frac{r(nn)(1-r(11))}{r(11)(1-r(nn))}$$ (vgl. Lord/Novick 1968, 112, 119).

In welchen Dimensionen sich die Anzahl <49> der erforderlichen Beobachter bewegt, wird ersichtlich, wenn man das in unserer Untersuchung aufgetretene geringste (Variante J; Urteile) mit dem hoechsten durchschnittlichen Zuverlaessigkeitsmass (bei Haeufigkeitsbeobachtungen) vergleicht.

Wollte man eine Gesamtreliabilitaet von r(nn) = ,80 erhalten, so waeren bei r(11) = ,5841 mindestens 3, bei r(11) = ,0078 hingegen nicht weniger als 509 (!) Beobachter einzusetzen; fuer ein angestrebtes r(nn) = ,90 lauten die entsprechenden Zahlen 7 und 1 145. (Fuer die Bestimmung weiterer Werte sei auf das Nomogramm bei Lord/Novick (1968, 113) verwiesen).

In der methodenorientierten Literatur zum Ratingproblem fanden wir keine systematische Eroerterung der Frage nach der erforderlichen Hoehe der Reliabilitaetskoeffizienten. Da ihre Beantwortung, wie gesagt, in einer Normsetzung besteht, kann auch keine allgemeinverbindliche Aussage dazu gemacht werden. Ein oftmals brauchbares Kriterium liefert jedoch die Beachtung des Zusammenhangs mit einem angestrebten Validitaetskoeffizienten. Aus der bekannten Beziehung

$$[19]\quad r(vv) = \frac{r(v)\,n}{1+(n-1)r(11)}$$

mit r(vv):=Gesamtvaliditaet
r(v) :=Einheitsvaliditaet

ergibt sich naemlich, dass bei gegebenem r(v) (durchschnittliche Validitaet pro Schaetzung) die Verluste in der Gesamtvaliditaet durch Verringerung der Raterzahl dann am geringsten sind, wenn r(11) moeglichst hoch ist. Gerade fuer den Fall einer aufwendigen Raterschulung mag dieser Gesichtspunkt von Bedeutung sein.

Zur Illustration dieses Effektes ziehen wir noch einmal die in Variante J (Urteile) und in den unabhaengigen Haeufigkeitsbeobachtungen ermittelten Reliabilitaetskoeffizienten heran (vgl. auch Lord/Novick 1968, 116-117);

| | Werte fuer r(vv) bei r(v)=,30 Beobachterzahl: | | | | |
|---|---|---|---|---|---|
| r(11) | 10 | 5 | 3 | 2 | r(vv;10)-r(vv;2) |
| ,5841 | ,379 | ,376 | ,353 | ,337 | ,042 |
| ,0078 | ,917 | ,661 | ,516 | ,423 | ,494 |

3-17 Veraenderung der Validitaet bei verschieden hohen Koeffizienten fuer die Reliabilitaet pro Rater<50>

Man sieht, dass im Bereich n < 10, also gerade dort, wo eine Vielzahl von Ratinguntersuchungen zu lokalisieren ist, bei geringem r(11) extreme Validitaetsverluste auftreten koennen, wenn die Beobachterzahl schwankt.

Wendet man zur Verbesserung der Reliabilitaet das von Smith (1974) vorgeschlagene Verfahren der Eliminierung inkonsistenter Rater an, so muss man damit rechnen, dass der Gewinn an Messzuverlaessigkeit durch Verluste in der Validitaet ueberkompensiert wird. Eine Entscheidung in dieser Frage ist jedoch erst nach Beseitigung des Bedeutungsfehlers sinnvoll. Wie aus den Ergebnissen der verschiedenen Schaetzverfahren hervorgeht, schwankt ja r(11) erheblich in Abhaengigkeit von der jeweiligen Strategie. Dabei gehen die unterschiedlichen Werte, die r(11) annimmt, auf Datenaenderungen zurueck, die dann ihrerseits wechselnde Personen als die "schlechtesten" Rater ausweisen.<51>

### 3,4,3,2, Die Verschiedenheit der Faktorenstrukturen als Indikator deskriptiver Unzulaenglichkeit

Am deutlichsten treten die Folgen des Bedeutungsfehlers hervor, wenn man untersucht, zu welchen Ergebnissen Faktorenanalysen fuehren, die - wie dargelegt wurde - haeufig zur Verrechnung und Interpretation von Ratingdaten aus der Unterrichtsbeobachtung herangezogen werden. Hauptsaechlich zwei theoretisch unterschiedliche Fragestellungen sind auseinanderzuhalten, die dabei verfolgt werden koennen. Die eine ist durch die Suche nach (latenten) Ursachen fuer die beobachteten Merkmale gekennzeichnet (Kausalanalyse), die andere durch die Suche nach Gemeinsamkeiten (gelegentlich: Idealtypen) in den beobachteten Merkmalen (beschreibende Analyse i.e.S.). Beide haben deskriptiven Charakter (i.w.S.) insofern, als mit ihnen Vorgaenge bzw. Zustaende von Realitaet in den Blick genommen werden.

Bei Untersuchungen von der hier diskutierten Art bestehen die Ergebnisse meist in der Angabe von unabhaengig variierenden Ursachen oder Eigenschaften, die das Verhalten der beteiligten Personen oder ihre Interaktion erklaeren oder beschreiben sollen. Die einzelnen Aussagen erwachsen aus der Deutung von Faktorenloesungen, in denen die "gemessenen" Einzelmerkmale nach bestimmten formalen Prinzipien zu Teilgruppen zusammengefasst werden. Der Uebergang von der Teilgruppe zur Ursache bzw. zum uebergreifenden Aspekt wird in solchen Faellen durch eine kreative Leistung des Forschers ermoeglicht, die im wesentlichen darin besteht, unter moeglichst genauer Beachtung der quantitativen Modellkonstellation den durchgehend gemeinsamen "Zug" der Einzelmerkmale zu bestimmen. Er bezieht sich dabei auf die Bedeutung der in eine Teilgruppe (Faktor) einbezogenen Beobachtungswoerter und spricht ueblicherweise dann von einer "gut interpretierbaren" Faktorenloesung, wenn es ihm gelingt, in allen Teilgruppen (theoretisch) plausible Beziehungen zwischen ihren jeweiligen Merkmalen zu entdecken und sprachlich einfach zu artikulieren.

Auf die Rolle des Bedeutungsfehlers bei der Faktoreninterpretation wollen wir hier nicht eingehen (vgl. dazu 4,2,), Unsere Aufmerksamkeit richtet sich vielmehr jetzt darauf, ob und in welchem Ausmass die Faktorenloesungen voneinander abweichen, die auf der Analyse unserer "Originaldaten" einerseits und der verschiedenen Schaetzvarianten fuer das Ausmass des Bedeutungsfehlers andererseits beruhen. Es ist klar, dass voneinander abweichende Resultate auch zu unterschiedlichen Faktoreninterpretationen fuehren. Soweit aber diese Deutungen, die ja auf

der Basis von Beschreibungen i.e.S. stehen, selbst als Deskriptionen (eines komplexen und/oder latenten Merkmals) Geltung beanspruchen, koennen sie nicht zugleich wahr sein, es sei denn - was hier qua Design ausgeschlossen ist - , sie waeren nicht Beschreibungen desselben Sachverhalts.

Sollte unsere Untersuchung, wie nach den bisherigen Ergebnissen zu erwarten ist, zu ungleichen Faktorenloesungen fuehren, so ist daraus zu folgern, dass der Bedeutungsfehler bei der Erfassung der (Unterrichst-)Realitaet nicht in Kauf genommen werden kann; er fuehrt dann zu Beschreibungen, die nicht "mit der Wirklichkeit uebereinstimmen", die - aussagenlogisch gewendet - mit der wahren Beschreibung nicht aequivalent und also falsch sind.

Wir fuehren die Untersuchung in zwei Schritten durch. Im ersten betrachten wir die faktorenanalytische Seite des Problems, also Aspekte des formalen Modells, auf das die Beobachtungsergebnisse abgebildet werden. Dabei beschraenken wir uns auf die Eroerterung der Frage nach der Zahl der zu extrahierenden Faktoren und auf den Vergleich der Loesungen.<52> (Es ist in diesem Zusammenhang nicht erforderlich, zwischen den beiden Sprachmodellen (vgl. 3.4.2.3.) zu unterscheiden, da hier nicht die Relationen zwischen Begriffen eine Rolle spielen, sondern die Variationen der durch sie bezeichneten Entitaeten.) Im zweiten Schritt wird die materiale Problemseite beleuchtet; es geht dabei um den objektsprachlichen Inhalt der erziehungswissenschaftlichen Aussagen, die im Anschluss an die statistische Analyse formuliert werden.

In den Lehrbuechern zur Faktorenanalyse finden sich eine Reihe von Angaben darueber, wie auf rein formale Weise eine Entscheidung hinsichtlich der Faktorenzahl modellgerecht zu faellen ist (vgl. z.B. Harman 1976, Revenstorf 1976, Ueberla 1971). Man kann - aus praktischer Sicht - dabei zwischen zwei Gruppen unterscheiden, den praezisen, aber aufwendig zu berechnenden, ggf. testbaren Kriterien und den leicht zu handhabenden Faustregeln. Die letztgenannten scheinen nach wie vor verbreitet Anwendung zu finden, sich in der Forschungspraxis der Sozialwissenschaften zu "bewaehren", was nicht zuletzt daran liegen duerfte, dass sie tatsaechlich einen gewissen Zusammenhang mit den genauen Verfahren aufweisen.<53>

Fuer einen ersten Vergleich der vorliegenden Datengruppen ("Originaldaten" und Schaetzvarianten) ziehen wir die gaengigsten Faustregeln heran, um festzustellen, ob in faktorenanalytischer Sicht unter diesem Aspekt aehn-

liche Datenstrukturen vorliegen. Einige dieser Regeln sind direkt auf die Hauptkomponentenloesungen anzuwenden, einige zusaetzlich auch auf die Faktorenloesungen:

Basis Hauptkomponentenloesung
1. "Scree"-Test
2. Zahl der Eigenwerte groesser/gleich 1,0
3. Zahl der Eigenwerte, die mindestens 5% der Gesamtvarianz repraesentieren
4. Zahl der Eigenwerte in der Abfolge, die erforderlich sind, um mindestens 90% der Gesamtvarianz aufzuklaeren
5. Zahl der Eigenwerte in der Abfolge, deren Summe gerade noch unterhalb der Gesamtkommunalitaet bleibt

Basis Faktorenloesung
6. wie (2.)
7. wie (3.).

3-18 Kriterien zur Bestimmung der Zahl der zu extrahierenden Faktoren

Die Ergebnisse sind in Tab. 6 zusammengestellt. Sie weisen auch bei den robusteren Kriterien noch beachtliche Schwankungen ueber alle 13 Datengruppen hinweg auf, und zwar in den folgenden "Groessenordnungen":

| Kriterium | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Streuung | 1...8 | 5...12 | 3...8 | 12...21 | 3...6 | 2...4 | 1...3 |

3-19 Schwankungsbereiche der Faktorenzahl in verschiedenen Extraktionskriterien

Legt man zusaetzlich das sog. (m/2)-Kriterium an (Faktorenzahl kleiner als die Haelfte der Variablenzahl), so scheiden lediglich die nach Kriterium (4.) ermittelten Moeglichkeiten - bis auf eine bei Variante F (vgl. Tab. 6) - aus.

Um den Eindruck dieses Ueberblicks noch ein Stueck weit zu praezisieren, bilden wir - auf dieser prima-facie-Stufe - ein "Mass" fuer die Verschiedenheit der Datengruppen, in das fuer je zwei Datengruppen die Differenzen zwischen den Faktorenzahlen und die Zahl der

Uebereinstimmungen (0-Differenz) in jedem Kriterium eingehen. Um dabei auf der Plausibilitaetsebene (und dem Skalenniveau der Faktorenzahl!) zu bleiben, wird folgende Berechnung fuer jedes Paar von Datengruppen vorgenommen; (a) Summe der Uebereinstimmungen in der nach den sieben Kriterien sich ergebenden Faktorenzahl; (b) Summe der Differenzen in den nicht uebereinstimmenden Faktorenzahlen; (c) Summe aus Uebereinstimmungen (a) und Differenzen (b).

Bei 7 Kriterien waeren 7 Uebereinstimmungen (gemaess (a)) "ideal". Dieser Fall tritt nicht auf. Am "aehnlichsten" sind sich die Gruppen "Originaldaten" und Variante C, die 6 Uebereinstimmungen aufweisen. Laesst man fuer den Fall der Nicht-Uebereinstimmung eine Abweichung von 1 Faktor zu, nimmt man also "in kauf", dass die Faustregeln um plus/minus 1 um "ihren wahren" Wert schwanken, dann darf bei einer Gesamtbetrachtung aller Kriterien die Differenz zwischen Uebereinstimmungszahl (z.B. 5) und Kriterienzahl (7 Kriterien) durch den Umfang der Abweichungen (b) nicht ueberschritten werden (=2).

Werte fuer (c), die groesser als 7 sind, besagen also, dass - nach den Faustregeln - das verglichene Paar von Datengruppen als verschieden zu betrachten ist. Da die Kriterien nicht immer zu eindeutigen Ergebnissen fuehren (vgl. Tab. 6), haben wir zunaechst nach dem fuer die Uebereinstimmung (a) und die Differenz (b) jeweils "guenstigsten" Wert gesucht, also ein "grosszuegiges" Mass angelegt, das "Verschiedenheit" erst im extremen Fall signalisiert. Daneben sind auch die "unguenstigen" Werte herausgesucht worden. Das Resultat ist in Tab. 7 in der unteren Dreiecksmatrix zusammengefasst. Dort steht ein Kreuz (x) fuer Aehnlichkeit im guenstigsten und zwei Kreuze (xx) fuer Aehnlichkeit auch im unguenstigsten Falle. Bei aller Zurueckhaltung, mit der solche rezeptartig gewonnenen Informationen zu behandeln sind, kann doch gesagt werden, dass der Befund sich abzuzeichnen beginnt, den wir gleich anschliessend mit einem praezisen Instrumentarium erfassen wollen; die um den Bedeutungsfehler bereinigten Daten koennen zu anderen Ergebnissen fuehren als die urspruenglichen Daten; m.a.W.: die Analyse und Interpretation der unbehandelten Daten fuehrt moeglicherweise zu fehlerhaften Aussagen. Im "guenstigsten" Falle zeigen 6 der von uns versuchsweise errechneten Schaetzvarianten - die ja den gesamten Streuungsbereich des Bedeutungsfehlers einkreisen - "Aehnlichkeit" mit den Originaldaten, im unguenstigsten sind es 2 (von 12).

Auch die Schaetzvarianten untereinander scheinen nicht viel Gemeinsamkeit aufzuweisen. Von den insgesamt 66

Paarvergleichen sind 11 "positv", davon 3 auch im "unguenstigsten" Falle. Zwei der zuletzt genannten (Varianten G-J, G-L) unterscheiden sich vom urspruenglichen Datenset, d,h, sie fuehren zu anderen Ergebnissen als die Ausgangsinformationen.

Um der Gefahr einer unzulaessigen Interpretation zu entgehen, wenden wir uns unter dem gleichen Aspekt, naemlich der Frage nach der Uebereinstimmung zwischen Roh- und Schaetzwertkonfigurationen, einem praezise bestimmten Verfahren zu. Es beruht auf dem Vergleich der Faktorenmuster, die der mathematisch exakten Berechnung der Hauptkomponentenloesungen entstammen. Das Vergleichsverfahren besteht in der Rotation der Loesungsmatrizen auf maximale Uebereinstimmung; es liefert den sog. Tucker-Koeffizienten (-1<phi<+1), der ein Aehnlichkeitsmass (vergleichbar mit dem Korrelationskoeffizienten) darstellt. Wir gehen mit Tucker (zit. nach Harman 1976, 345) davon aus, dass Werte unter ,9000 gegen die Annahme einer Faktorenkongruenz sprechen.

Die Bestimmung des Aehnlichkeitskoeffizienten ist nur fuer Matrizen des gleichen Typs moeglich. Aus diesem Grunde kann eine exakte Aussage dann nicht gemacht werden, wenn Faktorenloesungen mit unterschiedlicher Faktorenzahl verglichen werden sollen. Diese Fragestellung wird in der Forschungspraxis zwar auftreten koennen, sie spielt aber in unserem Zusammenhang keine Rolle; es waere denkbar, dass fuer zwei verschiedene Schaetzvarianten gemaess gesicherten Extraktionskriterien (z.B. Bartlett-Test) unterschiedlich viele Faktoren bestimmt werden muessten und dass zwischen zwei solchen Faktoren Uebereinstimmung bestuende. Die jeweiligen Loesungen insgesamt wuerden sich dennoch unterscheiden und das bedeutet, dass sie zu inkompatiblen Deskriptionen fuehren. Unsere Aufgabe besteht demzufolge darin, diejenigen Faktorenmuster in die Pruefung einzubeziehen, die im Rahmen praktisch vorliegender bzw. zu erwartender Interpretationsspielraeume liegen.

Das Verfahren der Wahl ist fuer diesen Fall die Aehnlichkeitsrotation der Loesungsmatrizen aus der Hauptkomponentenanalyse, wie oben dargestellt. Es ergibt biasfreie Werte, die direkt gedeutet werden koennen.

In einer breiter angelegten Studie untersuchten wir alle 1- bis 6-Faktorenloesungen und bestimmten fuer sie ueber alle Datengruppen die Phi-Koeffizienten (also 6 X(13 X 12/2) = 468 Paarvergleiche). Bei dieser Vorgehensweise eruebrigt sich die praezise Bestimmung der zu extrahierenden Faktorenzahl vor allem dann, wenn sich bei den einzelnen Paarvergleichen fuer die 6 Loesungen

einheitliche Ergebnisse einstellen. Zugleich sind damit die praktisch zu erwartenden Faelle weitgehend abgedeckt; eine Interpretation von mehr als drei Faktoren ist bei Untersuchung von Lehrverhalten auch bei groesserer Variablenzahl (vgl. Ryans 1960) nicht ueblich. Zudem tendieren die in Tab.6 aufgelisteten Faustregel-Kriterien fuer die Bestimmung der Obergrenzen (Spalten 1 und 5) zu Faktorenzahlen um maximal vier.

Von den zwoelf Schaetzvarianten zeigen vier eine Aehnlichkeit mit dem Originalset. Dies entspricht der Vermutung, die aus den oben erfolgten Vorueberlegungen erwachsen war, wo sich bis zu sechs Mal Uebereinstimmung andeutete; sie bleibt fuer die Varianten C, E; H, I erhalten. Wie sich aus Tab. 7 (obere Dreiecksmatrix) ergibt, in die alle phi>,900 fuer 1,,,6-Faktoren-Loesungen eingetragen wurden, sind diese Gruppen erwartungsgemaess auch untereinander aehnlich. Darueberhinaus finden sich keine stabilen, ueber mehrere Loesungen hinweg hohen Uebereinstimmungen, d.h. die uebrigen acht Varianten (A, B, D, F; G, J, K, L) fuehren zu Ergebnissen, die weder mit den Originaldaten noch untereinander uebereinstimmen. Damit erhaelt man insgesamt neun prinzipiell moegliche, sich jedoch gegenseitig ausschliessende Resultate, von denen nur eines (die aehnlichen Varianten C, E; H, I) mit demjenigen kongruent ist, das man auf der Basis der urspruenglichen Informationen gewonnen haette.

Bei dem zuletzt genannten Fall handelt es sich um jeweils maessig rigorose Prozeduren in jeder der beiden Schaetzgruppen A-F und G-L (vgl. 3,4,2,3,). Schon weniger rigide Verfahren fuehren also zu Loesungen, die nicht mehr mit den in der Forschungspraxis ueblicherweise interpretierten Ergebnissen kompatibel sind. So muendet etwa die Untersuchung der Variante D in einen Befund, der sich - unter voellig anderen Modellvoraussetzungen - in Kap. 3,3, schon einmal eingestellt hat; die Rueckuebertragung der faktorenanalytischen Ergebnisse in die Lehr-Lern-Sprache fuehrt zu der Aussage, dass die beobachteten Merkmale in keinem Zusammenhang miteinander stehen, dass sie, m.a.W., unabhaengig voneinander variieren und somit nicht irgendeiner gemeinsamen Ursache zugeschrieben werden koennen. Ein Blick auf das Eigenwertdiagramm macht diesen Sachverhalt deutlich (Abb. 3-20; Eigenwerte der Korrelationsmatrix fuer die Schaetzvariante D). Man sieht, dass die Eigenwerte praktisch auf einer Geraden liegen, die von links nach rechts abfaellt. Diese Anordnung weist, wie Cattell (1966) gezeigt hat, darauf hin, dass der Analyse Zufallskorrelationen zugrundeliegen. Tatsaechlich erreicht die Einzelrest-Varianz hier die groessten Werte unter allen Bedeutungsfehlerschaetzungen (vgl. Tab. 6); sie

betraegt bei der 2-Faktoren-Loesung 91,4%.

Erwiese sich das bei dieser Variante eingeschlagene Bereinigungsverfahren als empirisch begruendet und bildeten die Ausgangsdaten eine zulaengliche Stichprobe aus dem Lehrverhalten von (Hochschul-)Lehrern, so muessten daraus weitreichende Konsequenzen gezogen werden. Es waere erforderlich, in der Theoriebildung zu diesem Komplex noch einmal "von vorne" zu beginnen und nach neuen Merkmalen zu suchen, die in eine Kausalrelation mit dem Unterrichtsergebnis einzustellen waeren. Man koennte ebenso auch den Schluss ziehen, dass die Aufklaerung der Ursache-Wirkungsbeziehungen in diesem Bereich - zumindest gegenwaertig und mit der gewuenschten Erklaerungsleistung - nicht bewerkstelligt werden kann, da die relevanten Effekte zu gering und/oder zu zahlreich sind.

Wir werfen, ohne an dieser Stelle in eine Diskussion dieser Frage einzutreten, jetzt noch einen Blick auf weitere denkbare Konsequenzen fuer den materiellen Gehalt von Aussagen auf der Basis derartiger Untersuchungsverfahren. Dazu vergleichen wir die varimax-rotierte 3-Faktoren-Loesung (lambda >= 5% der Gesamtvarianz) der Originaldaten mit Variante F, welche die hoechste Varianzaufklaerung erlaubt. Einen ersten Ueberblick zur Bestimmung des Interpretationsrahmens gibt die Zusammenstellung der absoluten und relativen Faktoranteile innerhalb der beiden Analysen (angegeben sind gerundete Einzelwerte, die bei der Summenbildung zu Abweichungen fuehren koennen);

| Faktor | Anteil | Org.dat. | Variante F |
|---|---|---|---|
| I | Eigenwert (lambda) | 5,97 | 12,52 |
| | % der Ges.varianz | 22,1 | 46,4 |
| | % d. aufgekl.Var. | 56,0 | 75,6 |
| II | Eigenwert | 3,18 | 2,34 |
| | % der Ges.varianz | 11,8 | 8,7 |
| | % d. aufgekl.Var. | 29,8 | 14,1 |
| III | Eigenwert | 1,51 | 1,70 |
| | % der Ges.varianz | 5,6 | 6,3 |
| | % d. aufgekl.Var. | 14,1 | 10,3 |
| Summen | Eigenwert | 10,65 | 16,56 |
| | % der Ges.varianz | 39,4 | 61,3 |
| | % d. aufgekl.Var. | 100,0 | 100,0 |

3-21 Kennwerte der 3-Faktoren-Loesungen von Orginaldaten und Variante F

Neben der erheblichen Differenz der beiden Eigenwertssummen fallen die unterschiedlichen Proportionen der Faktoren untereinander auf; waehrend bei den urspruenglichen Daten der erste Faktor gut die Haelfte der aufgeklaerten Varianz aufnimmt und sich die beiden anderen den Rest teilen, so dass ungefaehr eine Relation von 4;2;1 zwischen den Faktoren entsteht, traegt in Variante F der erste Faktor drei Viertel der aufgeklaerten Varianz und Faktor II und III sind - grob gesprochen - etwa gleich gewichtig; die Relation lautet hier ungefaehr 6;1;1. Absolut gesehen sind in beiden Analysen die Faktoren II und III etwa gleich repraesentiert; ihr Varianzanteil ist ziemlich gering.

Unabhaengig von der inhaltlichen Bedeutung der Faktoren besteht somit der erste Hauptunterschied zwischen beiden Analysen darin, dass sie zu "Hintergrundvariablen" fuehren, denen verschieden hohe Bedeutung zugesprochen werden muss. Nach Analyse der Originaldaten koennen ca. 40% der beobachteten Verhaltensvariationen auf Veraenderungen in drei "Hauptursachen" zurueckgefuehrt werden oder - nicht kausal, sondern im engeren Sinne "beschreibend" ausgedrueckt - ; es sind drei "Hauptzuege" im beobachteten Verhalten zu identifizieren, die zusammen ca. 40% des gesamten wahrgenommenen Verhaltens (unter den vorgegebenen Aspekten) wiedergeben. Fuer Variante F gelten die entsprechenden Aussagen mit einem Anteil von etwa 60%; d.h. dass nach ihr, wenn sonst keine Unterschie-

de bestuenden, den Faktoren eine staerkere Wirksamkeit bzw. - beschreibungsbezogen - hoehere Praedominanz zugeschrieben werden muesste.

Insbesondere die beiden ersten Faktoren unterscheiden sich in dieser Hinsicht erheblich. Das heisst aber - immer noch bei unterstellter gleicher Bedeutung - , dass nur eines der beiden Ergebnisse empirisch "richtig" sein kann. In der kausalen Deutung entsteht die Unvertraeglichkeit hinsichtlich der behaupteten Wirksamkeit, des Grades der "Koordiniertheit" des Verhaltens, in der beschreibenden hinsichtlich des "gemessenen" Auspraegungsgrades der Merkmale.

Wenden wir nun unsere Aufmerksamkeit der inhaltlichen Bestimmung der einzelnen Faktoren in den beiden Analysen zu, so zeigt bereits ein Blick auf die Ladungsmuster, dass in den beiden Analysen Sekundaervariablen von recht unterschiedlicher Bedeutung konstituiert werden (zur Bedeutung der Variablen vgl. Tab. 2):

| Orginaldaten III | Orginaldaten II | Orginaldaten I | <1> Variable | Variante F I | Variante F II | Variante F III |
|---|---|---|---|---|---|---|
| | ,72 | | 1 | ,76 | | |
| | ,82 | | 2 | ,77 | | |
| | ,81 | | 3 | ,77 | ,30 | |
| | | | 4 | ,58 | | |
| ,38 | ,63 | | 5 | ,83 | | |
| | ,86 | | 6 | ,72 | | |
| | | ,77 | 7 | | ,69 | |
| | | ,88 | 8 | | ,81 | |
| | | ,82 | 9 | | ,61 | |
| | | | 10 | ,45 | | |
| | | ,52 | 11 | ,30 | ,56 | |
| | | ,84 | 12 | | ,65 | |
| | | ,36 | 13 | ,70 | | |
| | | ,78 | 14 | | ,78 | |
| | | ,30 | 15 | ,77 | ,32 | |
| ,85 | | | 16 | ,79 | | |
| ,49 | | | 17 | ,83 | | |
| | | | 18 | | | ,99 |
| | | | 19 | ,76 | | |
| | | | 20 | ,73 | | |
| ,53 | | | 21 | ,72 | | |
| ,56 | | | 22 | ,71 | | |
| ,31 | | | 23 | ,86 | | |
| | | | 24 | | | ,83 |
| | ,36 | | 25 | ,87 | ,32 | |
| | | | 26 | ,79 | | |
| ,73 | | | 27 | ,76 | ,35 | |

<1> Alle Ladungen >,30 sind eingetragen.

3-22 Faktorenmuster der varimax-rotierten 3-Faktor-Loesungen von Orginaldaten und Variante F.

In den Originaldaten trennen die drei Faktoren im wesentlichen zwischen den Variablengruppen (Global-, Partialurteile, Haeufigkeitsbeobachtungen), waehrend in Variante F zunaechst naeherungsweise ein Hauptfaktor isoliert wird, dem ein zweiter folgt, der gewisse Aehnlichkeiten mit dem ersten der Originaldaten aufweist und ein dritter, der nur zwei Variablen aufnimmt. Der Versuch, diese Konstellationen auf den Begriff zu bringen, ist - wie gesagt - nicht ohne weiteres intersubjektiv zugaenglich. Deshalb braucht die von uns versuchte Deutung, was ihre rein sprachliche Seite betrifft, auch nicht geteilt zu werden. Entscheidend ist, dass jeder Versuch einer die jeweiligen Ladungsmuster repraesentierenden Zusammenfassung darauf abheben muss, die Differenzierung des Modells begrifflich widerzuspiegeln, ganz gleich, welche Formulierungen im einzelnen gebraucht werden.

Mit dieser Einschraenkung koennte man folgende, aus den beiden Analysen erwachsende Faktorenbezeichnungen vornehmen:

| Orginaldaten | Faktor | Variante F |
|---|---|---|
| Einfuehlung und Offenheit | I | Haltung und Auftreten |
| Menschliche Qualitaet | II | Einfuehlung und Behutsamkeit |
| Lebhaftigkeit | III | Lernsteuerung |

3-23 Interpretation der 3-Faktoren-Loesungen von Orginaldaten und Variante F

Die vergleichende Betrachtung der beiden Analysen macht deutlich, dass sich in ihnen zwei unterschiedliche Betrachtungsweisen konkretisieren, die auf ebenfalls verschiedene Konzeptualisierungen des Prozesses der Verhaltensgenese (kausale Deutung) bzw. der Verhaltensstruktur (beschreibende Deutung) zurueckgehen, Waehrend die Faktorisierung der Originaldaten den Eindruck vermittelt, es werde mit Hilfe der Beobachtungswoerter zwischen (drei) trennbaren Hauptaspekten differenziert, legt das Resultat aus Variante F eher die Vermutung nahe, dass die Beobachtungswoerter die Beachtung einer vorrangigen Dimension stimulieren, welche die beiden anderen teils mit umfasst (Faktor II; Variablen 3, 11, 15, 25, 27), teils als peripher erweist (Faktor III).

Ohne die Werte im einzelnen zu deuten, stellen wir hier die Ergebnisse zweier weiterer Schaetzvarianten dar, um zu zeigen, wie weit der "Spielraum der Unsicherheit" gedacht werden muss, der durch die fehlende Kenntnis der Wirkungsweise des Bedeutungsfehlers eroeffnet wird. Nach dem gleichen Kriterium wie fuer die beiden eroerterten Beispiele (lambda >= 5% der Gesamtvarianz) sind aus Schaetzung J und L zwei Faktoren zu extrahieren, fuer die folgende Ladungen errechnet werden:

| Variante J | | <1> | Variante L | |
|---|---|---|---|---|
| II | I | Variable | I | II |
| | | 1 | | |
| | | 2 | | |
| | | 3 | | |
| | -,30 | 4 | | |
| | | 5 | | |
| | ,46 | 6 | | |
| | | 7 | | ,42 |
| | ,32 | 8 | | ,51 |
| | | 9 | | ,60 |
| | | 10 | | |
| | | 11 | | ,61 |
| | | 12 | | ,66 |
| | | 13 | | -,40 |
| | | 14 | | ,44 |
| | | 15 | | |
| ,85 | | 16 | ,80 | |
| ,41 | ,41 | 17 | ,56 | |
| | ,43 | 18 | | |
| | | 19 | | |
| | | 20 | | |
| ,53 | | 21 | ,47 | |
| ,56 | ,34 | 22 | ,68 | |
| ,31 | | 23 | ,40 | |
| | ,70 | 24 | ,32 | |
| | ,43 | 25 | ,42 | |
| ,72 | | 26 | ,72 | |
| | | 27 | | |

<1> Alle Ladungen >,30 sind eingetragen.

3-24 Faktorenmuster der varimax-rotierten 2-Faktoren-Loesung von Variante J und Variante L

Es zeigt sich zunaechst eine gewisse Aehnlichkeit zwischen Faktor II in J und Faktor I in L, die darueberhinaus auch mit Faktor III der Originaldaten zu bestehen scheint. Ein genaueres Studium der Ladungen foerdert aber auch hier Nuancierungen zutage, die fuer eine aufzubauende oder zu pruefende Theorie nicht folgenlos bleiben koennten. Deutlicher fallen die uebrigen Unterschiede zwischen diesen beiden und den oben dargestellten Ladungskonfigurationen aus. Sie setzen sich fort in der Gegenueberstellung der Kennwerte von J und L einerseits und Originaldaten und F andererseits:

| Faktor | Variante J | Anteil<1> Eigen- | % der Ges.var. | % d. aufgekl.Var. |
|---|---|---|---|---|
| I | Kooperative Aktivitaet | 3,01 | 11,1 | 58,9 |
| II | Rhetorische Differenziertheit | 1,36 | 5,0 | 41,1 |
| | Summen: | 4,37 | 16,2 | 100,0 |
| | Variante L | | | |
| I | Differenziertheit des Ausdrucks | 3,22 | 11,9 | 61,4 |
| II | Einfuehlsame Zurueckhaltung | 2,02 | 7,5 | 38,6 |
| | Summen: | 5,23 | 19,4 | 100,0 |

<1> Gerundete Einzelwerte, die bei Summierung zu Abweichungen fuehren koennen.

3-25 Kennwerte und Interpretation der 2-Faktoren-Loesungen der Varianten J und L

Damit ergibt sich, dass zwar Aehnlichkeiten zwischen den einzelnen Ergebnissen gefunden werden koennen, dass sich jedoch zugleich Differenzen im Hinblick auf ihre theoretischen Implikationen sowie ihren deskriptiven Gehalt in wesentlichen Punkten einstellen, obwohl ein und derselbe beobachtete Sachverhalt zugrunde liegt. Je nach Fragestellung muessen aufgrund der verschiedenen Analysen unterschiedliche, untereinander unvertraegliche Aussagen formuliert werden. Dass diese theoretisch und auch prak-

tisch nicht trivial sind, ergibt sich aus den Konsequenzen, die fuer unseren Fall der Erfassung von Lehrverhalten hier stichwortartig angedeutet werden koennen.

In der kausalen Bedeutung betreffen sie nichts geringeres als die psychologische Rekonstruktion der Persoenlichkeitsstruktur und die sozialpsychologische Deutung des Einflusses von Lehrpersonen auf Lernende. Praktisch unterscheiden sie sich in den Antworten, die auf Fragen nach den vom Lehrer anzustrebenden bzw. gezielt zu variierenden Verhaltensweisen, nach deren Erwerbbarkeit, nach der curricularen Gestaltung von Lehrerstudiengaengen usf. impliziert werden. In der beschreibenden Deutung der Faktorenanalysen liegt die theoretische Relevanz bspw. in der abweichenden Bestimmung von Merkmalen, die in der (Unterrichts-)Realitaet wahrnehmbar differenziert und "gemessen" werden koennen. Eine praktische Folge davon waere eine Kontroverse ueber die Kriterien, unter denen etwa Pruefungslehrproben im Studium oder Referendariat beurteilbar sind.

## 3,5, Der empirische Gehalt von Ratinginformationen

### 3,5,1, Die Relation von Sprachstruktur und Realitaetsbeschreibung

Die Beziehung zwischen Bedeutungsueberschneidung und Informationsgehalt von Ratingdaten laesst sich in einer einfachen Relation ausdruecken: je hoeher das Ausmass der Bedeutungsgemeinsamkeit einer gegebenen Wortmenge, desto geringer der mit ihr erfassbare empirische Gehalt. Vor allem bei den - hier so bezeichneten - Beobachtungsurteilen sorgt diese Beziehung, wie wir gesehen haben, fuer erhebliche Informationseinbussen. Diese kommen jedoch nicht unmittelbar in den quantitativen Aussagen zum Ausdruck, die unter Verwendung solcher Woerter ueber die Realitaet gemacht werden. Der "empirische Kern" in den Daten ist ueberlagert von "Fehlerschichten", die das Qualitaetsmerkmal "reliabel, valide und objektiv" zwar suggerieren, aber nicht erfuellen. Vor der Auswertung ist daher fuer solche Informationen auf jeden Fall eine "Reinigungsprozedur" indiziert, die u.a. die Beseitigung von sprachlichen "Ablagerungen" gewaehrleisten muss. Geschieht dies nicht, so muss - nach allem, was unsere bisherige Analyse ergab - damit gerechnet werden, dass die ermittelten Ergebnisse eher ueber (verbreitete) Sprechgewohnheiten als ueber beobachtete Sachverhalte informieren.

Um eine Vorstellung ueber das relative Ausmass der sprachlichen "Verunreinigung" zu gewinnen, wird in diesem Abschnitt untersucht, ob es moeglich ist, bei Kenntnis der semantischen Struktur der Beobachtungswoerter gewisse Voraussagen darueber zu machen, welche Ergebnisse sich bei deren Anwendung einstellen - unabhaengig davon, welche Merkmale der tatsaechlich beobachtete Sachverhalt aufweist. Je weniger Voraussage und empirischer Befund divergieren, desto "maechtiger" ist die "Sprachschicht" ueber der eigentlich interessierenden empirischen Aussage, d.h. desto geringer ist der empirische Gehalt der zur Verfuegung stehenden Information.

Die quantitative Durchfuehrung der Analyse erfolgt in zwei Abschnitten. Zuerst fragen wir, wie "gut" die in den Beobachtungsratings ausgedrueckte Realitaetsbeschreibung aus den ebenfalls vorliegenden Wortaehnlichkeiten haette vorhergesagt werden koennen. Dabei gehen wir, weil ausreichend grosse Stichproben zur Sprachanwendung jedes einzelnen Beobachters nicht vorliegen (jedes Wort wurde von jedem Beobachter ja nur einmal "angewendet"), wie in Kap. 3,3, von der theoretischen Annahme

eines Sprachkollektivs aus. Sodann pruefen wir unter Verwendung der publizierten Werte, wie die Verhaeltnisse bei der Vorhersage von Untersuchungsergebnissen anderer Arbeiten liegen. In erster Linie fassen wir dabei die Studien von Mueller-Wolf/Fittkau (1971) und Mueller-Wolf (1977) ins Auge, zu denen bei der Konstruktion unseres Designs ja eigens eine gewisse Parallelitaet hergestellt wurde. Die dazu ermittelten Resultate diskutieren wir knapp auch im Hinblick auf weitere Publikationen auf diesem Gebiet.

Die Frage nach den Beziehungen zwischen Sprachstruktur und Beschreibungen kann fuer unsere Untersuchung so formuliert werden: Wie aehnlich sind die Aussagen, die aufgrund der Kenntnis der Bedeutungsueberschneidungen zwischen den Beobachtungswoertern formuliert werden koennen, mit denjenigen, die aus den (unkorrigierten) Unterrichtsratings hervorgehen? In der Sprache unseres formalen Modells lautet sie: In welchem Masse besteht Kongruenz zwischen den Ladungsmustern aus der Faktorenanalyse von Wortaehnlichkeitsmatrix und ("empirischer") Korrelationsmatrix?

Wie in Kap. 3.3. beschrieben, interpretiert man zum Aufbau der Wortaehnlichkeitsmatrix die Mittelwerte der aus den Paarvergleichen stammenden Aehnlichkeitsangaben (0,..9) als Winkelmasse, deren Kosinus (vgl. Gleichung [3] und [4]) wertgleich mit dem Korrelationskoeffizienten ist. Da Aehnlichkeitsangaben fuer die "Haeufigkeitsbeobachtungen" nicht vorliegen (vgl. 3-3 und Tab. 2), setzen wir fuer sie den Wert Null ein; dies entspricht einer besonders vorsichtigen Deutung.<54> Damit entsteht eine Kovarianzmatrix von der in Tab. 3 dargestellten Art; sie unterscheidet sich von ihr nur durch die zusaetzliche Aufnahme der "Haeufigkeitsbeobachtungen". Von analoger Struktur ist die "empirische" Korrelationsmatrix (27 X 27), die z.T. (15 X 15) ebenfalls in Tab. 3 enthalten ist.

Berechnet man nun fuer beide Matrizen eine (exakte) Hauptkomponentenanalyse und vergleicht - in Anlehnung an die in Kap. 3.4. diskutierte Bestimmung der Faktorenzahl - die 3-Faktoren-Loesungen ueber eine Aehnlichkeitsrotation (Tuckers Phi), so erhaelt man die gesuchte Massangabe fuer den fraglichen Zusammenhang. Fuer die Bildung der Wortaehnlichkeitsmatrix stehen drei Ausgangsmasse des Mittelwerts zur Verfuegung: arithmetisches Mittel, Median, Modus; somit lassen sich drei Aehnlichkeitskoeffizienten ermitteln. Darueberhinaus ist es sinnvoll, dieses Mass auch fuer die "Beobachtungsurteile" alleine zu bestimmen (15 X 15-Matrix), weil hierbei die ungesicherten Annahmen ueber die (Un-)Aehnlich-

keit der "Haeufigkeitsbeobachtungen" wegfallen. Es ergeben sich folgende Werte:

| Vergleichsbasis | Berechnungsbasis arith. Mittel | Median | Modus |
|---|---|---|---|
| Alle Woerter (Var. 1 - 27 | ,627 | ,619 | ,576 |
| Beobachtungsurteile (Var. 1 - 15 | ,718 | ,720 | ,705 |

3-26 Aehnlichkeitskoeffizienten fuer den Vergleich von Sprach- und Beschreibungsstruktur

Wir muessen uns im vorliegenden Kontext auf die direkte Interpretation dieser Koeffizienten beschraenken, da die inhaltliche Deutung der Faktorenmuster, wie sie in (3.4.3.2.) durchgefuehrt wurde, nicht zu vergleichbaren Aussagen fuehrt.<55> Man kann mit Blick auf die obigen Werte nicht sagen, dass im Wissen von einer Sprachstruktur das Wissen darueber, was in ihr ausgedrueckt werden kann, bereits vollstaendig enthalten sei. Diese Behauptung stuende auch im Widerspruch zu unserer alltagspraktischen Erfahrung mit der Sprache. Doch laesst sich nicht leugnen, dass gewisse, in besonderen Faellen sogar betraechtliche "Ueberdeckungen" zu konstatieren sind. Gerade die Untersuchungen zum Lehrverhalten sind es, die zu einem grossen Teil zu den "besonderen" Faellen zu rechnen sind (vgl. Kap. 1.). Fuer sie gilt - wenn unsere Ergebnisse auch nur groessenordnungsmaessig zutreffen - , dass mit dem eingesetzten Wortinventar nur ein begrenztes Mass an faktischer Merkmalsvariation ausgedrueckt zu werden vermag; der "Rest" versinkt - um es etwas ueberspitzt zu sagen - im Sumpf der Bedeutungsueberschneidungen.

Es ist plausibel, dass die Werte bei der Beschraenkung auf die Urteilswoerter insgesamt hoeher liegen als bei Zugrundelegung aller Woerter, da ihre Anwendung mit der Beachtung einer groesseren Anzahl von Einzelmerkmalen einherzugehen pflegt; die Menge der zum Urteil gehoerenden Merkmale ist groesser als bei den "Haeufigkeitsbeobachtungen" und (bei einer endlich grossen Zahl der ueberhaupt in die menschliche Beobachtung einbezogenen Merkmale) die Ueberschneidungswahrscheinlichkeit folglich hoeher. (Wir weisen der Klarheit wegen noch einmal

darauf hin, dass wir - wie bereits eingangs eroertert - dabei den moeglichen Wertungsgehalt solcher Ausdruecke ausser Betracht lassen und insofern einen sicherlich mit Recht bestreitbaren Vertrauensvorschuss leisten.)

Man kann die tabellierten Phi-Koeffizienten so aehnlich wie Korrelationskoeffizient en interpretieren (mit dem Unterschied, dass ihrer Berechnung keine Abweichungswerte zugrundeliegen, was sie tendenziell niedriger ausfallen laesst). Mit einiger Vorsicht lassen sich die Ergebnisse - unter Beachtung der obengenannten sprachtheoretischen Voraussetzungen - daher auch so zusammenfassen: Fuer den Fall der Einbeziehung aller Woerter ist das faktorenanalytisch ermittelte Ergebnis einer empirischen Untersuchung - bei Kenntnis der Sprachstruktur - zu 33% (Modus), 48% (Median) oder 39% (arithm. Mittel) voraussehbar <56>; bei der - haeufig geuebten - Beschraenkung auf die "interessanteren" Urteile belaufen sich die entsprechenden Werte sogar auf 50% oder 52% (Median und arithm. Mittel).

### 3.5.2. Die Vorhersagbarkeit empirischer Befunde aus der Struktur von Erhebungsinstrumenten

Zum Vergleich unserer "Prognosen" mit den Ergebnissen der empirischen Arbeit von Mueller-Wolf/Fittkau (1971) und Mueller-Wolf (1977) wurden zunaechst fuer alle drei Sprachaehnlichkeitsmatrizen (Basis: arithm. Mittel, Median, Modus) wie dort fuer die Korrelationsmatrizen je eine 3- bzw. 2-Faktoren-Loesung berechnet (mit der Einheitsvarianz als Kommunalitaetenschaetzung und anschliessender Varimax-Rotation). Sodann erfolgte eine visuelle Anpassungsschaetzung, da fuer beide Untersuchungen nicht die vollstaendigen Ladungsmuster publiziert sind.

Zwei Beschraenkungen beeintraechtigen von vornherein die "Prognoseleistung" unserer Daten: zum einen legten wir den Pbn., wie berichtet, stets nur ein Wort bzw. einen Ausdruck zum Vergleich mit anderen und zur Anwendung vor; wir benutzten also keine Mehrfachbenennungen, wie dies z.T. in den beiden zitierten Arbeiten der Fall ist (vgl. 2-1); zum anderen wurden in unsere Erhebung nicht alle Items von dort aufgenommen, was sich vor allem auf die Chance des Zustandekommens von bestimmten Faktoren auswirkt.

Fuer die Gegenueberstellung mit der 3-Faktoren-Loesung (1977) erweist sich die Mittelwertsberechnung der Aehn-

lichkeitsangaben als geeignetste (vgl. Tab. 8a). Sie reproduziert die Faktoren des Lehrverhaltens qualitativ (Abfolge und Variablenzuordnung) fast perfekt. Nur zwei Items (4; "AUTOKRATISCH", 9; "TRIFFT ENTSCHEIDUNGEN SELBST") laden nicht auf den erwarteten Faktor. Wegen des insgesamt hoeheren Werteniveaus in der zugrundeliegenden "Korrelationsmatrix" ergeben sich bei den Sprachaehnlichkeitsfaktoren noch zusaetzliche Ladungen, die aus den empirischen Daten - deren Niveau regelmaessig niedriger liegt - nicht "destillierbar" sind. Was den Vergleich der Ladungshoehen in den einzelnen Variablen betrifft, so kann nicht von einer besonders guten Anpassung gesprochen werden. Allerdings waere dies auch kaum zu erwarten gewesen.

Ergaenzt man jedoch die in Tab. 8a fehlenden Ladungsangaben durch den Wert 0,0 und betrachtet insofern nur die tatsaechlich auch in die Interpretation einbezogenen Variablen, so fuehrt die Zielrotation der Ladungsmatrix "Sprachstruktur" auf die Ladungsmatrix aus der Mueller-Wolf-Erhebung zu einem beinahe Kongruenz signalisierenden Aehnlichkeitskoeffizienten von R =,85.

Zu vergleichbaren Ergebnissen fuehrt die Inspektion der 2-Faktoren-Loesungen, in die wegen der Beschraenkung auf Ladungen >= ,70 weniger Variable aufgenommen werden koennen (vgl. Tab. 8b). Auch hier stellt sich weitgehende qualitative Konkordanz ein; sieben von zehn Items laden erwartungsgemaess auf Faktoren, die auch in der Rangfolge uebereinstimmen.

Um die dem Wortfeld inhaerente Struktur, die sich in den Faktorenanalysen herausschaelt, einer "Stabilitaetspruefung" zu unterwerfen, haben wir die Aehnlichkeitsmasse zusaetzlich einer minimum-spanning-tree-Clusteranalyse unterworfen. In diesem Verfahren wird zunaechst diejenige Anordnung der Elemente (Woerter) gesucht, fuer die in der Ebene drei Merkmale erfuellt sind: (1) alle als Punkte dargestellten Elemente werden durch Strecken verbunden; (2) diese Linien schliessen sich nirgends zum "Kreis"; (3) die Summe der Streckenlaengen bildet ein Minimum. Von diesem Graphen mit (15-1=) 14 "Enden" (kuerzester Dendrit) werden einzelne Strecken entfernt und dadurch Punktecluster abgetrennt. Wir suchten diejenige 3-Cluster-Aufteilung, fuer welche die Summen der Aehnlichkeiten pro Cluster ueber alle drei Cluster hinweg maximal sind (vgl. Calinski/Harabasz, zit. nach Wishart 1978, 68-71).

Diese Prozedur unterscheidet sich betraechtlich vom faktorenanalytischen Modell und kann insofern als formales (!) "Aussenkriterium" gelten. Sie reproduziert die dort

gefundene Struktur bei Verwendung der arithmetischen Mittelwerte als Basis dennoch recht gut (Abb. 3-27; Kuerzester Dendrit 3-Cluster-Loesung). Bei dieser Beurteilung ist zu beruecksichtigen, dass im Unterschied zur Faktorenanalyse hier eine dichotome Zuordnungsentscheidung fuer jedes Item gefaellt wird; Variablen, die in der Faktorenanalyse auf mehrere Faktoren laden, werden nur einmal eingruppiert. Maximiert man in diesem Spielraum die Trefferquote, so ergibt sich, dass nur zwei Items nicht erwartungsgemaess zugeordnet werden (6: "ANREGEND", 9: "TRIFFT ENTSCHEIDUNGEN SELBST"). Damit erweist sich die Vermutung als erhaertet, dass der durch Ratingurteile vermittelte Blick auf die Realitaet weitgehend durch die semantischen Beziehungen zwischen den verwendeten Woertern (bzw. Ausdruecken) verstellt wird.

In den bereits zitierten Quellen von Mueller-Wolf/Fittkau und Mueller-Wolf werden eine Reihe von inhaltlichen Uebereinstimmungen mit anderen Untersuchungen festgestellt (1971, 171; 1977, 101-102). Auf sie ist in dem Masse, in dem die Parallelitaet jeweils besteht, das soeben gewonnene Resultat uebertragbar. (Man beachte vor allem die Bezuege zu der umfangreichen Arbeit von Keil und Piontkowski (1973).)

Eine Studie von Coffman (1954), in die fast zweitausend Studenten einbezogen wurden, sollte hier noch erwaehnt werden. Sechs (Scale No. 5, 9, 12, 14, 15, 19) seiner 19 Items (278-280) haben eine naeherungsweise Entsprechung in den von uns verwendeten Variablen ("ANREGEND", Ziff. 11, 7, ,15, 3, 5 in Tab. 2). Stellt man die bei Coffman berichtete 4-Faktoren-Loesung (281) neben unsere 2-Faktoren-Loesung der Wortaehnlichkeiten (Tab. 8b), so zeigt sich auch hier - bis auf Scale No. 5 - Uebereinstimmung in der Zuordnung.

Zu einem aehnlichen Ergebnis gelangt der Vergleich mit einer Untersuchung von Harari/Zedeck (1973), der allerdings nicht auf der Modellebene praezisiert werden kann. Sie gelangen unter Verwendung von Ratingurteilen abschliessend zu der Auffassung, dass nur mit diesem Verfahren - in der Hand der Studenten - eine effiziente und wissenschaftlich zulaengliche Feststellung des Lehrverhaltens von Hochschullehrern moeglich sei (265).

Wir koennen diesem Urteil aufgrund der Ergebnisse unserer Analyse nicht folgen. Sie zeigt, dass die Sprachstruktur, was ja auch von Mueller-Wolf untersucht und festgestellt wurde (1977, 104-105), eine auffallende Aehnlichkeit mit der "Realitaetsstruktur" aufweist. Dieser Befund laesst den von Mueller-Wolf gezogenen Schluss, dass damit tatsaechlich eine Aehnlichkeit zwi-

schen zwei verschiedenen Entitaeten konstatierbar sei, aber nicht zu, Vielmehr handelt es sich, wie wir gezeigt haben, um ein Artefakt, das dadurch zustande kommt, dass - in einem gelaeufigen Bilde ausgedrueckt - die Welt durch gefaerbte Glaeser betrachtet und daher in der Weise fuer farbig gehalten wird, wie diese Brille es suggeriert. Nicht das erkenntnistheoretische Problem, sondern der methodische Mangel, der damit bezeichnet ist, kann gemildert, wenn nicht sogar beseitigt werden.

# 4. Ueberlegungen zum Status und zur Praezisierung von Unterrichtsforschung

## 4.1. Grenzen und Gefahren der Anwendung des faktorenanalytischen Modells in systematischer Sicht

Die semantische Unklarheit von Beschreibungen des Lehrverhaltens ist beileibe kein Problem, das erst in letzter Zeit ins Bewusstsein der Forscher gedrungen waere. In dem Mitte der sechziger Jahre erschienenen Band zur "Psychologie der Erziehungsstile", der Beitraege namhafter Psychologen enthaelt, findet man immer wieder Hinweise auf die mangelnde begriffliche Praezision in einschlaegigen Untersuchungen (vgl.z.B. Ahrens, 32 ff.; Weinert, 95; Diskussion zu den Ausfuehrungen von R. Tausch, 220 ff.). In seinem Vorwort charakterisiert der Herausgeber den "Markt der Veroeffentlichungen" zu diesem Problemkreis als "ziemlich chaotisch"; die "Techniken zur Messung des Erziehungsverhaltens (seien) ...nach den Testguete-Kriterien zumeist unzureichend" (Herrmann 1966, 8).

Soweit wir heute diesen "Markt" ueberblicken, muessen wir bekennen, dass das - sicherlich zutreffende - Urteil von damals nicht revisionsbeduerftig ist. Die Erziehungswissenschaft hat sich auf ein unter solchen Gesichtspunkten zu fuehrendes "fachliches Gespraech" (Herrmann, ebd.) mit den interessierten Psychologen nicht so eingelassen, dass als Konsequenz eine methodische und vor allem auch theoretische Entwicklung zu hoeherer Praezision zu konstatieren waere.

Hier ist nicht der Ort, die Gruende fuer diesen Gang der Dinge zu ventilieren. Wir rechnen es aber zu unserer Aufgabe, nicht bei der Feststellung der Unzulaenglichkeiten stehenzubleiben, sondern die Frage nach der Moeglichkeit ihrer Beseitigung zu eroertern. Zu diesem Zweck erscheint es nuetzlich, das Hauptergebnis unserer Untersuchung noch einmal vor Augen zu fuehren. Es besteht im wesentlichen in der Feststellung, dass die fehlende (oder doch wenigstens: nicht kontrollierte) Ueberschneidungsfreiheit der Bedeutungen von Beobachtungswoertern durch die Abbildung auf ein quantitatives formales Modell (z.B. Faktoranalyse) nicht etwa behoben, sondern via Rueckuebertragung in die objektsprachlichen theoretischen Saetze hineingetragen wird - dies zum Nachteil fuer deren empirischen Gehalt: sie "beschreiben" in Wirklichkeit nicht das beobachtete Geschehen, sondern sind Ausdruck der (personenspezifischen) semantischen Struktur der eingefuehrten Woerter. Die haeufig vorkom-

mende Wendung in Arbeiten zur Unterrichtsforschung, man sei auf so und so viele "gut interpretierbare" Faktoren gestossen, signalisiert in solchen Faellen daher weniger den Eintritt eines Erkenntnisfortschritts als vielmehr die (uneingestandene oder unbewusste) Projektion des eigenen Sprachverstaendnisses; fraegt man unabhaengig von den Daten einer Untersuchung nur nach der sachlichen Zusammengehoerigkeit von verbreiteten Beobachtungswoertern, so erhaelt man wahrscheinlich - eine informelle Studie weist diese Tendenz auf - weitgehend dieselbe Zuordnung, wie sie sich in der Faktorenanalyse ergibt.

Wir kommen damit auf einen Punkt zu sprechen, der sich an frueherer Stelle (vgl. 3.4.3.2.) bereits als problematisch erwiesen hat. Sucht man mit einer Faktorenanalyse naemlich nicht lediglich nach "Stellvertretervariablen", also nach Merkmalen, deren Erhebung das Messen anderer Merkmale ersparen soll, so muss man - wie ebenfalls angedeutet - Klarheit darueber haben, welche theoretische Funktion die "Interpretation" der Analyseergebnisse uebernimmt. Im Falle der Suche nach Kausalbeziehungen wird man i.d.R. ausschliesslich Merkmale einbeziehen, die in dieser Hinsicht auf ein und derselben Stufe stehen, also zum Ursache- oder Wirkungsbereich gehoeren (fuer mehrstufige Kausalanalysen sind Modelle aus der Korrelations-/ Regressionsstatistik angemessener). Das bedeutet aber, dass die Interpretation zur Benennung der (hypothetischen) Wirkung oder der Ursache fuehren muss, unabhaengig davon, ob das gesamte Unterfangen oder ein Teil davon auf der operationalen oder der Konstruktebene abgewickelt wird (vgl. dazu auch Herrmann 1973).

Der Status der hier untersuchten und vergleichbarer weiterer Woerter und Ausdruecke aus der Unterrichtsforschung ist in dieser Hinsicht unklar. Sie werden als Bezeichnungen fuer interne Strukturen der Lehrperson ("Dispositionen", Attitueden), fuer ihr (sichtbares i.e.S.) Verhalten und fuer die Eindrucksbildung bei den Hoerern (auch: Beobachtern!) verwendet; hin und wieder bleibt ihre Deutung voellig im Dunkeln.

Im "einfacheren" Falle der Beschreibung (i.e.S.; vgl. 3.4.3.2.) tritt die Frage nach der Unterscheidung von theoretischer und Beobachtungssprache in den Vordergrund. Ob sie im Sinne einer strikten Trennung oder eines stufenweisen Uebergangs gesehen wird, kann an dieser Stelle ausser Betracht bleiben. In jedem Falle gilt, dass die integrative Leistung des faktorenanalytischen Modells zugleich auch die Interpretationsrichtung bestimmt; sie fuehrt stets zur theoretischen bzw. Konstruktebene hin, niemals umgekehrt.<57> Das bedeutet eine gewisse Unuebersichtlichkeit fuer Analysen, in die

Termini unterschiedlichen Inferenzniveaus eingehen. Wir kommen auf diesen Punkt gleich noch zu sprechen. Das semantische Problem, das wieder zum Bedeutungsfehler hinfuehrt, haengt, wie wir zuerst zeigen wollen, direkt damit zusammen.

Zunaechst ist darauf hinzuweisen, dass der Uebergang von operational gefassten Beobachtungswoertern zu (theoretischen) Konstrukten (Subsumtionsrichtung) wie - aus logischen Gruenden - auch der umgekehrte Weg (Operationalisierungsrichtung) einen willkuerlichen, also regelungs- und konsensbeduerftigen Akt darstellt. Hielte man sich theoretisch streng an die Beschreibungsfunktion (i.e.S.), so lieferte die Faktorenanalyse der Werte aus der Erfassung operationalisierter Merkmale Hinweise fuer Moeglichkeiten zur Bildung theoretischer Begriffe. Dabei ist es offensichtlich, dass gegebene Faktorenloesungen solche Begriffsbildungen nicht erzwingen, sondern lediglich "nahelegen" (vgl. z.B. die oben eingefuehrten Formulierungen: 3-23 und 3-25).

Hier stellt sich nun heraus, dass das "Finden" eines Faktors nicht ohne Rekurs auf die Bedeutung der Variablenbezeichnungen vonstatten geht. Der gesuchte "gemeinsame Zug" erwaechst oft aus der (interpretierten) Bedeutungsueberschneidung der gemeinsam ladenden Variablennamen. Es ist plausibel, dass die Interpretation um so leichter faellt, je groesser der (mengentheoretische) Durchschnitt in den zu den Woertern gehoerenden Merkmalsmengen (samt Gewichts- und Haeufigkeits-/Intensitaetsdimension) ausfaellt. Das heisst aber nichts anderes, als dass viele bedeutungsaehnliche Beobachtungswoerter, die gemeinsam zu einem Faktor gehoeren, diesen als "gut interpretierbar" erscheinen lassen. Wie wir gezeigt haben, stehen die Chancen dafuer, dass solch ein Ergebnis eintritt, guenstig; Bedeutungsueberschneidungen gewaehrleisten im formalen Modell hohe Korrelationen, die allerdings nicht aus Merkmalskovariationen am Forschungsobjekt resultieren. So scheint der Forscher tatsaechlich in ein Dilemma zu geraten. Versucht er naemlich, die semantischen Artefakte zu vermeiden, kommt er in die Schwierigkeit, "unzusammenhaengende" Aspekte (wie Sprechlautstaerke, Satzkonstruktion, Kopfbewegung, Mundwinkelstellung, Bewegung im Raum etc.) "zusammenfassen" zu muessen.

Bevor eine Loesung dieses Problems gesucht wird, ist ein weiterer Gesichtspunkt zu erwaehnen. Da das faktorenanalytische Modell die Strategie der (Rest-)Varianzmaximierung pro Faktor verfolgt, kann bei Kenntnis von Bedeutungsueberschneidungen der Beobachtungswoerter und der damit zu erwartenden Korrelationen von vornherein das

Entstehen eines Faktors "provoziert" werden. Je groesser naemlich die Zahl von derart zusammengehoerigen Woertern ist, die als Beobachtungsaspekte in ein Erhebungsinstrument aufgenommen werden, desto sicherer und klarer werden sie sich zu einem Faktor vereinigen. Jedoch kommt es auch hier auf das Sprachverstaendnis des jeweiligen Forschers an. Eine gegebene Faktorenloesung kann von verschiedenen Individuen unterschiedlich interpretiert werden. Wo der eine noch Deutungsmoeglichkeiten "sieht", mag ein anderer die Loesung schon verwerfen und eine geringere Faktorenzahl extrahieren (vgl. Ueberla 1971, 137).

In diesem Kontext koennen wir vom Auftreten des Meta-Bedeutungsfehlers sprechen: er entsteht bei der Interpretation von zu Gruppen zusammengefassten Beobachtungswoertern, die hinsichtlich ihrer "Gemeinsamkeiten" von verschiedenen Personen unterschiedlich beurteilt werden, und dies unabhaengig davon, wie der oder die Beobachter solche Beziehungen bei der Datenerfassung tatsaechlich realisierten. Die letzteren betrachten die Realitaet und bringen bei deren Beschreibung den objektsprachlichen Bedeutungsfehler hervor, die erstgenannten untersuchen die Beschreibungswoerter, also sprachliches Material, und erzeugen dabei den metasprachlichen Bedeutungsfehler.

Das in vielen Untersuchungen zur Unterrichtsforschung (vgl. z.B. Tausch 1966) replizierte Resultat uebereinstimmender Faktoren buesst vor diesem Hintergrund viel von seinem Charakter empirischer Bestaetigung ein; es kann durchaus auch unter dem Aspekt der "Kongenialitaet" der beteiligten Forscher im Hinblick auf ihr "Sprachgefuehl" gedeutet werden.

Angesichts dieser Moeglichkeit waere es sicherlich nicht nur reizvoll, sondern auch methodologisch erhellend, wenn man bei bevorstehenden Untersuchungen mit faktorenanalytischen Auswertungsplaenen den zustaendigen Forschern vor der Datenanalyse die Aufgabe stellte, die einbezogenen Items in alle denkbaren, ihnen sinnvoll erscheinenden Kombinationen zu gruppieren. Die uebrigen waeren damit von vornherein ausgeschlossen, wuerden also im Falle ihres Auftretens zu Konstruktinnovationen fuehren koennen; theoriefremde ad-hoc-Deutungen waeren ausgeschlossen. Es ist durchaus vorstellbar, dass eine empirische Erhebung der wissenschaftlichen Arbeitsstrategien auf diesem Gebiet Kenntnisse verschaffen wuerde, die ihrerseits den Bedarf zur Erhoehung der Sprachpraezision unterstreichen koennten.

Mit diesen Ueberlegungen sollte zugleich auch deutlich

geworden sein, dass die wuenschenswerte Praezisierung keinesfalls erreicht wird, wenn man sowohl operationale als auch theoretische Termini in die Beobachtung aufnimmt, wie dies haeufig geschieht und in dieser Arbeit aus methodologischen Gruenden wiederholt wurde. Die dabei gewonnenen Gruppierungen erwachsen aus dem Sprachgebrauch der Beobachter und sind insofern empirisch fundiert. Sie bieten also keine Loesung des Operationalisierungs- bzw. Subsumtionsproblems, das seiner "Natur" nach ja analytisch ist. So waere es unzulaessig und irrefuehrend, aufgrund einer Faktorenanalyse, in die eine derartige Woerterkombination eingegangen ist, eine "Behauptung" ueber die "echte" Bedeutung bestimmter Termini aufzustellen. Diese muss vielmehr stets durch einen (sprach-)normierenden Akt ("Definition") der beteiligten Forscher "gestiftet" werden.

## 4.2. Terminologische und konzeptuelle Aspekte des Beschreibungsproblems

Wer sich darauf einlaesst, analytisch-empirische Kausalforschung im Realitaetsbereich Unterricht zu betreiben, muss als erstes anstreben, moeglichst exakte und reliable Beschreibungen von Merkmalen, die ihrerseits in eine hypothetische Ursache-Wirkung-Konstruktion eingeordnet sind, zu beschaffen. Betrachtet man unter diesem Gesichtspunkt noch einmal die Itembeispiele, die im Rahmen dieser Arbeit verwendet wurden, so wird man nach allem, was die sozialpsychologische Forschung zu diesem Komplex an Erkenntnissen gewonnen hat (vgl. z.B. Irle 1975 und zum speziellen Problem; Hofer 1969), sagen koennen, dass die sogenannten Beobachtungsurteile zum Bereich der Wirkung von Lehrverhalten gehoeren und daher nicht als dessen Beschreibung gedeutet werden duerfen.<58>

Es ist nicht auszuschliessen, dass ueber die Begriffe "Verhaltensstil", "Erziehungsstil", "Interaktionsstil", die auf den ersten Blick als geeignete sprachliche Rekonstruktionen fuer die Interpretation von statistisch gebildeten Kunstvariablen (naemlich Faktoren) erscheinen, sich ein systematischer Fehler in die Hypothesenbildung zur Unterrichtsforschung einschleicht. Sollen mit ihnen naemlich Zusammenfassungen von operationalen Merkmalen aus einer Kausalstufe intendiert sein, so waere damit ein Ziel ins Auge gefasst, das in dieser Forschungsstrategie gar nicht enthalten ist. Faktorisierungen dienen dazu, latente Merkmale ("Dispositionen" etc.) im formalen Modell zu rekonstruieren. Diese sind aber als "Verursacher" (seltener als Wirkungen) dessen anzu-

sehen, was der unmittelbaren Beobachtung zugaenglich ist.

Eine nicht-kausale Interpretation von streng operational erhobenen Merkmalsdaten des Verhaltens muss daher als Sinnlosigkeit abgewiesen werden. Die dennoch durchgefuehrten Versuche scheinen darauf hinzuweisen, dass in den Interpreten eine geisteswissenschaftliche Tradition lebendig ist, die ihnen die Vorstellung von der wesensmaessigen Existenz eines unkonkreten Verhaltenscharakters plausibel erscheinen laesst.

Es ist klar, dass dieser Gedanke, auch wenn er im modernen Gewande des Stilbegriffs auftritt, nur unter ontologischen Voraussetzungen in die Unterrichtsforschung eingebracht werden koennte, die mit denjenigen einer analytisch-empirischen Konzeption nicht vereinbar sind. Zugleich haette man damit eine alternative wissenschaftstheoretische Position eingenommen, in deren methodologischem Konzept eine komplexe statistische Modellanalyse keinen Platz hat.

Die begriffliche Naehe des Verhaltensstils etwa zu dem des kognitiven Stils mag eine terminologische Parallelisierung zwar plausibel erscheinen lassen. Tatsaechlich handelt es sich jedoch nur beim zweiten Ausdruck um die Bezeichnung einer Entitaet, die als (theoretisches) Konstrukt in die Kausalanalyse eingebracht werden kann; sie bezieht sich auf einen als Ursache oder als Wirkung gedachten Komplex in der latenten Persoenlichkeitsstruktur, an dessen Status selbstverstaendlich auch die Tatsache nichts aendert, dass seine Auspraegung - wie ueberall in der Testpsychologie - ueber Verhaltensindikatoren zu erfassen versucht wird (vgl. Beck 1978, 647).

Stilbegriffe auf der im strengen Sinne sichtbaren Verhaltensebene gewinnen demzufolge nur einen Sinn, wenn sie eine "additive" Funktion von der Art uebernehmen, dass durch sie praezise angegebene Haeufigkeits- bzw. Intensitaetsauspraegungen mehrerer Einzelmerkmale angegeben werden. Sie bleiben damit auf derselben Sprachebene wie die von ihnen eingeschlossenen Termini und dienen dann der Sprachvereinfachung. Ihre Bedeutung erhalten sie durch Definition (nicht: Subsumtion oder Operationalisierung) und weisen qua Aequivalenzrelation daher auch keinerlei "Bedeutungsueberschuss" gegenueber den im Definiens enthaltenen Praedikaten auf.

Die Einfuehrung so konstruierter Begriffe in die Unterrichtsforschung ist im gegenwaertigen Stadium moeglicherweise verfrueht. Sie erscheint uns erst dann sinnvoll, wenn es gelungen ist, kausal relevante, im Hin-

blick auf bestimmte Explananda erklaerungskraeftige Merkmalskombinationen des Verhaltens zu identifizieren. Ob man aber ueberhaupt in dieser Richtung die Forschung vorantreiben sollte, waere zuvor zu diskutieren. Wir koennen diese Frage hier nicht verfolgen und beschraenken uns daher auf den Hinweis fuer eine Alternative, Sie bestuende in der methodischen (nicht: theoretischen!) Abkehr von der Stimulus - Organismus - Response - Strategie (S-O-R) samt ihren Erweiterungen zugunsten eines Organismus (i) - Verhalten (i) - Organismus (j) - Konzepts (O(i)-V(i)-O(j)), in dem gerade das sichtbare (Lehr-)Verhalten forschungspraktisch umgangen, also nicht erfasst wird. Erhoben werden dagegen die (interessierenden) O-Zustaende der beteiligten Personen (i,j) und dies - selbstverstaendlich - gemaess dem bewaehrten S-O-R-Prinzip. Auf diese Weise liesse sich das schwierige Problem der Suche nach relevanten Verhaltensmerkmalen umgehen; das Kausalglied "Verhalten" in der Ursache-Wirkungskette wuerde zur "neglected box" erklaert, also forschungsmethodisch (vorlaeufig) uebersprungen.

## 4.3. Das Lehrverhalten und seine Wahrnehmung als Objekte der Unterrichtsforschung - Eine kausalanalytische Re-Interpretation von "Verhaltensratings"

Es ist jetzt moeglich, unsere Daten noch einmal in einem neuen Licht zu betrachten. Trennt man sie naemlich gemaess der Kausalidee in eine Ursache- und eine Wirkungsgruppe (Haeufigkeitsbeobachtungen (Variablen 16-27) und Beobachtungsurteile (Variablen 1-15)), so laesst sich die Frage stellen, welche Erklaerungsleistung mit ihnen zu erbringen ist. Nach der soeben skizzierten Terminologie fuer Forschungskonzepte liegt hier ein V(i)-O(j)-Ansatz vor mit den bereits diskutierten Einschraenkungen hinsichtlich der Praezision in der theoretischen Isolierung und der methodischen Erfassung der Merkmale. Nicht voellig auszuschliessen ist auch ein Konfundierungseffekt zwischen den beiden Merkmalsgruppen, die ja aus derselben Datenquelle stammen; allerdings braucht bei dem relativ hohen Operationalisierungsgrad der V-Aspekte (Variablen 16-27) diese Gefahr wohl nicht allzu hoch veranschlagt werden.

Problematisch wird dagegen erneut die Bedeutungsfrage. Unterstellt man naemlich, dass die sogenannten Beobachtungsurteile in Wahrheit Informationen ueber die Eindrucksbildung (vgl. Witte 1966) enthalten, so waere jetzt nach der Dimensionalitaet dieser persoenlichkeitsinternen Teilstruktur zu fragen und festzustellen, wel-

che Urteilsskalen in welcher Weise auf sie bezogen sind, Wuerde man, um es an einem Beispiel zu veranschaulichen, von einem dreidimensionalen Raum der Personwahrnehmung ausgehen und waeren unsere 15 Urteile auf sie im Verhaeltnis 12;0:3 faktorenrein verteilt, so gelangte man bei der statistischen Analyse (beispielsweise Berechnung der kanonischen Korrelationskoeffizient en zwischen V- und O-Merkmalen) zu entsprechend verzerrten Ergebnissen, Alle multivariaten statistischen Modelle, die auf einer Faktorisierung der Datenmatrizen beruhen, sind hinsichtlich der Merkmalsgewichtung sensibel und muessen daher in dieser Hinsicht vorsichtig gehandhabt werden,

Hier gilt es, eine scharfe Trennungslinie zwischen Kausalanalyse und Persoenlichkeitsstrukturforschung zu ziehen, Die methodischen Bedenken, die wir soeben formuliert haben, betreffen den zuerst genannten Fall, Sie beziehen sich auf das, was man den theoretischen Bedeutungsfehler nennen koennte, Dieser aeussert sich in der statistisch relevanten Gewichtung der in eine Untersuchung einbezogenen Variablen, Er entsteht dadurch, dass unterscheidbare Merkmale (z,B, Wahrnehmungsdimensionen, Attitueden) durch unterschiedlich viele "reine" oder unausgewogen komplexe Messungen erfasst und in die Analyse eingebracht werden,

Etwas anders liegen die Dinge dagegen, wenn es darum geht, die Dimensionalitaet einer (Teil-)Struktur der Persoenlichkeit ueberhaupt erst zu bestimmen, Dies ist naemlich nicht zuletzt eine Entscheidungsfrage, die im Kontext metatheoretischer Kriterien (z,B, Fruchtbarkeit, Einfachheit) reflektiert werden muss, Die Diskussion ihrer methodischen Problematik faellt jedoch nicht in den Rahmen unserer Thematik,

Die in unserer Analyse vorgenommene Datenkorrektur stuetzt sich auf Informationen ueber den Sprachgebrauch der Sprecher, Hat man sich nach den bisher diskutierten Einwaenden dazu entschlossen, die Ratings zum Lehrverhalten nicht als Beschreibungen, sondern als (projizierte) Indikatoren fuer Zustaende von internen Merkmalen zu deuten, so gewinnen auch die Sprachaehnlichkeitsangaben einen neuen Status, Sie koennen dann als indirekte Auskunft ueber die Dimensionalitaet des sprechereigenen Wahrnehmungsraums betrachtet werden, Der Rang der Aehnlichkeitsmatrix der Urteilswoerter, die zur Grundlage der Berichtigung (in Kap, 3,4,) gemacht wurde, gibt in dieser Sicht eine annaehernde Schaetzung der Dimensionalitaet der individuellen Situationswahrnehmung ab (mit den ebenfalls in Kap, 3,4, diskutierten formalen Einschraenkungen), Eine Wertekorrektur in Orientierung an dieser Matrix erzeugt demnach tendenziell dimensionsrei-

ne Daten <59>, die wegen der interindividuellen Unterschiede natuerlich bei der Aggregation diese Eigenschaft verlieren ( wie die in 3,4,3,2, wiedergegebenen Ladungsmatrizen auch zeigen),

Sieht man von den diskutierten (sowie den lediglich erwaehnten) Einschraenkungen ab (insbesondere interindividuelle Sprachunterschiede, Praezision der V-Variablen-"Messung"), so waere eine globale Kausalanalyse zunaechst statistisch als kanonische Korrelationsberechnung anzulegen, Beruecksichtigt man ausserdem intervenierende differentielle Aspekte, wie z,B, Leistungsmotivation, "autoritaere Einstellung", Einstellung zum Studium, Geschlecht, so muss eine ein- bis mehrfaktorielle multiple Kovarianzanalyse unter Beruecksichtigung moeglicher Wechselwirkungen durchgefuehrt werden,

Es gehoert nicht zu den Zielsetzungen dieser Arbeit, derartige Analysen durchzufuehren und zu interpretieren, Sie liegen vielmehr im Zustaendigkeitsbereich kausal-empirischer Unterrichtsforschung, Dass sie jedoch unter Beruecksichtigung des Semantikproblems in der zuletzt diskutierten Form des theoretischen Bedeutungsfehlers zu signifikanten Ergebnissen fuehren koennen, zeigt eine uebersichtsmaessige informelle Globalanalyse der bereinigten Daten, Sie umfasst die "sprachstufenuebergreifend" korrigierten Urteilswerte (Basis: individuelle Aehnlichkeitsmatrizen der Variablen 1-15: Schaetzvarianten A-F) als "abhaengige Variablen", die nicht veraenderten (Original-)Werte aus den "Haeufigkeitsbeobachtungen" als "unabhaengige Variablen" und ggf, die erwaehnten differentiellen Merkmale als qualitative "Faktoren" (vgl, 3,1,4,), Obwohl in den unveraenderten Ausgangsdaten keine ueberzufaelligen Effekte aufgefunden werden koennen, stellen sie sich in einigen der bereinigten Datensaetze ein (z,B, bei $p<,05$; $R=,454$ im Korrelationsdesign fuer Variante D; $R=,442$ im Kovarianzdesign mit Faktor "autoritaere Einstellung" fuer Variante F),

Weder fuer die Datenerhebung noch zur Interpretation dieser Effekte ist die oben kritisierte theoretische (und damit faktorielle) Zusammenfassung der Verhaltensmerkmale erforderlich, Sie bleiben vielmehr unveraendert multivariat "erhalten", Ihre Bedeutung fuer das Zustandekommen der Ergebnisse laesst sich aus der Inspektion der entsprechenden Koeffizientenmatrizen (Beta-Gewichte, Kontraste) ermitteln,

Dass die aufgeklaerten Varianzen mit ca, 20% relativ niedrig bleiben, mag zunaechst an den besprochenen methodischen Ungenauigkeiten des Untersuchungskonzepts liegen, das ja nicht speziell fuer diese Fragestellung

entworfen wurde. Sodann ist zu beruecksichtigen, dass nur eine geringe Zahl der tatsaechlich diskriminierbaren und wahrscheinlich relevanten Verhaltensmerkmale einbezogen wurde. Schliesslich kann nach einer immer mehr verbreiteten plausiblen Auffassung angesichts der Komplexitaet der Unterrichtssituation und der grossen Zahl von Faktoren, die sie beeinflussen, nicht damit gerechnet werden, dass einige wenige Hauptursachen eine befriedigende Erklaerungsleistung zu erbringen vermoegen. Gerade deshalb kommt es darauf an, Verfahren zu konstruieren, die - ohne Artefakte zu erzeugen - sensibel und praezise genug sind, um die Vielzahl der vergleichsweise schwachen Kraefte zu erfassen, die unsere Wirklichkeit praegen.

# T A B E L L E N .

"Systematische" Fehler bei der Erhebung von Daten über Personen mit Schätzskalen[1]

| lfd. Nr. | BEZEICHNUNG[2]: Kurzbeschreibung (Literaturhinweise[3]) | Entstehungsort/-zeit (relativ zum "Meßinstrument") | Persönlichkeitsbereich |
|---|---|---|---|
| 1 | BEDEUTUNGSFEHLER: Unkontrollierte Übereinstimmung der (sprachlich gefaßten) Beobachtungsaspekte | input | eher intellektuell |
| 2 | HINTERGRUNDEFFEKT: Nicht ausschaltbare variable Einflüsse des Person-Umfeldes (Faßnacht 1979, 42) | | |
| 3 | INTERVENTIONSFEHLER: Modifikation der "abgespeicherten" Wahrnehmungen, wenn zwischen Beobachtung und Urteilsabgabe Zeit verstreicht (ggf. auch Modifikation der bereits (intern) gefaßten Schätzurteile; in diesem Fall: output-Fehler) (Faßnacht 1979, 151; Friedrichs/Lüdtke 1971, 70) | | |
| 4 | ZEITRAUMFEHLER: Unkontrollierte Zeitspanne der Beobachtung (Faßnacht, 1979, 151) | | |
| 5 | BEOBACHTUNGSABFOLGEEFFEKT ("primary-recency-effect"): Unterschiedliche Reihenfolge der Ereignisse führt zu Variationen in den Einschätzungen (v.Cranach/Frenz 1969, 281) | Verarbeitung/ Entscheidung | |
| 6 | GENERALISIERUNGSFEHLER: Ein (für der Beobachter) "eindruckvolles" Ereignis wird hinsichtlich seiner Bedeutsamkeit/Häufigkeit falsch "hochgerechnet" und beeinflußt so das Schätzurteil mit falschem "Gewicht" (Faßnacht 1979, 55) | | |
| 7 | LAIENFEHLER: Beobachter werden zur Abgabe von Urteilen "gezwungen" unter Aspekten, die sie (aufgrund von Unwissenheit/fehlender Sprachkompetenz) nicht klar identifizieren können (Friedrichs/Lüdtke 1971, 65) | | |
| 8 | "LOGIKFEHLER" ("logical error"): Abgabe des Schätzurteils aufgrund vermuteter (kausaler/funktionaler) Zusammenhänge zwischen beobachteten Ereignissen und (davon unterscheidbaren) zu beurteilenden Merkmalen (Guilford 1954, 279; Faßnacht 1979, 54: "theoretischer Fehler") | | |
| 9 | NACHBARSCHAFTSEFFEKT ("proximity error"): Beeinflussung durch räumliche/zeitliche Nähe von nicht unter den Beobachtungsaspekt fallenden Ereignissen (Guilford 1954, 280) | | |
| 10 | "REDUKTIVE KODIERUNG": Unkontrollierte Zuordnung von Beobachtung(sbegriffen) zu Urteilsdimensionen (Klassen, abstrakte/theoretische Begriffe) (Faßnacht 1979, 151; Friedrichs/Lüdtke 1971, 53, 61, 85) | | |
| 11 | SKALENNIVEAUFEHLER: Unterstellung einer unzutreffenden Variationseigenschaft des Gegenstandes (z.B. unterstellt: Intervallskala, tatsächlich: Ordinalskala) (Faßnacht 1979, 98, 150; Friedrichs/Lüdtke 1971, 78) | | |
| 12 | VORWISSENSFEHLER: Beeinflussung der Wahrnehmung/Schätzung durch Kenntnisse über die soziale Position/Rolle etc. der zu beurteilenden Person (Faßnacht 1979, 55) | | |
| 13 | "HOFEFFEKT" ("halo-effect"): Beeinflussung der Urteile durch eine generelle Einstellung gegenüber der zu beobachtenden Person (Guilford 1954, 279) | | eher emotional |
| 14 | "KONTRASTFEHLER" ("contrast error"): Neigung des Beobachters, sich selbst als Vergleichsmaßstab einzubringen (Beachtung von eigenen/nicht eigenen Eigenschaften) (Guilford 1954, 279-280; Walter 1977, 91; nach Faßnacht (1979, 55): "Kontaktfehler") | | |
| 15 | NACHSICHTIGKEITSEFFEKT ("Effekt der Milde/Großzügigkeit", "error of leniency"): Soziale Beziehungen zwischen Beobachter und Beobachtetem beeinflussen die Urteile (ergänzend Sumaski (1977, 68): Effekt der "Strenge") (Guilford 1954, 278) | | |
| 16 | WERTUNGSEFFEKT: Einfluß der (subjektiven) Wertschätzung des Beobachters für die Fragestellung überhaupt, den (jeweiligen) Beobachtungsaspekt (vgl. Ziff. 14) und das Ergebnis (i.S. sozialer Erwünschtheit: Walter (1977, 91); generell (also auch etwa Forscherintentionen): Faßnacht (1979, 56)) | | |
| 17 | "ZENTRALE TENDENZ" ("central tendency"): Neigung zur Abgabe von "durchschnittlichen" Urteilen (ergänzend Sumaski (1977, 68): Extremisierungstendenz) (Guilford 1954, 278) | output | |
| 18 | INTERPERSONELLE DIFFERENZEN (der Beobachter) in 1 - 17 | | |
| 19 | INTRAPERSONELLE DIFFERENZEN (eines Beobachters zu verschiedenen Zeitpunkten) in 1 - 18 | | |
| 20 | INTERPERSONELLE BEEINFLUSSUNG (der Beobachter untereinander) vor allem in 1 - 3, 8, 12, 13, 15 - 17 | | |

1) Die Aufstellung ist nicht mit dem Anspruch der Vollständigkeit versehen. Die Gruppierung erfolgt gemäß dem Abfolge-/Ortkriterium (vgl. Sp. 3) von Singleton (1972), innerhalb der Gruppe nach dem (vermuteten) Persönlichkeitsbereich der Fehlerentstehung, innerhalb der Persönlichkeitsbereiche unsystematisch ("Bezeichnungsalphabet"). Die gesamte Einteilung beruht auf Plausibilität; sie ist im vorliegenden Zusammenhang nicht von systematischer Bedeutung. Im übrigen vgl. Abschn. (2.).

2) Eingeführte (zitierte) Bezeichnungen sind in Anführungszeichen gesetzt.

3) Die Hinweise beschränken sich i.d.R. auf die Angabe von Sekundärliteratur.

TABELLE 2

LISTE DER GEORDNETEN ITEMS UND SKALENBEISPIELE

(IN KLAMMERN UEBEREINSTIMMUNGSWERTE AUS MUELLER-WOLF/FITTKAU 1971, 170,)

I, GLOBALURTEILE
- 1, EMOTIONAL WARM (,56)
- 2, NACHSICHTIG
- 3, LIBERAL
- 4, AUTOKRATISCH (,68)
- 5, QUALITAET ALS HOCHSCHULLEHRER (,70)
- 6, SOZIALINTEGRATIV (,55)

II, PARTIALURTEILE
- 7, DOZENT RESPEKTIERT DIE STUDENTEN ALS PARTNER UND PERSOENLICHKEITEN (,79)
- 8, DOZENT ZEIGT FUER DIE SCHWIERIGKEITEN DER STUDENTEN VERSTAENDNIS (,76)
- 9, BETEILIGT STUDENTEN AN ENTSCHEIDUNGEN (,73)
- 10, BEFEHLEND (,72)
- 11, ENTSPANNT (,69)
- 12, ER BRINGT ZUM AUSDRUCK, NOCH EIN LERNENDER ZU SEIN (,72)
- 13, DIE STUDENTEN ERMUTIGEND (,65)
- 14, SENSITIV FUER DIE GEFUEHLE DER STUDENTEN (,65)
- 15, SELBSTSICHER (,64)
- (- ANREGEND (,80); VGL, ANM, 37)

III, HAEUFIGKEITSBEOBACHTUNGEN
- 16, BEIM SPRECHEN LAUTSTAERKE WECHSELN
- 17, BEIM SPRECHEN GESTEN MIT DEN ARMEN MACHEN
- 18, FRAGEN AN STUDENTEN STELLEN
- 19, WAEHREND DER VERANSTALTUNG AN EINER STELLE STEHEN (SITZEN) BLEIBEN
- 20, GLEICHBLEIBENDE SPRACHLICHE WENDUNGEN BENUTZEN
- 21, LAUT SPRECHEN
- 22, GESICHTSAUSDRUCK VERAENDERN
- 23, BEIM SPRECHEN PAUSEN MACHEN
- 24, ARBEITSANWEISUNGEN AN STUDENTEN GEBEN
- 25, BEIM SPRECHEN ZU DEN STUDENTEN HINSCHAUEN
- 26, STIMMLAGE (-HOEHE) WECHSELN
- 27, KOERPER BEIM SPRECHEN RUHIG HALTEN

SKALENBEISPIELE

ALLE SKALEN SIND NEUNSTUFIG, SIE UNTERSCHEIDEN SICH JE NACH FRAGESTELLUNG IN DER BESCHRIFTUNG:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| WORTVERGLEICH INNERHALB GR, | KEINE AEHNL,-KEIT | | SEHR UNAEHNLICH | | | | SEHR AEHNLICH | | GLEICHE BEDEU-TUNG |
| ZW, GRUPPEN | KEIN ZUSHG, | SEHR SCHWACH | | | | | | | SEHR ENG |
| BEOBACHTUNG MIT I, | GAR NICHT | SCHWACH AUSGEPRAEGT | | | | | STARK AUSAUSGEPRAEGT | | |
| MIT II, | GAR NICHT | IN SEHR GE-RINGEM MASSE | | | | | | | IN HOHEM MASSE |
| MIT III, | NIE | SEHR SELTEN | | | | | | | SEHR HAEUFIG |
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

TABELLE 3

EXPERIMENTELLE(UNTERE MATRIX) UND ARTIFIZIELLE KORRELATIONEN (R(EX),R(A)) FUER RATING - URTEILE<1>

| VAR. NR. <2> | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | MIT | 81 | 61 | 31 | 68 | 79 | 78 | 88 | 62 | 49 | 65 | 67 | 83 | 91 | 56 |
| EMOTIONAL | MED | 87 | 61 | 26 | 70 | 78 | 84 | 93 | 62 | 29 | 63 | 67 | 87 | 94 | 51 |
| | MOD | 87 | 64 | 00 | 76 | 94 | 98 | 98 | 00 | 00 | 50 | 94 | 98 | 98 | 00 |
| 2 | | | 79 | 26 | 67 | 78 | 79 | 90 | 59 | 40 | 71 | 48 | 73 | 85 | 53 |
| NACHSICHTIG | 66 | 1. | 86 | 00 | 77 | 77 | 77 | 94 | 50 | 17 | 77 | 34 | 77 | 87 | 34 |
| | | | 94 | 00 | 77 | 77 | 94 | 98 | 50 | 00 | 94 | 00 | 77 | 94 | 00 |
| 3 | | | | 27 | 71 | 83 | 93 | 87 | 95 | 61 | 67 | 48 | 68 | 64 | 53 |
| LIBERAL | 65 | 73 | 1. | 22 | 77 | 88 | 97 | 90 | 98 | 32 | 62 | 39 | 71 | 65 | 49 |
| | | | | 00 | 94 | 94 | 98 | 98 | 1. | 00 | 50 | 00 | 77 | 64 | 00 |
| 4 | | | | | 36 | 30 | 56 | 45 | 50 | 96 | 39 | 35 | 48 | 46 | 82 |
| AUTOKRATISCH | -15 | -17 | -19 | 1. | 34 | 23 | 36 | 31 | 25 | 99 | 22 | 21 | 34 | 30 | 92 |
| | | | | | 17 | 00 | 17 | 17 | 00 | 1. | 00 | 00 | 00 | 00 | 98 |
| 5 | | | | | | 81 | 91 | 93 | 91 | 50 | 72 | 74 | 93 | 85 | 88 |
| QUALITAET | 56 | 58 | 57 | 00 | 1. | 84 | 96 | 96 | 95 | 35 | 78 | 79 | 96 | 90 | 92 |
| | | | | | | 93 | 98 | 98 | 98 | 00 | 87 | 93 | 98 | 98 | 98 |
| 6 | | | | | | | 96 | 94 | 95 | 50 | 78 | 57 | 89 | 89 | 59 |
| SOZ.-INT. | 66 | 68 | 73 | -17 | 62 | 1. | 97 | 97 | 97 | 28 | 84 | 54 | 94 | 93 | 55 |
| | | | | | | | 98 | 98 | 98 | 00 | 93 | 00 | 98 | 98 | 50 |
| 7 | | | | | | | | 97 | 98 | 22 | 85 | 50 | 87 | 90 | 54 |
| PARTNER | 22 | 30 | 27 | -04 | 30 | 34 | 1. | 98 | 98 | 15 | 87 | 42 | 91 | 92 | 48 |
| | | | | | | | | 98 | 98 | 00 | 87 | 00 | 94 | 94 | 34 |
| 8 | | | | | | | | | 82 | 23 | 84 | 53 | 93 | 94 | 50 |
| VERSTAENDNIS | 15 | 22 | 18 | -15 | 23 | 26 | 80 | 1. | 85 | 15 | 89 | 51 | 95 | 95 | 47 |
| | | | | | | | | | 86 | 00 | 94 | 00 | 98 | 98 | 00 |
| 9 | | | | | | | | | | 18 | 73 | 47 | 83 | 79 | 55 |
| ENTSCHEIDUNGEN | 00 | 10 | 12 | -00 | 16 | 16 | 65 | 69 | 1. | 07 | 75 | 46 | 88 | 85 | 52 |
| | | | | | | | | | | 00 | 76 | 00 | 94 | 94 | 50 |
| 10 | | | | | | | | | | | 20 | 25 | 25 | 23 | 67 |
| BEFEHLEND | 16 | 17 | -01 | 12 | 09 | 00 | -17 | -09 | -06 | 1. | 13 | 19 | 19 | 17 | 69 |
| | | | | | | | | | | | 00 | 00 | 17 | 00 | 93 |
| 11 | | | | | | | | | | | | 44 | 80 | 85 | 54 |
| LERNEND | 22 | 18 | 25 | -07 | 30 | 26 | 52 | 52 | 5, | -06 | 1. | 39 | 85 | 91 | 51 |
| | | | | | | | | | | | | 34 | 94 | 94 | 50 |
| 12 | | | | | | | | | | | | | 53 | 63 | 70 |
| ENTSPANNT | 14 | 14 | 16 | -07 | 31 | 22 | 28 | 35 | 21 | 12 | 16 | 1. | 50 | 61 | 70 |
| | | | | | | | | | | | | | 00 | 64 | 86 |
| 13 | | | | | | | | | | | | | | 89 | 59 |
| ERMUTIGEND | 03 | 13 | 12 | -07 | 18 | 17 | 59 | 69 | 79 | 01 | 53 | 26 | 1. | 90 | 61 |
| | | | | | | | | | | | | | | 98 | 77 |
| 14 | | | | | | | | | | | | | | | 54 |
| GEFUEHLE | 27 | 20 | 15 | -12 | 24 | 30 | 62 | 74 | 62 | 02 | 49 | 69 | 34 | 1. | 50 |
| | | | | | | | | | | | | | | | 34 |
| 15 | | | | | | | | | | | | | | | |
| SELBSTSICHER | 10 | 08 | 11 | -00 | 27 | 15 | 21 | 30 | 17 | 10 | 09 | 20 | 63 | 24 | 1. |

<1> DIE ANGEGEBENEN WERTE SIND GERUNDET. DER FUEHRENDE PUNKT IST WEGGELASSEN.

<2> MIT: BASIS ARITH. MITTEL, MED: MEDIAN, MOD; MODUS (VGL. 3,3.).

TABELLE 4

RESTKORRELATIONEN ("WAHRE" WERTE) FUER RATING - URTEILE <1>

OBERE MATRIX: MIT (1,ZEILE), MED (2,ZEILE) <2>
UNTERE MATRIX: MOD <2>

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 EMOTIONAL | 1. | / | 22 | / | / | / | / | / | / | / | / | / | / | / | / |
| | | / | 21 | -08 | / | / | / | / | / | 01 | / | / | / | / | / |
| 2 NACHSICHTIG | / | 1. | / | / | / | / | / | / | / | / | / | / | / | / | / |
| | | | / | -17 | / | / | / | / | / | 01 | / | / | / | / | / |
| 3 LIBERAL | / | / | 1. | / | / | / | / | / | / | / | / | / | / | / | / |
| | | | | / | / | / | / | / | / | / | / | / | / | / | / |
| 4 AUTOKRATISCH | -17 | -17 | -18 | 1. | / | / | / | / | / | / | / | / | / | / | / |
| | | | | | / | / | / | / | / | / | / | / | / | / | / |
| 5 QUALITAET | / | / | / | / | 1. | / | / | / | / | / | / | / | / | / | / |
| | | | | | | / | / | / | / | / | / | / | / | / | / |
| 6 SOZ.-INT. | / | / | / | / | / | 1. | / | / | / | / | / | / | / | / | / |
| | | | | | | | / | / | / | / | / | / | / | / | / |
| 7 PARTNER | / | / | / | / | / | / | 1. | / | / | / | / | / | / | / | / |
| | | | | | | | | / | / | / | / | / | / | / | / |
| 8 VERSTAENDNIS | / | / | / | / | / | / | / | 1. | / | / | / | / | / | / | / |
| | | | | | | | | | / | / | / | / | / | / | / |
| 9 ENTSCHEIDUNGEN | 00 | / | / | -00 | / | / | / | / | 1. | / | / | / | / | / | / |
| | | | | | | | | | | / | / | / | / | / | / |
| 10 BEFEHLEND | 17 | 17 | -01 | / | 09 | 00 | -17 | -09 | -06 | 1. | / | / | / | / | / |
| | | | | | | | | | | | / | / | / | / | / |
| 11 LERNEND | / | / | / | -07 | / | / | / | / | / | -06 | 1. | / | / | / | / |
| | | | | | | | | | | | | / | / | / | / |
| 12 ENTSPANNT | / | 15 | 16 | -07 | / | 23 | 28 | 35 | 21 | 12 | / | 1. | / | / | / |
| | | | | | | | | | | | | | / | / | / |
| 13 ERMUTIGEND | / | / | / | -07 | / | / | 26 | / | / | / | / | / | 1. | / | / |
| | | | | | | | | | | | | | | / | / |
| 14 GEFUEHLE | / | / | / | -11 | / | / | / | / | / | 02 | / | / | / | 1. | / |
| | | | | | | | | | | | | | | | / |
| 15 SELBSTSICHER | 10 | 08 | 11 | / | / | / | / | 30 | / | / | / | / | / | / | 1. |

<1> DIE ANGEGEBENEN WERTE SIND GERUNDET, DER SCHRAEGSTRICH BEDEUTET "ECHTE" NULL-KORRELATION.

<2> MIT: BEI BASIS ARITH. MITTEL, MED: BEI MEDIANBASIS, MOD: BEI MODUSBASIS (VGL. 3.3.).

TABELLE 5

DATEN ZUR RELIABILITAETSANALYSE DER SCHAETZVARIANTEN

| VAR. NR. | <2> KOEFF. | ORG. DAT. | VARIANTE A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|---|
| <1> 1-27 | R(NN) | 9889 | 9910 | 9841 | 9892 | 9746 | 9880 | 9515 |
| | R(11) | 4651 | 4942 | 4598 | 4677 | 2397 | 4551 | 2277 |

| | | ORG. DAT. | VARIANTE G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|
| 1- 6 | R(NN) | 7004 | 9089 | 7065 | 6564 | 8782 | 5069 | 4729 |
| | R(11) | 0414 | 0826 | 0371 | 0317 | 0545 | 0160 | 0108 |
| 7-15 | R(NN) | 9785 | 9770 | 9791 | 9790 | 1844 | 9737 | 9025 |
| | R(11) | 3819 | 3352 | 3746 | 3717 | 1664 | 3234 | 0876 |
| 1-15 | R(NN) | 9762 | 9648 | 9728 | 9754 | 0566 | 9635 | 9437 |
| | R(11) | 3181 | 2009 | 2794 | 2988 | 0078 | 2199 | 1319 |
| 16-27 | R(NN) | 9930 | 9930 | 9930 | 9930 | 9930 | 9930 | 9930 |
| | R(11) | 5841 | 5841 | 5841 | 5841 | 5841 | 5841 | 5841 |

<1> VARIABLEN 1 BIS 6; "GLOBALURTEILE", 7-15; "PARTIAL-URTEILE", 1 BIS 15; "URTEILE", 16-27; "HAEUFIGKEITS-BEOBACHTUNGEN" (VGL. ZUR BEDEUTUNG DER VARIABLEN TAB. 2)

<2> R(NN):=R(HORST)(VGL. FORMEL 2); R(11):=RELIABILITAET PRO RATER (VGL. FORMEL 1)

TABELLE 6

FORMALE BESTIMMUNG DER ZAHL DER ZU EXTRAHIERENDEN FAKTOREN

| SPALTE | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| | BASIS HAUPTKOMPONENTENLOESUNG | | | | | FAKTOREN-LOESUNGEN <1> | |
| DAT. BASIS | SCREE | L(I)>1 <3> | L(I)>5% D.GES.V. | SL(I)>90% D.GES.VAR <4> | SL(I)>SH2 <5> (SH2 IN % D.GES.VAR) | L(I)>1 | L(I)>5% D.GES.V. |
| ORG.-DATEN | 3-4 | 8 | 6 | 17 | 4 (53,5%) | 4 | 3 |
| A | 4/7 | 8 | 6 | 18 | 4 (50,7%) | 4 | 3 |
| B | 4 | 6 | 3 | 14 | 4 (67,6%) | 3 | 2 |
| C | 5 | 8 | 6 | 17 | 4 (52,8%) | 4 | 3 |
| D | (1) <2> | 12 | 7 | 21 | 3 (25,6%) | 2/4 | 1 |
| E | 3/8 | 7 | 5 | 16 | 4 (55,9%) | 4 | 3 |
| F | 5 | 5 | 4 | 12 | 6 (74,6%) | 4 | 3 |
| G | 3-4 | 10 | 7 | 19 | 4 (41,5%) | 3 | 2 |
| H | 3-4 | 8 | 5 | 17 | 4 (51,4%) | 4 | 3 |
| I | 3 | 9 | 5 | 17 | 4 (51,2%) | 4 | 3 |
| J | 3 | 11 | 8 | 20 | 4 (37,3%) | 2-3 | 2 |
| K | 3/8 | 8 | 6 | 18 | 4 (46,3%) | 4 | 3 |
| L | 3 | 10 | 6 | 20 | 3 (36,9%) | 4 | 2 |

<1> DIESE ANGABEN BERUHEN AUF DER AUSWERTUNG DER 6-,-1-FAKTOREN-LOESUNGEN.
<2> DAS EIGENWERTDIAGRAMM DIESER DATENGRUPPE ZEIGT KEINEN "KNICK" (VGL. DAZU KAP. 3.4.3.2. UND ABB. 3-18).
<3> L:=LAMBDA (EIGENWERT); I:=ZAHL DER EIGENWERTE
<4> S:=SUMME
<5> H2:=KOMMUNALITAET

TABELLE 7

VERGLEICH DER DATENGRUPPEN UNTER FAKTORENANALYTISCHEN GESICHTSPUNKTEN

OBERE DREIECKSMATRIX; AEHNLICHKEITSKOEFFIZIENTEN DER FAKTORENMATRIZEN AUS DER HAUPTKOMPONENTENLOESUNG<2>
UNTERE DREIECKSMATR.; "VERSCHIEDENHEITSBESTIMMUNG" MIT "FAUSTREGELN" ZUR BESTIMMUNG DER FAKTORENZAHL(VGL.3.4.2.3.)<1>

| DAT. GRUPPE | A | B | C | D | E | F | G | H | I | J | K | L | ORG.-DATEN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | - | | | | | | | | | | | | |
| B | | - | | | | | | | | | | | |
| C | X | | - | | 92(3) | | | 99(3) | 99(2) | | | | 99(6) |
| D | | | | - | | | | | | | | | |
| E | | | | | - | | | 93(3) | 93(3) | | | | 98(6) |
| F | | | | | | - | | | | | | | |
| G | | | | | | | - | | | | | | |
| H | X | | X | | X | | | - | 99(2) | | | | 99(6) |
| I | X | | | | | | | XX | - | | | | 98(6) |
| J | | | | | | | XX | | | - | | | |
| K | X | | | | | | | X | X | | - | | |
| L | | | | | | | XX | | | | | - | |
| ORG.-DATEN | X | | X | | X | | | XX | XX | | X | | - |

<1> X;"AEHNLICHKEIT" IM "GUENSTIGSTEN" FALLE; XX: "AEHNLICHKEIT" IM "UNGUENSTIGSTEN" FALLE

<2> DIE ANGABEN BEZIEHEN SICH AUF DEN KLEINSTEN PHI-KOEFFIZIENTEN AUS DEN 6...1-FAKTOREN-LOESUNGEN; IN KLAMMERN IST ANGEGEBEN, UM WELCHE LOESUNG ES SICH HANDELT.

TABELLE 8A

FAKTORENLADUNGEN FUER VERGLEICHBARE VARIABLEN AUS EMPIRISCHEN UNTERSUCHUNGEN UND DER SPRACHSTRUKTUR - ANALYSE

| VAR. NR. | MNEMOTECHN. KURZFORM <4> | MUELLER-WOLF (1977,98 - 101) VAR. | LADUNGEN<2> I | II | III | SPRACHSTRUKTUR-ANALYSE(BASIS: MITTELWERTE LADUNGEN<2> I | II | III |
|---|---|---|---|---|---|---|---|---|
| 7 | PARTNER | LV31 | ,78 | | | ,95 | | |
| 14 | SENSITIV | LV18 | ,75 | | | ,78 | ,54 | |
| 1 | EMOTIONAL | LV22 | ,75 | | | ,64 | ,55 | |
| 8 | VERSTAENDNIS | LV32 | ,74 | | | ,91 | | |
| 13 | ERMUTIGT | LV 3 | ,70 | | | ,79 | ,50 | |
| 6 | SOZ.-INT. | LV16 | ,63 | | | ,89 | | |
| 12 | LERNENDER | LV29 | ,57 | | | ,77 | | |
| 3 | LIBERAL | LV24 | ,54 | | ,40 | ,87 | | |
| 4 | AUTOKRAT. | LV10 | ,51 | | ,49 | ,20<3> | | |
| 15 | SELBSTSICHER | LV 2 | | ,72 | | | ,64 | ,68 |
| -<1> | ANREGEND | LV 6 | | ,72 | | ,64 | ,63 | |
| 5 | QUALIFIKAT. | LV30 | ,48 | ,62 | | ,69 | ,62 | |
| 11 | ENTSPANNT | LV 7 | | ,52 | | | ,84 | |
| 10 | BEFEHLEND | LV 8 | | | ,66 | | | ,99 |
| 9 | ENTSCHEIDGN. | LV15 | | | ,64 | ,91 | | ,18<3> |

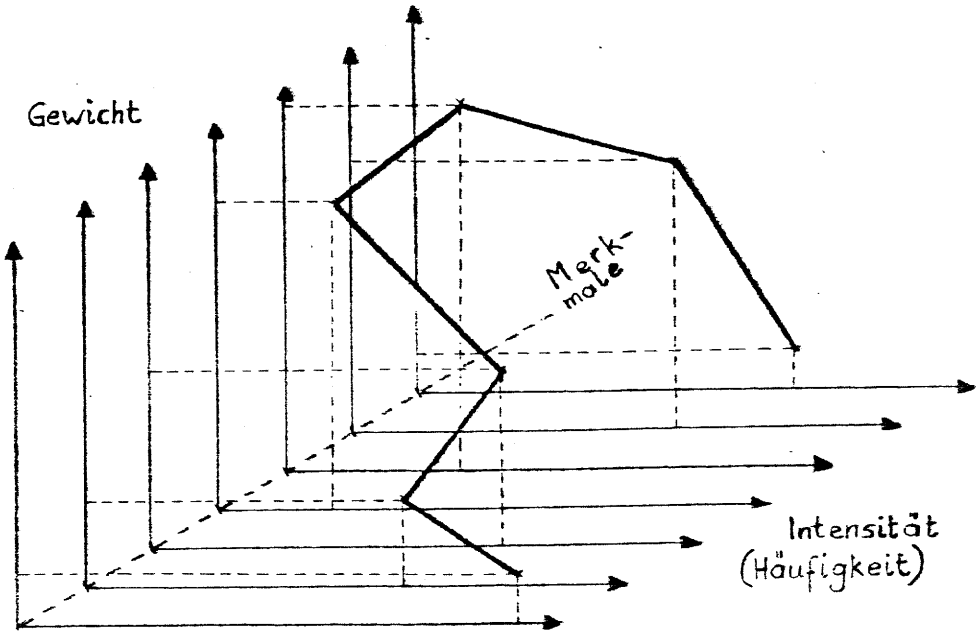
<1> FUR DIESE VARIABLE WURDEN NUR DIE SPRACHAEHNLICHKEITSWERTE ERMITTELT.
<2> ANGEGEBEN SIND LADUNGEN >=,40.
<3> DIESER WERT LIEGT UNTERHALB DER FUER UEBEREINSTIMMUNG ZU ERWARTENDEN HOEHE.
<4> VGL. DAZU TAB. 2.

TABELLE 8B

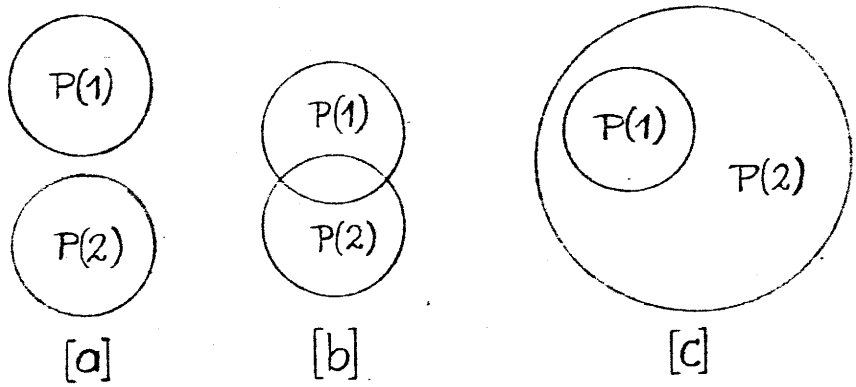
| VAR. NR. | MNEMOTECHN. KURZFORM <4> | MUELLER-WOLF/ FITTKAU('71,172) VAR. | LADUNGEN<2> | | SPRACHSTRUKTUR-ANALYSE(BASIS: MEDIANE LADUNGEN<2> | |
|---|---|---|---|---|---|---|
| 10 | BEFEHELEND | 8. | .85 | | .04<3> | .72 |
| 7 | PARTNER | 31. | .85 | | .98 | |
| 8 | VERSTAENDNIS | 32. | .84 | | .98 | |
| 9 | ENTSCHEIDGN. | 15. | .83 | | .94 | |
| 12 | LERNENDER | 29. | .82 | | .84 | |
| 14 | SENSITIV | 18. | .79 | | .94 | |
| 4 | AUTOKRAT. | 10. | .79 | | .12<3> | .99 |
| 13 | ERMUTIGT | 3. | .76 | | .94 | |
| -<1> | ANREGEND | 6. | | .75 | .84 | .32<3> |
| 15 | SELBSTSICHER | 2. | | .72 | | .85 |

<1> SIEHE TAB. 8A
<2> ANGEGEBEN SIND LADUNGEN >=.70.
<3> SIEHE TAB. 8A
<4> VGL. DAZU TAB. 2.

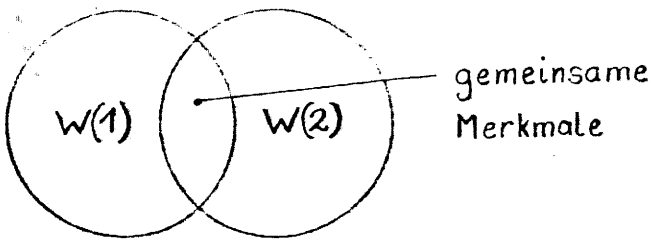
# A B B I L D U N G E N



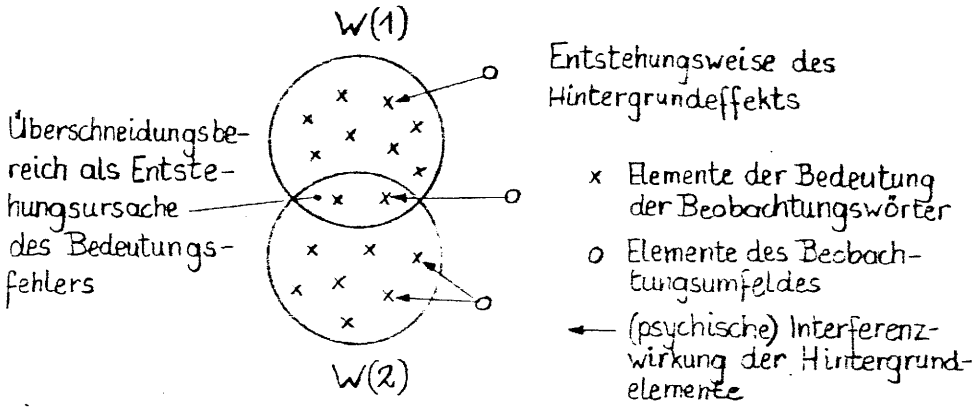

2-2 Bedeutungsprofil eines Wortes



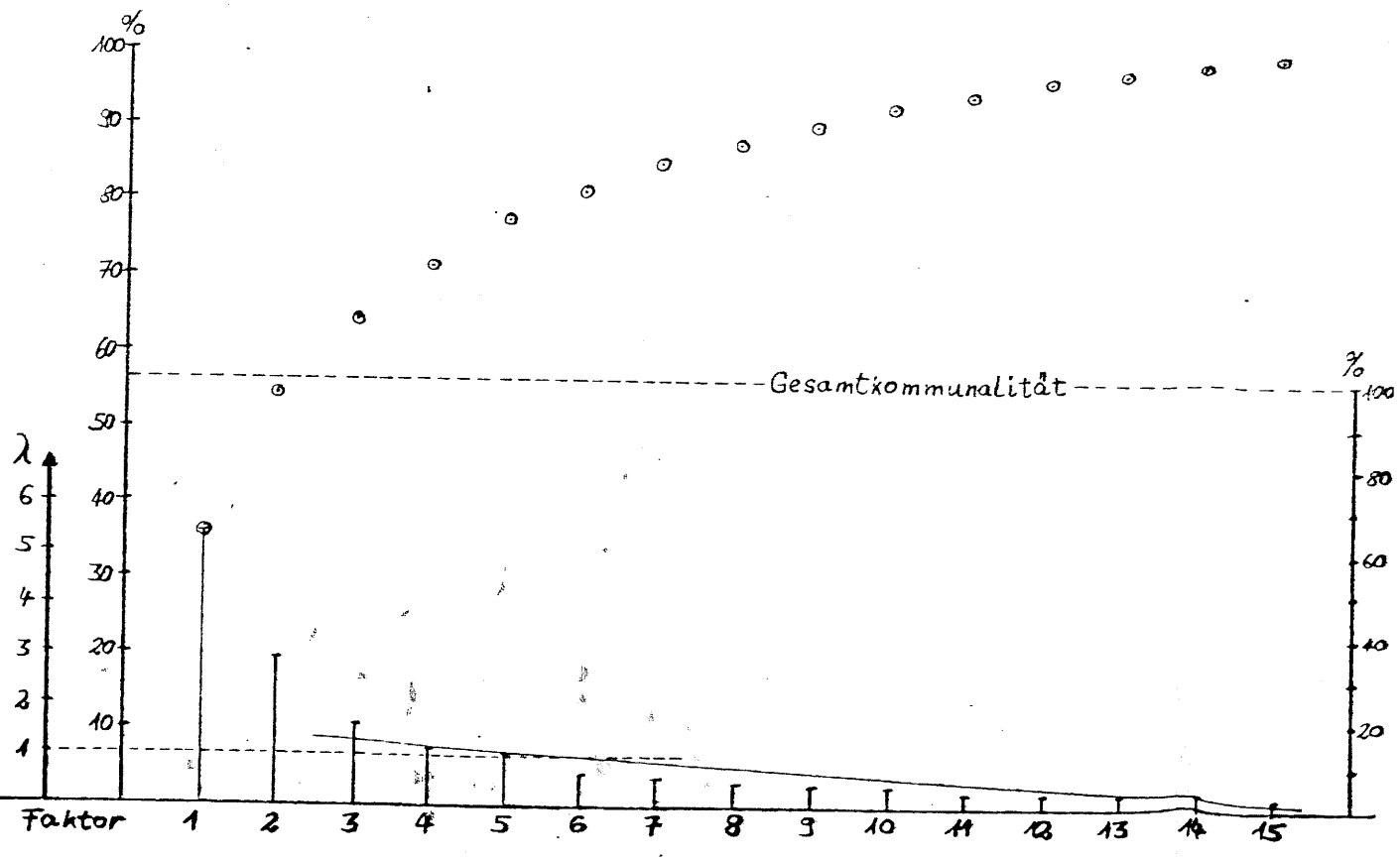

2-3 Verhältnis der Merkmalsmengen bei verschiedener Verwendung desselben Wortes durch zwei Personen



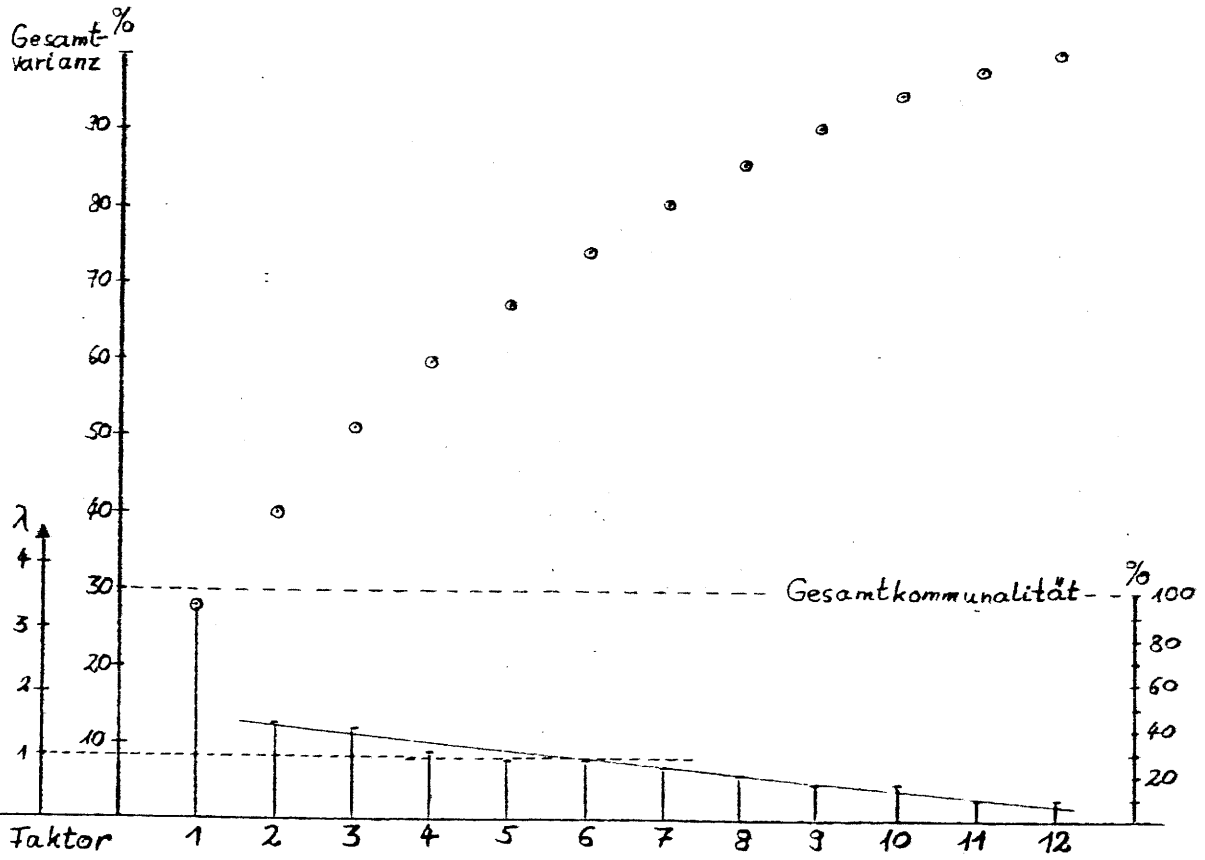

2-4 Die Bedeutungsgemeinsamkeit zweier Wörter



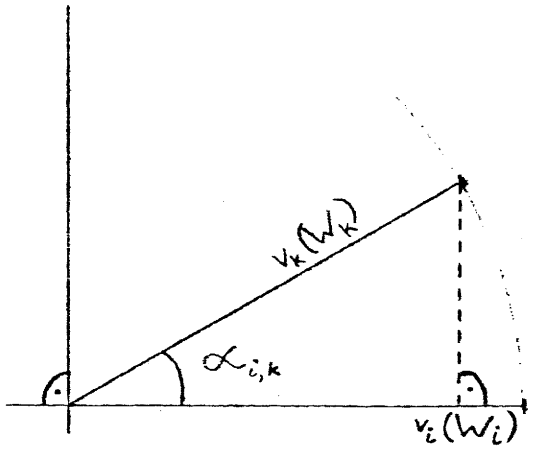

2-5 Bedeungsfehler und Hintergrundeffekt



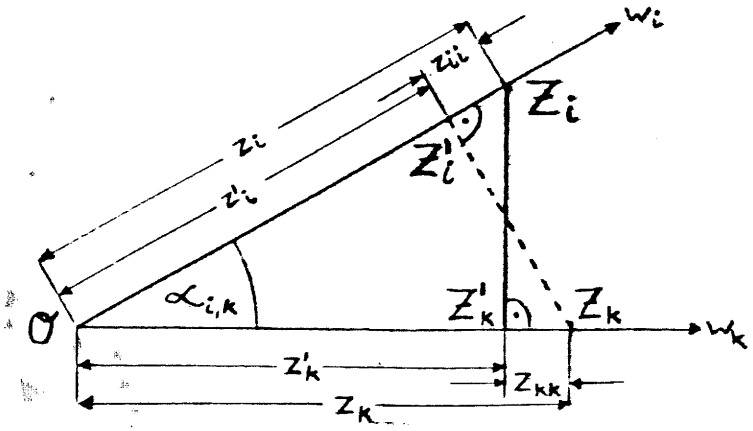

3-4: Bestimmung der Faktorenzahl für Urteilsdaten (Stabilitätsprüfung)



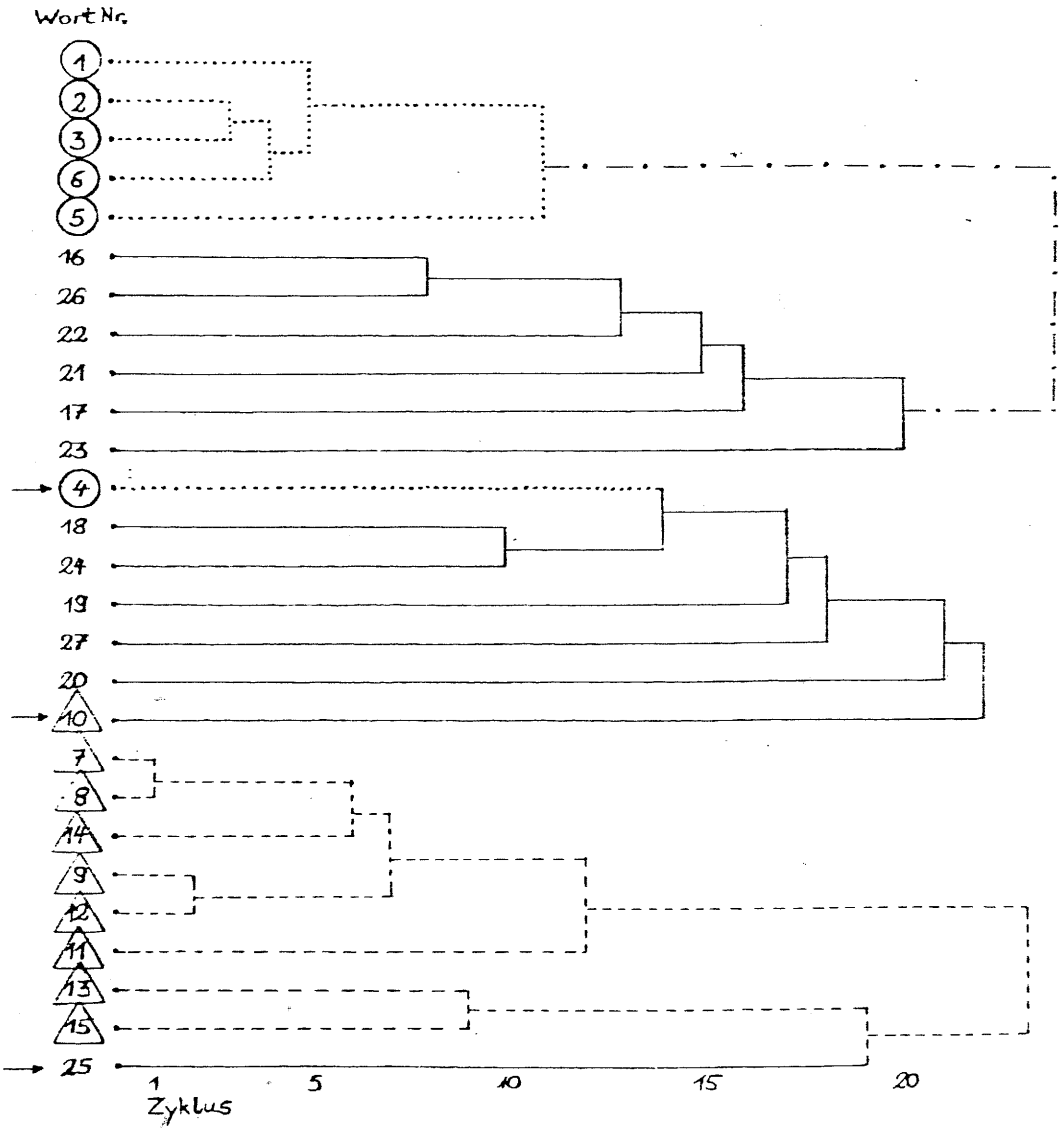

3-5: Bestimmung der Faktorenzahl für Häufigkeitsdaten (Stabilitätsprüfung)



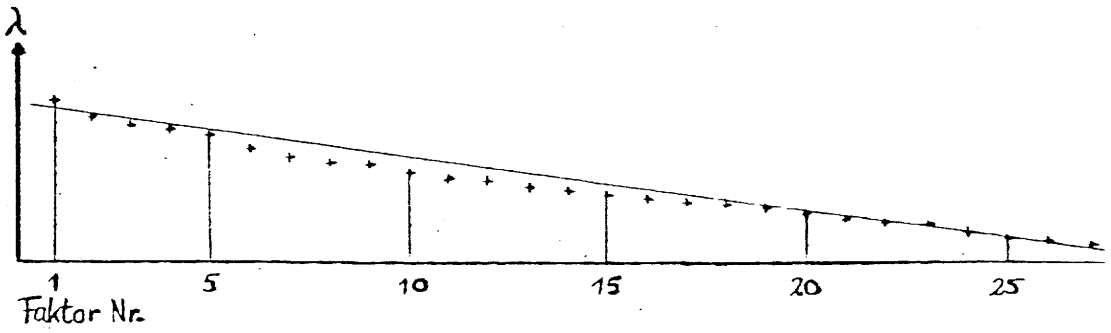

3-7 Geometrische Darstellung der Ähnlichkeitsmaße



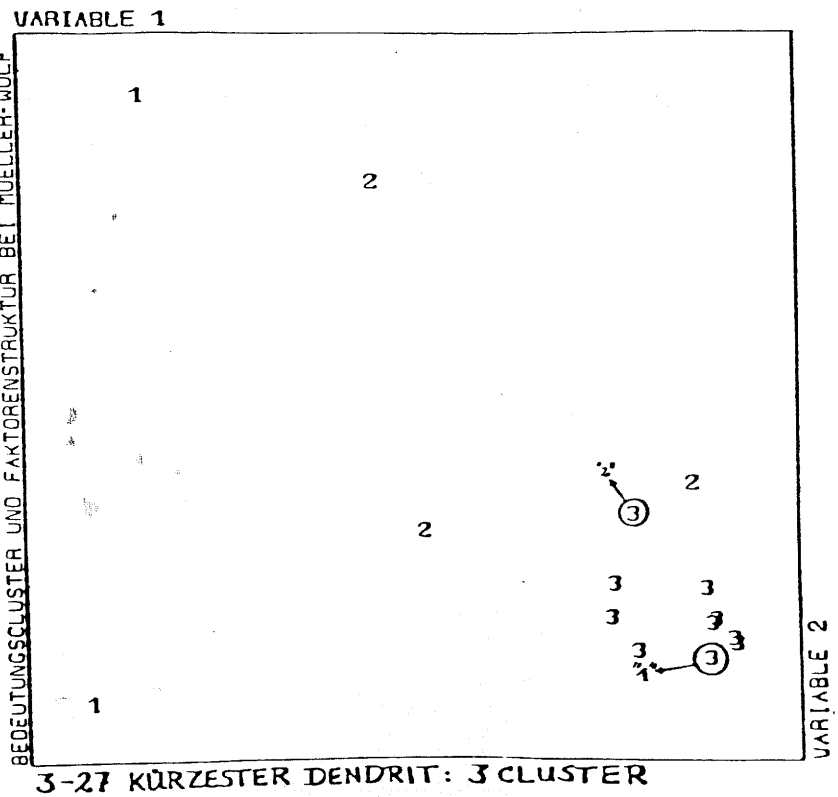

3-9 Zweidimensionales Modell der Ähnlichkeitsrelation




3-13 Dendrogramm der Beobachtungswörter




3-20 Eigenwerte der Korrelationsmatrix für die Schätzvariante D




3-27 KÜRZESTER DENDRIT: 3 CLUSTER

# A N M E R K U N G E N

<1>

Die mit dem Vorgang der Wahrnehmung und ihrer Verarbeitung verbundenen erkenntnistheoretischen Probleme werden hier nicht naeher behandelt (vgl. dazu z.B. Bohnen 1972). Das gleiche gilt im wesentlichen fuer die Fragestellung der modernen Sprachphilosophie, die in der Nachfolge von Wittgenstein behandelt wird (vgl. z.B. Searle 1969; Austin 1976).

<2>

Selbstverstaendlich koennte auch der umgekehrte Effekt eintreten, der darin bestehen wuerde, dass Begriffe trotz vermuteter Bedeutungsgleichheit zu gering korrelierten Daten fuehren. Dass dieser - augenfaelligere - Fall von sprachlich beeinflussten Ergebnissen in Untersuchungen zur Unterrichtsforschung nicht diskutiert wird, liegt sicherlich daran, dass tautologische Skalen vom Instrumentenkonstrukteur von vornherein vermieden werden, d.h. dass dieser von der Verschiedenheit der aufgenommenen Merkmale ueberzeugt ist und die "Tautologiehypothese" daher gar nicht prueft.

<3>

Wir vermeiden hier und im weiteren den Ausdruck "Begriff", um Assoziationen mit einzelnen Varianten der Deutung des Zusammenhangs zwischen "Wort" (als sinnlich wahrnehmbare Chiffre) und "Bedeutung" (als interne Sinnkonstitution) auszuschliessen. Vgl. zur philosophischen Problematik den Ueberblick bei Wagner (1973); kritisch zur Vieldeutigkeit des "Begriffs": Gaetschenberger (1977, 20 ff.). Weitere Angaben finden sich im Text dieses Abschnitts.

<4>

Funktion und Bedeutung der zweiten Gruppe von Termen, der Namen, werden im vorliegenden Zusammenhang nicht diskutiert (vgl. dazu z.B. Lyons 1977, 147 f.).

<5>

Zu den Abstaenden der Skalenpunkte und zur Polung der Skalen in der zitierten Arbeit vgl. Mueller-Wolf (1977, 197, 203); zu methodischen Aspekten der Skalenkonstruktion vgl. z.B. Hasemann (1971, 820 ff.). Auf drei be-

sonders problematische Punkte an den ausgewaehlten Beispielen soll hier nur hingewiesen werden: (1) Durch die Vorgabe von mehreren Woertern an jedem Skalenende koennen ueber die in dieser Untersuchung diskutierten Fehler hinaus Ungenauigkeiten hervorgerufen werden; (2) die Vergabe eines Null-Punktes provoziert u.U. Interpretationsverschiedenheiten bei den Beobachtern (vgl. das Beispiel bei Fenner 1973, 128-129) und laesst darueberhinaus die Frage entstehen, ob hier nicht ein systematischer Verstoss gegen die Konstruktvaliditaet vorliegt; die Suggestion einer kontinuierlichen Veraenderlichen durch Skalendarstellung steht im Widerspruch zur Vorstellung von einer "Nicht-Auspraegung" bzw. eines "Nicht-Vorkommens" des Merkmals (wie ja auch beim Thermometer der Nullpunkt nicht das Fehlen des Waermemerkmals bedeutet). Die Schwierigkeit besteht - inhaltlich gesprochen - darin, dass diese Stelle der Skala entweder als Nicht-Vorhandensein oder als mittlere Auspraegung des Merkmals gedeutet werden kann. Im ersten Falle liegt ein Verstoss gegen die Konstruktvaliditaet vor, im zweiten ist die "Erlaeuterung" des mittleren Skalenpunktes unzutreffend (vgl. auch Hasemann 1971, 830-831). (3) Vor allem bei Polaritaetsskalen (Ordinal-, Intervallniveau) bleibt voellig offen, mit welcher (Durchschnitts-/Ideal-)Vorstellung der Beobachter seine Wahrnehmungen vergleicht. Es fehlt hier gewissermassen das "secundum comparationis" - ein Mangel, von dem bei Personenbeurteilungen auch Verhaeltnisskalen moeglicherweise beeintraechtigt werden.

<6>

Es waere beispielsweise denkbar, dass Geschlecht, Kleidung, Bewegung im Raum ("spatial behavior") o.ae. nicht in das subjektive Bedeutungsfeld "freundlich - unfreundlich" eingeschlossen sind, also bei der Generierung dieses Urteils nicht "in Betracht" gezogen werden (vgl. Anm. 7). Auf die Dimensionalitaet des Wahrnehmungsraums kommen wir weiter unten (Abschn. 2.1.2.) zu sprechen. Vgl. zum Verhaeltnis von Wort und zugehoerigen Merkmalen Cicourel (1970, 30, 36-37).

<7>

Damit wird die Vorstellung verbunden, dass einzelne wahrgenommene Merkmale von unterschiedlicher Relevanz fuer das Zustandekommen des Urteils sind. So koennte etwa von einem bestimmten Beobachter sowohl Mimik ("facial expressions") als auch Verbalverhalten ("linguistic behavior") zur Urteilsbildung herangezogen, die Mimik aber

fuer "wichtiger" gehalten werden.

<8>

Es haengt vom Merkmal und vom Differenzierungsvermoegen des Beobachters ab, in welcher Weise er den Ausprae-gungsgrad "wahrnimmt", Kategorisierung (z.B. Dichotomi-sierung) fuehrt zu Haeufigkeitsinformationen (z.B. Stirnfalten; gezeigt vs. nicht gezeigt), Metrisierung fuehrt zu Intensitaetsinformationen (z.B. Sprechlaut-staerke; gering bis stark). Beide Strategien koennen (meist) in Bezug auf ein und dasselbe Merkmal angewendet werden (z.B. Stirnfalten; "gar nicht" ueber "sanft" ("smooth"; vgl. Weick 1968, 384) bis "tief, gefurcht"; Sprechlautstaerke; "leise" vs. "laut"). - Auf einen moeglichen Spezialfall ist hier noch hinzuweisen; es ist denkbar, dass bestimmte Merkmale, die im Bedeutungsfeld eines "Urteilswortes" enthalten sind, mit - individuel-len - Schwellenwerten versehen sind in der Weise, dass sie unter- bzw. oberhalb kritischer Intensitaets-/Haeu-figkeitswerte nicht (Gewichtungsfaktor 0,0; vgl. Anm. 6), aber jenseits dieser Grenzen beachtet werden.

<9>

Wir untersuchen hier nicht weiter, ob die in einer Beo-bachtungssprache vorkommenden Woerter von unterschiedli-cher Komplexitaet in dem Sinne sind, dass ihr Bedeu-tungsraum durch wenige oder mehr Dimensionen aufgespannt wird. Unter einem physiologischen Aspekt muesste in die-sem Zusammenhang dargestellt werden, welche Dimensiona-litaet der Reize mit einzelnen peripheren sensorischen Endorgane kodiert werden kann und wie Woerter und Bedeu-tungen ueber Sinneserregungen konstituiert werden (vgl. zur Farbwahrnehmung z.B. Lenneberg 1977, 411 ff.).

<10>

In der Sprache der Mengenlehre waere - praeziser - zu sagen, dass das Wort als Name gilt fuer eine Relation zwischen einer relevanten Merkmalsmenge und der Menge der geordneten Paare aus Gewicht und Intensitaet bzw. Haeufigkeit.

<11>

Wir halten nicht eine streng mathematische Ausdruckswei-se durch, da dies fuer die Darstellung unseres Problems

nicht erforderlich ist.

<12>

Im Falle einer Haeufigkeitsinterpretation muesste statt der Intervall- eine Ordinalskala gezeichnet werden. Die Merkmalskoordinate ist hier als "Nominalskala" abgebildet; diese Ungenauigkeit laesst sich durch graphische "Aufloesung" in eine entsprechend der Maechtigkeit der Merkmalsmenge grosse Zahl von zweidimensionalen Koordinatensystemen beseitigen.

<13>

Die hier diskutierten Gewichte duerfen nicht verwechselt werden mit den bei Friedrichs/Luedtke (1973, 38) diskutierten Objektivitaets-Gewichten fuer die Qualitaet der unterschiedlichen Beobachter.

<14>

Die Problematik des Aequivalenzkoeffizienten, wie sie hier dargestellt wird, ist im uebrigen auch im Verfahren der Bestimmung der internen Validitaet ("Konzepttreue") enthalten, da die "wahren Werte" auf derselben Basis hervorgebracht werden wie die zu pruefenden Messwerte (vgl. z.B. Langer/Schulz v.Thun 1974, 141 ff.). - Zur Charakterisierung des Reliabilitaetskonzepts im Kontext von Beobachtungen vgl. Fassnacht (1979, 60 ff.).

<15>

Diese Auffassung wird z.B. bei Medley/Mitzel (1971, Sp. 661), Sumaski (1977, 71), Dechmann (1978, 238) vertreten.

<16>

Vgl. zur Unterscheidung von Gegenstand und Aspekt Gruemer (1974, 75-76), Brunner (1978, 150).

<17>

Wir lassen hier die Frage offen, ob die Modifikation in der "perzeptiven" oder der "apperzeptiven" Phase erfolgt.

<18>

Diejenigen Einfluesse, die von unterschiedlichen (situativen oder ueberdauernden) Auspraegungen des sensorischen Apparates und der Reizleitungsbahnen herruehren, werden hier nicht diskutiert. Vgl. z.B. zum Einfluss der Reaktionszeit Wiens u.a. (1966). - Auch wissenschaftssoziologisch erfassbare Effekte, die z.B. im Zusammenhang mit der Unterscheidung von Untersuchungsleitung und -durchfuehrung identifiziert werden koennen (Forscherehrgeiz, Beeinflussung des Feldes durch die Messung, Datenfaelschung etc.), bleibt ausser Betracht (vgl. dazu Barber 1973). - Literaturangaben zu empirischen Untersuchungen ueber das Ausmass der einzelnen Fehler aus Tab. 1, Ziff. 2-20 finden sich bei Guilford (1954, 294-295).

<19>

Selbstverstaendlich wirken sich Bedeutungsueberschneidungen auch bei Anwendung parameterfreier Analyseverfahren aus, weil sie ja den zu untersuchenden Daten innewohnen. Wir diskutieren diese Frage hier aus zwei Gruenden nicht weiter; zum einen wegen der geringeren Verbreitung solcher Analysestrategien (vgl. die bereits erwaehnten Zusammenstellungen bei v.Cranach/Frenz 1969, 283-285, 305-307), zum anderen wegen des dafuer erforderlichen raeumlichen Aufwandes (die Vielfalt der vorliegenden Konzepte (vgl. Lienert 1973) wuerde sich im Umfang der Eroerterung niederschlagen). Dennoch waere gerade in diesem Bereich eine sorgfaeltige Untersuchung nuetzlich, weil dort Analysestrategien zur Verfuegung gestellt werden, deren Voraussetzungen eher mit den Gegebenheiten des Ratings zusammenpassen, als dies bei parametrischen Verfahren normalerweise der Fall ist (z.B. hinsichtlich des Skalenniveaus und der zu unterstellenden Verteilungen).

<20>

Wir brauchen hier die Frage der sog. Hintergrundsfaktoren (Faktoren zweiter Ordnung) nicht zu diskutieren. Sie spielen nur im Falle von Erklaerungen eine Rolle und repraesentieren dort komplexe theoretische Konstrukte. Unter der Idee der Entwicklung eines Beobachtungsinstruments waere daher der Einsatz solcher Verfahren ein "Schritt in die falsche Richtung", weil damit stets hoehere Inferenzleistungen der Beobachter verbunden waeren, die ihrerseits einen unguenstigen Einfluss auf die Da-

tenqualitaet haben (vgl. stellvertretend fuer viele Quellen v.Cranach/Frenz 1969, 283-285 und die dort angegebene Literatur), also genau auf das, was durch die Massnahme verbessert werden sollte.

<21>

Zwar bezieht sich Fassnacht in seinen Ausfuehrungen explizit auf die Gegenueberstellung von Beobachtung und Experiment. Diese Begriffe stehen bei ihm aber als genus proxima fuer Schaetzverfahren und Tests (vgl. 1979, 148 und 45 ff.).

<22>

Im "deskriptiven Fall" spielt dann der Bedeutungsfehler die von uns oben gekennzeichnete Rolle. Fuer den "Wertungsfall" gilt das gleiche im Hinblick auf die Identifizierung des Wertobjekts. Dazuhin tritt eine zweite Bedeutungsfehler-Variante, die sich auf die Relation zwischen empfundener Einstellung und Wertpraedikat bezieht (wir untersuchen sie hier nicht naeher), waehrend die uebrigen "emotionalen" Komponenten (vgl. Tab. 1, Ziff. 13-17) in diesem Zusammenhang nicht als Fehler, sondern als Messobjekte betrachtet werden muessen.

<23>

Neben dem Itemtext wird auch die Instruktion bei Diehl und Kohr so gefasst, dass diese Unklarheit bestehen bleibt; ">>Machen Sie bitte bei jeder Aussage durch Ankreuzen in einem der vier Kaestchen kenntlich, wie weit nach ihrer Meinung diese Aussage fuer die Veranstaltung zutrifft oder nicht zutrifft.<< Auf jedes Item konnte durch Ankreuzen mit >>stimmt<<, >>stimmt ueberwiegend<<, >>stimmt ueberwiegend nicht<< oder >>stimmt nicht<< reagiert werden" (1977, 64). Diese Formulierung waere geeignet, die (intersubjektive) Wahrheitsfaehigkeit der Antworten zu suggerieren; dagegen wird bei Itemtexten wie "Der dargebotene Stoff kam einem oft ziemlich unwichtig vor" (1977, 66) deutlich, dass "subjektive Daten" ermittelt werden koennen. - Zum Problem der Wertungsbegriffe in der Umgangssprache vgl. die Bemerkung bei Friedrichs/Luedtke (1973, 41).

<24>

Das gilt natuerlich auch fuer die Variable "Interak-

tionsgeschehen" (Mueller-Wolf 1977, 33), die als Wirkung von Verhaltens-, aber als Ursache von Einstellungsausspraegungen betrachtet werden kann. - Es bleibt im uebrigen unklar, welcher kausale Rang bei Mueller-Wolf (1977, 32) der ">>Bedeutung<< des Lehrverhaltens" im Unterschied zur ">>Auswirkung<<" zugeschrieben wird. - Charakteristisch fuer diese "kausale Unklarheit" sind auch einige Aussagen von Mueller-Wolf (1977), die sich auf die Frage beziehen, ob das von Studenten beurteilte Lehrverhalten als unabhaengige Messung zu den (von denselben Studenten geaeusserten) Angaben ueber dessen Wirkungen in Beziehung gesetzt werden duerfe. Offenbar von Kritikern darauf aufmerksam gemacht, versucht der Autor die Gefahr der "Konfundierung" seiner Daten dadurch zu umgehen, dass er neutrale Beobachter als Datenlieferanten einbezieht, Waehrend im Falle des Selbst- und Fremdratings (vgl. Langer/Schulz v.Thun 1974, 102) der betroffenen Studenten von Konfundierung gesprochen wird (49), soll dieser Fehler beim (neutralen) Fremdrating nicht auftreten (35). Tatsaechlich bestehen auch nur mittlere Uebereinstimmungen zwischen den beiden Ratergruppen (107). Dennoch fuehrt ein Vergleich der Zusammenhaenge zwischen Lehr- und Studierverhalten, getrennt nach Ratergruppen, zu "weitgehende(r), grundsaetzliche(r) Uebereinstimmung" (121-122; im Original kursiv). U.E. ist diese Inkonsistenz auf das Analyseverfahren und die - unzulaessige - Interpretation des Korrelationskoeffizienten als Intervalldatum (vgl. 122) zurueckzufuehren. Anderenfalls waere die Einschaetzung der unterschiedlichen Datenqualitaet nicht durchzuhalten (vgl. Anm. 25).

<25>

Im letzteren Falle muesste man eher von einem psychotherapeutischen Eingriff als von der "Herstellung" eines Untersuchungsinstruments sprechen.

<26>

Die Ausfuehrungen von Mueller-Wolf (1977) zu diesem Punkt sind undurchsichtig, z.T. widerspruechlich. Er hebt zunaechst hervor, dass die hohe Uebereinstimmung der (ungeschulten) Rater (ca.: ,90) ein Hinweis auf die Unabhaengigkeit des Verfahrens "von den Spezifitaeten des Messvorgangs" seien (103). Um diesen Aspekt der "Objektivitaet" aber zu sichern, werden auch (geschulte; 103) "neutrale Rater" (34) eingesetzt, die zu ueber ,90 (103) zuverlaessig sind. - Wenn zunaechst schon auffallen muesste, dass die Schulung keine wesentlichen "Ver-

besserungen" bringt (Waren die neutralen Rater nicht "objektiv"? Waren die betroffenen Studenten "neutral"?), so haette erst recht die Untersuchung der Uebereinstimmung zwischen diesen beiden Gruppen stutzig machen muessen; sie belaeuft sich - in den Faktoren der Vorlesungsuntersuchung - auf maximal ,57 (107, 109)! Das bedeutet, dass recht verschieden geurteilt wurde, was Mueller-Wolf auch zu der Bemerkung veranlasst, die Ratings enthielten "nicht nur... subjektive Komponente(n)...., sondern auch eine nicht objektive, realitaetsbezogene Komponente" (108-109; Unterstreichung im Original kursiv; das Woertchen "nicht" repraesentiert sicherlich die Fehlleistung eines unverstaendigen (?) Setzers). - Im "Ueberblick" (12) liest sich das anders (naemlich wie oben; 103): "Es zeigt sich, dass Studenten in der Lage sind, ein brauchbares Feedback... zu geben; Die Beurteilungen der Studenten weisen recht hohe bis hohe Uebereinstimmungen mit den Einschaetzungen der neutralen Beobachtern auf." Dagegen (33): In der Untersuchung geht es "durchaus auch um >>subjektive Befindlichkeit<< und nicht nur um sog. objektive Befunde." Dennoch soll "die Hinzuziehung der neutralen Beurteiler" eine "ansatzweise" Pruefung des ">>objektiven Gehalt(s)<< (der) Beurteilungen der Studenten" (35) ermoeglichen. Konsequenzen aus dieser Pruefung werden aber nicht gezogen. Die kaum zulaessige Behauptung "hoher" Uebereinstimmung kollidiert mit der - ebenfalls zentralen - Versicherung: "Nicht >>objektives<< Lehrverhalten werde erfasst, sondern ein subjektiv wahrgenommenes Lehrverhalten..." (49).

<27>

Zum Zusammenhang von theoretischen und methodischen Problemen vgl. Blalock Jr. (1978).

<28>

Das hier Gesagte betrifft im Prinzip nur den Begruendungszusammenhang von Wissenschaft. Das bedeutet allerdings nicht, dass diese Ueberlegungen ohne Ausstrahlung auf den Entdeckungszusammenhang waeren (vgl. Hyman 1964, 46). Zwar liegen fuer ihn keine Regeln im strengen Sinne vor, aber der methodologische und metatheoretische Anspruch des wissenschaftlichen Procedere regieren auch diesen Bereich (insofern unterlaege man einem Irrtum, wenn man glaubte, dass im Entdeckungszusammenhang einer analytisch-empirisch begriffenen Wissenschaftsauffassung "alles" moeglich sei). So implizieren der Wunsch nach Erhoehung von Unterrichtseffizienz und das Interesse an ihren Bedingungen bereits die Idee der Unterscheidung

von Ursache und Wirkung und damit die Unterscheidung entsprechender Merkmalsgruppen. Damit ist aber bereits klar, dass Sprachprobleme von der Art des Bedeutungsfehlers auftreten koennen.

<29>

Wir diskutieren - wie in Abschn 1. angekuendigt - hier nicht ausfuehrlich die Problemlage in verschiedenen Paradigmata. Der an dieser Stelle gegebene Hinweis erscheint uns lediglich deshalb von Bedeutung, weil in einem Teil der von uns benutzten Literatur (z.B. Friedrichs/Luedtke 1973, Mueller-Wolf 1977, Dechmann 1978) Bezuege zum action-research -Konzept und damit zur kritischen Theorie der Frankfurter Schule hergestellt werden.

<30>

Es wird sich weiter unten herausstellen, dass es sich hier aus der Sicht einer analytisch-empirischen Position gar nicht um eine Alternative handelt.

<31>

Hier liegt eine andere Art von Konfundierung vor als sie von Hofer wegen der mangelnden Unabhaengigkeit der studentischen Rater zurecht befuerchtet wird (vgl. Mueller-Wolf 1977, 35 und Anm. 24) und von der Muel-ler-Wolf - zu unrecht - glaubt, sie durch "neutrale" studentische Rater ausgeschaltet zu haben (ebd.). "Neutrale" und "nicht-neutrale" Studenten unterliegen gemeinsam einer Art von Effekt, den McCall als Ethnozentrismus (vgl. Gruemer 1974, 64-65) bezeichnet.

<32>

Es kann deshalb auch nicht verwundern, dass Mueller-Wolf bei der Auswertung seiner Untersuchung fuer die Praxis (1977, 125-130, 186-195) nicht zu Resultaten gelangt, die - in diesem Falle - bestimmte Verhaltensweisen von Hochschullehrern im Hinblick auf bestimmte Effekte auszeichnen. Dies war - wie einleitend bemerkt wird (32) - ohnehin nicht beabsichtigt; dennoch sind im Design "echte" Ursache-Wirkung-Analysen enthalten (Lehrverhalten/studentisches Verhalten). Und tatsaechlich werden auch Ratschlaege gegeben fuer effizienteres Lehrverhalten (194, 195), die aber u.E. in ihrem Informationsge-

halt nicht ueber das hinausgehen, was man in der paedagogischen Praxis an Hinweisen und Ratschlaegen bereits seit laengerem tradiert (vgl. z.B. Salzmann 1915, zuerst 1806; o.J. zuerst 1780).

<33>

Mueller-Wolf (1977, 104-105) unternimmt im uebrigen selbst eine von ihm so genannte Semantik-Studie, in der sich zwischen 'empirischer' und 'semantischer' Faktorenstruktur ein hoher Aehnlichkeitskoeffizient von .85 ergibt. Bei der Interpretation kommt er aber - wie es scheint - gar nicht auf den Gedanken, dass die 'empirischen' Faktoren nur scheinbar 'empirisch' seien; vielmehr ist er der - erstaunlichen - Auffassung, dass der semantische Befund den empirischen bestaetige (105)!

<34>

Nicklis (1978), Walter (1979), Nicklis (1979). - Das von den in unseren bisherigen Ausfuehrungen zitierten Autoren immer wieder betonte Erfordernis der Praxisrelevanz mag auch vor dem Hintergrund der Besorgnis um die Erhaltung der "Aussenlegitimitaet" von (Erziehungs-)Wissenschaft verstanden werden. Wuerde sie auf diesem Wege hergestellt - was wir, wenigstens fuer eine kurzfristige Perspektive nicht ausschliessen koennen - so waere damit die Moeglichkeit eines antagonistischen Verhaeltnisses zur Binnenlegitimitaet aufgezeigt (vgl. dazu auch Zabeck 1978). Allerdings liegen die Verhaeltnisse schon deshalb komplizierter, weil mit dem Bebriff der Legitimitaet selbst bestenfalls ein operationalisierungsbeduerftiger theoretischer Term der Soziologie vorliegt, dessen Bedeutungsbereich hoechst unklar ist (vgl. Hennis 1976).

<35>

Folgenden Primaer- und Sekundaerquellen wurden auswertbare Informationen entnommen:

Belschner/Spaeth 1977
Brunner 1978
Clemens-Lodde 1974
Coffman 1954
Cosgrove 1959
v.Cranach/Frenz 1969
Diehl/Kohr 1977
Fittkau 1972
Flanders 1970
Freedman/Stumpf 1978
Gage-Hdbch. 1971
Harari/Zedeck 1973
Hartmann 1975
Hoeder/Joost/Klyne 1975
Hofer 1969
Jahoda/Deutsch/Cook'51
Kennedy et al. 1978
Koskenniemi 1971
Lienert/Orlik 1966
van Lieshout 1973
Manning 1977
Mielke 1978
Mueller-Wolf 1977
Mueller-Wolf/Fittkau 1971
Neuberger 1972
Popp 1974
Prior 1971
Reiss jr. 1971
Ryans 1960
Schmidtke/Groffmann/Schaller'78
Schmitz 1975
Schomaker/Lehmann 1976
Schott 1973
Seitz 1977
Siebert 1977
Stegemann 1975
Steinzor 1949
Tausch/Tausch 1970(5),1973(7)
Travers; 2nd.Handbook 1973
Tscherner/Masendorf 1974
Watson/Potter 1962
Wilcke 1976
Wirth 1977

<36>

Nicht gerechnet sind dabei direkt beobachtbare Verhaltensdimensionen, wie Bewegung der Augenbrauen, Mundwinkel etc. (vgl. z.B. Grant 1969, 65 und die in Kap. 2 angegebene Literatur).

<37>

Bei der Erstellung des etwas komplizierten Erhebungsinstrumentes wurde im konventionellen Teil des Fragebogens das Item "anregend" versehentlich weggelassen, so dass alle Daten mit diesem Terminus aus dem Bedeutungsteil gestrichen werden mussten. Fuer den Vergleich mit anderen Arbeiten konnte dieses Item jedoch verwendet werden.

<38>

Obgleich wir aus den genannten Gruenden die Trennung von Global- und Partialurteilen sachlich und sprachlich fuer problematisch erachten, behalten wir diese Terminologie bei, um den begrifflichen Anschluss an die zur Debatte stehende Literatur zu halten.

<39>

In der vorliegenden Untersuchung wurde unterstellt, dass die Termini zur Haeufigkeits-/Intensitaetsbeobachtung ueberschneidungsfrei seien. Sie wurden daher auch nicht zur Beurteilung vorgelegt. Damit ergeben sich pro Beobachter ((27 x 26)/2 - (12 x 11)/2=) 285 Paarvergleiche.

<40>

Eine gewisse Schwierigkeit ergab sich aus der Notwendigkeit, die Vpn, wegen der Zweiteilung der Erhebung zu identifizieren. Wir gaben dazu bei beiden Terminen eine Karteikarte aus, die die Fragebogennummer enthielt und auf der jede Vpn, ihren Namen eintrug. So konnte am Schluss die Zuordnung vorgenommen werden. Das anfaengliche Misstrauen der Teilnehmer konnte zerstreut werden.

<41>

Die Frage nach der zu extrahierenden Faktorenzahl kann hier rein formal beantwortet werden, da es sich um eine - ebenfalls formale - Inspektion der Daten handelt.

<42>

Es kann leicht gezeigt werden, dass der von Ebel (vgl. Guilford 1954, 395) vorgeschlagene, mit dem Produkt-Moment-Koeffizienten verwandte Wert stets groesser ist als derjenige von Horst, den wir der Berechnung zugrunde legen (vgl. z.B. Langer/Schulz v.Thun 1974, 88):

$$\frac{Var(o) - Var(e)}{Var(o)} > 1 - \frac{SAQ(tot) - SAQ(o)}{(n-1)\,SAQ(o)}$$

$$\frac{Var(e)}{Var(o)} < \frac{SAQ(e) + SAQ(s)}{(n-1)\,SAQ(o)}$$

$$\frac{SAQ(e)(k-1)}{SAQ(o)(n-1)(k-1)} < \frac{SAQ(e)(k-1) + SAQ(s)(k-1)}{SAQ(o)(n-1)(k-1)},$$

fuer e: error, o: objects (hier: Merkmale),
s: systematic (hier: Rater), tot: total;
n: Zahl der Rater, k: Zahl der Objekte.

<43>

Dieser Zusammenhang wird auch aus der Beziehung zwischen r(11) und r(nn) deutlich;

$$r(11) = \frac{SAQ(E) - SAQ(e)(1-r(nn))}{SAQ(E) + SAQ(e)(1-r(nn))(n-1)}$$

fuer SAQ(E) = SAQ(s) + SAQ(e) (vgl. Anm. 42).

Man sieht, dass mit steigendem n (das auch ein Anwachsen von r(nn) mit sich bringt) der Zaehler langsamer waechst als der Nenner.

<44>

Analoge Deutungen gelten fuer Modus und Median, die ebenfalls als Korrelationsmasse betrachtet werden koennen.

<45>

Der Gedanke, dass prinzipiell auch diskutiert werden muesste, welche Probleme sich fuer den Fall ergeben, dass die in unserer Untersuchung benuetzten 27 Woerter in einem hoeherdimensionalen Raum abzubilden seien, soll hier lediglich erwaehnt werden. Er muesste im Kontext sprachtheoretischer Untersuchungen einer gruendlicheren Betrachtung unterzogen werden. Fuer den Zusammenhang der allgemeinen Modelldiskussion halten wir uns an die berichteten Ergebnisse zur Ermittlung der Dimensionalitaet der Personwahrnehmung, wonach in diesem Bereich mit weniger Dimensionen zu rechnen ist. Einschraenkend muessen wir jedoch darauf hinweisen, dass diese Dimensionen selbst nicht elementar in dem Sinne sind, wie er in (2,1,2,) eingefuehrt wurde, so dass der soeben formulierte Hinweis auch hier seine Berechtigung behalten koennte. - Wie sich weiter unten zeigen wird, versuchen wir durch das von uns eingeschlagene Verfahren dieses Problem zu umgehen.

<46>

Bei der Analyse des kollektiven Sprachraums (3,3,) waere mit der multidimensionalen Skalierung neben der Faktorenanalyse ein brauchbares Modell gegeben. Fuer die dort vorliegende Fragestellung ist seine Verwendung aber

nicht erforderlich. Wir haben unter anderen Aspekten Analysen der Daten mit diesem Modell durchgefuehrt, Ueber ihre Ergebnisse soll an anderer Stelle berichtet werden.

<47>

Die Varianten C, D und I, J duerften in dieser Hinsicht gleichrangig einzustufen sein, da sich in ihnen der Wechsel von Schaetzstrategie und Abfolgeregel jeweils konterkarieren.

<48>

Ein numerischer Vergleich mit den Ergebnissen aus den Varianten A-F waere nicht sinnvoll, weil bei ihrer Berechnung ja von voellig anderen sprach- und damit hier auch: wissenschaftstheoretischen Voraussetzungen ausgegangen wurde.

<49>

Die Groesse n muss nicht als ganzzahlige Variable interpretiert werden, wie dies im vorliegenden Falle getan wird. Allgemein drueckt sie irgendeinen beliebigen Bruchteil aus, um den ein vorliegendes Erhebungsinstrument, fuer das - hier - r(11) bekannt ist, veraendert ("verlaengert") wird.

<50>

Aus der groesseren absoluten Hoehe von r(vv) bei niedrigerem r(11) darf nicht geschlossen werden, dass geringere Werte von r(11) bezueglich der Validitaet guenstiger seien. Beachtet man naemlich das Prinzip r(vv)<=r(nn), so sieht man leicht, dass bei r(11)=,0078 keiner, bei r(11)=,5841 alle rechnerisch ermittelbaren Validitaetskoeffizienten diese Forderung erfuellen:

Werte von r(nn) fuer Tab. 3-15:

| r(11) | Beobachterzahl | | | |
|---|---|---|---|---|
| | 10 | 5 | 3 | 2 |
| .5841 | .934 | .875 | .808 | .737 |
| .0078 | .073 | .038 | .023 | .015 |

<51>

Die von Smith (1974) vorgeschlagene Strategie ist im uebrigen in einer Hinsicht als zu rigoros zu qualifizieren. Fuer ihre Anwendung muss naemlich unterstellt werden, dass der Rater als ein einzelnes Messinstrument fungiere, dass er also entweder stets oder niemals zutreffende Angaben liefere. Diese Rekonstruktion ist nicht ausreichend. Die Fehlerfrage muss zusaetzlich auf das beobachtete Merkmal relativiert werden; die Messgueteb bestimmt sich als Reliabilitaet der Masse fuer einen bestimmten Aspekt. Die Tatsache, dass ein Rater in Bezug auf den einen Gesichtspunkt nicht reliabel schaetzt, impliziert nicht, dass dies fuer alle Gesichtspunkte gelten muss. - Formal laeuft diese Ueberlegung auf die Beseitigung von Ausreisserwerten aus den fuer jedes Merkmal vorliegenden Messdaten hinaus und fuehrt zu hoeheren Reliabilitaetskoeffizienten als sie beispielsweise von de Freitas und Ribeiro (1977) gefunden werden.

<52>

Gemaess unserer spezifischen methodenkritischen Absicht verzichten wir auf die Beschaeftigung mit weiteren problematischen Punkten, die mit der Anwendung der Faktorenanalyse zusammenhaengen (Modellvoraussetzungen, Kommunalitaetenschaetzung, Rotationskriterium). Sie werden in den vorliegenden Untersuchungen haeufig nicht beachtet bzw. dogmatisch behandelt (Unterstellung der Voraussetzungsadaequatheit, des Nicht-Auftretens von Einzelrestvarianz (!) und (mit Einschraenkungen) der Moeglichkeit unabhaengiger Variationen empirischer Entitaeten). Unsere Absicht besteht gerade darin, diejenigen Probleme herauszustellen, die auch nach der Akzeptierung der "ueblichen Methoden" noch bestehen.

<53>

Dies liegt vielleicht auch daran, dass die gaengigen

Programmpakete, die auf den EDV-Anlagen von Universitaeten implementiert sind (z.B. BMD, Osiris, SPSS), standardmaessig solche "aufwendigen" Kriterien nicht berechnen.

<54>

Aus mathematischen Gruenden wurde nicht exakt r=0,0000 eingesetzt, sondern zufaellige Abweichungen davon im Schwankungsbereich -0,0020<r<+0,0020.

<55>

Die Deutung der Sprachaehnlichkeitswerte muendet in Aussagen ueber Strukturen der Bedeutungsueberschneidung, die der Beobachtungswerte in Aussagen ueber Ursachen/Charakteristika (vgl. 3,4,3,2, am Anfang) der (beobachteten) Realitaet.

<56>

Zur sprachtheoretischen Bedeutung der einzelnen Mittelwerte vgl. Kap. 3,3..

<57>

Die Interpretation einer Faktorenanalyse, die auf der Basis von Termini hoher Inferenzstufe Aussagen niedrigeren Inferenzniveaus macht, waere durch Hinzufuegen von Informationsgehalt zustande gekommen und daher logisch unzulaessig.

<58>

Welche Aspekte der Persoenlichkeit durch sie erfasst werden und ob dies auf die methodisch guenstigste Weise geschieht, bleibt unklar, braucht aber hier nicht weiter diskutiert zu werden.

<59>

Die 12 verschiedenen Schaetzvarianten sind auf die Inkonsistenz der Abbildung im euklidischen Distanzmodell abgestellt und nicht auf deren Beseitigung. Daher entstehen auch nicht streng dimensionsreine Matrizen mit minimalem Rang.

# L I T E R A T U R V E R Z E I C H N I S

## COMPUTER - PROGRAMME

CRAMER ET AL., MANOVA, VERS. RECHENZENTRUM UNIVERSITAET MANNHEIM (WERNER)

NIE, N.H. ET AL., SPSS, NEW YORK

RECHENZENTRUM UNIVERSITAET MANNHEIM, RUNOFF; TEXTAUFBEREITUNG

WISHART, D., CLUSTAN; CLUSTERANALYSE, EDINBURGH

ALLE WEITEREN PROGRAMME ZUR AUFBEREITUNG UND AUSWERTUNG DER DATEN WURDEN VOM VERF. ERSTELLT.

## LITERATUR

ADORNO, THEODOR W., DER POSITIVISMUSSTREIT IN DER DEUTSCHEN SOZIOLOGIE (1969), IN: BLUMENBERG, HANS U.A.(HRSG.), THEODOR W. ADORNO: AUFSAETZE ZUR GESELLSCHAFTSTHEORIE UND METHODOLOGIE, (SUHRKAMP) FRANKFURT 1970, 167-245

ADORNO, THEODOR W. U.A., DER POSITIVISMUSSTREIT IN DER DEUTSCHEN SOZIOLOGIE, (LUCHTERHAND) NEUWIED 1972

ALBERT, HANS, WISSENSCHAFT UND POLITIK, ZUM PROBLEM DER ANWENDBARKEIT EINER WERT- FREIEN SOZIALWISSENSCHAFT, IN: TOPITSCH, E.(HRSG), PROBLEME DER WISSENSCHAFTSTHEORIE, (SPRINGER) WIEN 1960, 201-232

ALBERT, HANS, TOPITSCH, ERNST, WERTURTEILSSTREIT, (WISS. BUCHGES.) DARMSTADT 1971

ALEAMONI, LAWRENCE AND SPENCER, RICHARD E., THE ILLINOIS COURSE EVALUATION QUESTIONNAIRE: A DESCRIPTION OF ITS DEVELOPMENT AND A REPORT OF SOME OF ITS RESULTS,IN: EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 33, 1973, 669-684

ALLERBECK, KLAUS R., MESSNIVEAU UND ANALYSEVERFAHREN - DAS PROBLEM "STRITTIGER INTERVALLSKALEN", IN: ZEITSCHRIFT FUER SOZIOLOGIE, 7, 1978, 199 - 214

ALTMAN, IRWIN AND HAYTHORN, WILLIAM W., THE ECOLOGY OF ISOLATED GROUPS, IN : BEHAVIORAL SCIIENCE, 12, 1967, 169 - 182

APEL, KARL OTTO, DIE KOMMUNIKATIONSGEMEINSCHAFT ALS

TRANSZENDENTALE VORAUSSETZUNG DER SOZIALWISSENSCHAFTEN, IN: NEUE HEFTE FUER PHILOSOPHIE, 1972, 2/3, 1-40

ARGYLE, MICHAEL, THE PSYCHOLOGY OF INTERPERSONAL BEHAVIOUR, (COX AND WYMAN) LONDON 1967

AUSTIN, JOHN L., ZUR THEORIE DER SPRECHAKTE, (RECLAM) STUTTGART 1976

BARBER, THEODORE XENOPHON, PITFALLS IN RESEARCH: NINE INVESTIGATOR AND EXPERIMENTER EFFECTS, IN: TRAVERS, ROBERT (ED.), SECOND HANDBOOK OF RESEARCH ON TEACHING,(RAND MCNALLY) CHICAGO 1973, 382-404

BAXTER, JAMES C. AND WINTERS, ELAINE P., GESTURAL BEHAVIOR DURING A BRIEF INTERVIEW AS A FUNCTION OF COGNITIVE VARIABLES, IN: JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY, 8, 1968, 303 - 307

BECKER, GEORG E., MICROTEACHING: TRAINING DES FRAGEVERHALTENS, IN: PROGRAMMIERTES LERNEN, 4, 1971, 174 - 183

BECKER, WESLEY C., CONSEQUENCES OF DIFFERENT KINDS OF PARENTAL DISCIPLINE, IN: HOFFMAN, MARTIN L. AND HOFFMAN, LOIS WLADIS (EDS.), REVIEW OF CHILD DEVELOPMENT RESEARCH, (RUSSEL SAGE FOUNDATION) NEW YORK 1964, 169-208

BELSCHNER, WILFRIED, SPAETH, HELMUTH, VERSUCH EINER KATEGORISIERUNG VON ERZIEHERISCHEN SITUATIONSDEFINITIONEN MITTELS CLUSTER-ANALYSE, IN: PSYCHOLOGIE IN ERZIEHUNG UND UNTERRICHT, 24, 1977, 49-53

BIDDLE, BRUCE J., THE INTEGRATION OF TEACHER EFFECTIVENESS RESEARCH, IN: DERS. UND ELLENA, W.J., CONTEMPORARY RESEARCH ON TEACHER EFFECTIVENESS, NEW YORK 1964, 1-40

BILLETER, ERNST PETER, GRUNDLAGEN DER ERFORSCHENDEN STATISTIK, (SPRINGER) WIEN 1972

BLALOCK JR., HUBERT, THEORIE, MESSUNG UND DIE REPRODUKTION VON EFFEKTEN IN DEN SOZIALWISSENSCHAFTEN, IN: ACHAM, KARL (HRSG.), METHODOLOGISCHE PROBLEME DER SOZIALWISSENSCHAFTEN, (WISS. BUCHGES.) DARMSTADT 1978, 403 - 415

BLASS, THOMAS (ED.), PERSONALITY VARIABLES IN SOCIAL BEHAVIOR, (LAWRENCE ERLBAUM ASS.) HILLSDALE, N.J. 1977

BOHNEN, ALFRED, ZUR KRITIK DES MODERNEN EMPIRISMUS,

BEOBACHTUNGSSPRACHE, BEOBACHTUNGSTATSACHEN UND THEORIEN, IN: ALBERT, HANS (HRSG.), THEORIE UND REALITAET, 2, AUFL., (MOHR) TUEBINGEN 1972, 171-190

BORNEWASSER, MANFRED, NATURWISSENSCHAFTLICHE UND VERHALTENSTHEORETISCHE ORIENTIERUNGEN IN DER SOZIALPSYCHOLOGIE, BIELEFELDER ARBEITEN ZUR SOZIALPSYCHOLOGIE, BERICHT NR. 4, APRIL 1976

BORNEWASSER, MANFRED, DIE KONSISTENZBEZIEHUNG ZWISCHEN EINSTELLUNGEN UND OFFENEM VERHALTEN, BIELEFELDER ARBEITEN ZUR SOZIALPSYCHOLOGIE, BERICHT NR. 16, FEBRUAR 1977

BRACHT, GLENN H., EXPERIMENTAL FACTORS RELATED TO APTITUDE TREATMENT INTERACTIONS, IN: REVIEW OF EDUCATIONAL RESEARCH, 40, 1970, 627 - 645

BRANNIGAN, CHRISTOPHER R., HUMPHRIES, DAVID A., HUMAN NON - VERBAL BEHAVIOR, A MEANS OF COMMUNICATION, IN: JONES, N. BLURTON (ED.), ETHOLOGICAL STUDIES OF CHILD BEHAVIOR, (UNIV. PRESS) CAMBRIDGE 1972, 37-64

BREDENKAMP, JUERGEN, DER SIGNIFIKANZTEST IN DER PSYCHOLOGISCHEN FORSCHUNG, (AKAD. VERL.GES.) FRANKFURT 1972

BRITTON, BRUCE K., LEXICAL AMBIGUITY OF WORDS USED IN ENGLISH TEXT, IN: BEHAVIOR RESEARCH METHODS & INSTRUMENTATION, 10, 1978, 1-7

BRUNNER, REINHARD, LEHRERVERHALTEN, (SCHOENINGH) PADERBORN 1978

BRYAN, ROY C., STUDENT RATING OF TEACHERS, IN: IMPROVING COLLEGE AND UNIVERSITY TEACHING, 16, 1968, 200 - 202

CAMPBELL, DONALD T., SYSTEMATIC ERROR ON THE PART OF HUMAN LINKS IN COMMUNICATION SYSTEMS, INFORMATION AND CONTROL, 1958

CARNAP, RUDOLPH, MEANING AND NECESSITY, (UNIV. OF. CHICAGO PRESS) CHICAGO 1947 (2. AUFL. 1956)

CATTELL, RAYMOND B., THE SCREE TEST FOR THE NUMBER OF FACTORS, IN: MULTIVARIATE BEHAVIORAL RESEARCH, 1, 1966, 245-276

CHAPPLE, E.D. ET AL., THE MEASUREMENT OF ACTIVITY PATTERNS OF SCHIZOPHRENIC PATIENTS, IN: JOURNAL OF NERVOUS AND MENTAL DISEASE, VOL. 137, 1960, 258-267

CICOUREL, AARON V., METHODE UND MESSUNG IN DER SOZIOLOGIE, (SUHRKAMP) FRANKFURT 1970

CLEMENS-LODDE, BEATE, EINSTELLUNGEN VON LEHRENDEN ZU INTRINSISCH UND EXTRINSISCH MOTIVIERENDEM LEHRVERHALTEN, (DISS.) MUENSTER 1974

CLIFFORD, GERALDINE J., A HISTORY OF THE IMPACT RESEARCH ON TEACHING, IN: TRAVERS, RICHARD M.W. (ED), SECOND HANDBOOK OF RESEARCH ON TEACHING, (RAND MC.NALLY) CHICAGO 1973, 1-46

COFFMAN, WILLIAM E., DETERMINING STUDENTS' CONCEPTS OF EFFECTIVE TEACHING FROM THEIR RATINGS OF INSTRUCTORS, IN: THE JOURNAL OF EDUCATIONAL PSYCHOLOGY, 45, 1954, 277-286

CONDON, W.S. AND OGSTON, W.D., SOUND FILM ANALYSIS OF NORMAL AND PATHOLOGICAL BEHAVIOUR PATTERNS, IN: THE JOURNAL OF NERVOUS AND MENTAL DISEASE, VOL. 143, 1966, 338 - 347

CORRELL, WERNER, DAS PAEDAGOGISCH-PSYCHOLOGISCHE PROBLEM DER BEZIEHUNG ZWISCHEN LEHRSTIL UND LERNLEISTUNG, IN: HERRMANN, THEO (HRSG.), PSYCHOLOGIE DER ERZIEHUNGSSTILE, GOETTINGEN 1966, 225 - 242

COSGROVE, DON J., DIAGNOSTIC RATING OF TEACHER PERFORMANCE, IN: JOURNAL OF EDUCATIONAL PSYCHOLOGY, VOL. 50, 1959, 200 - 204

V. CRANACH, MARIO, FRENZ, HANS-GEORG, SYSTEMATISCHE BEOBACHTUNG, IN: GRAUMANN, C. F. (HRSG.), SOZIALPSYCHOLOGIE, 1. HALBBAND, THEORIEN UND METHODEN, 7. BAND DES HANDBUCHS DER PSYCHOLOGIE, (HOGREFE) GOETTINGEN 1969, 269-331

DAVIS, HAZEL, EVOLUTION OF CURRENT PRACTICES IN EVALUATING TEACHER COMPETENCE, IN: BIDDLE, BRUCE J., ELLENA, WILLIAM J. (EDS.), CONTEMPORARY RESEARCH ON TEACHER EFFECTIVENESS, (HOLT, RINEHART, WINSTON) NEW YORK 1964, 41-66

DIEHL, JOERG M. UND KOHR, HEINZ-U., ENTWICKLUNG EINES FRAGEBOGENS ZUR BEURTEILUNG VON HOCHSCHULVERANSTALTUNGEN IM FACH PSYCHOLOGIE, IN: PSYCHOLOGIE IN ERZIEHUNG UND UNTERRICHT, 24, 1977, 61-75

DOERING, KLAUS W., LEHRERVERHALTEN UND LEHRERBERUF, 5. AUFL., (BELTZ) WEINHEIM 1973

EBEL, ROBERT L., ESTIMATION OF THE RELIABILITY OF

RATINGS, IN: PSYCHOMETRIKA, 16, 1951, 407-424

ENGELKAMP, JOHANNES, PSYCHOLINGUISTIK, (FINK) MUENCHEN 1974

ESSLER, WILHELM K., ANALYTISCHE PHILOSOPHIE 1, (KROENER) STUTTGART 1972

EYFERTH, K., METHODEN ZUR ERFASSUNG VON ERZIEHUNGSSTILEN, IN: HERRMANN, THEO (HRSG.), PSYCHOLOGIE DER ERZIEHUNGSSTILE, (HOGREFE) GOETTINGEN 1966, 17-31

FALTER, JUERGEN W., ZUR VALIDIERUNG THEORETISCHER KONSTRUKTE: WISSENSCHAFTSTHEORETISCHE ASPEKTE DES VALIDIERUNGSKONZEPTS, IN: ZEITSCHRIFT FUER SOZIOLOGIE, 6, 1977, 370-385

FASSNACHT, GERHARD, SYSTEMATISCHE VERHALTENSBEOBACHTUNG, (REINHARDT) MUENCHEN 1979

FEGER, HUBERT UND VAN TROTSENBERG, EDMUND, PARADIGMEN FUER DIE UNTERRICHTSFORSCHUNG, IN: INGENKAMP, KARLHEINZ UND PAREY, EVELORE (HRSG.), HANDBUCH DER UNTERRICHSTFORSCHUNG, TEIL I, (BELTZ) WEINHEIM 1970, SP. 269 - 366

FENNER, HANS-JOERG, VERFAHREN UND ERGEBNISSE ZUR OBJEKTIVIERUNG DES LEHRVERHALTENS, IN: NICKEL, HORST, LANGHORST, ERICH (HRSG.), BRENNPUNKTE DER PAEDAGOGISCHEN PSYCHOLOGIE, (HUBER) BERN 1973, 123 - 133

FITTKAU, BERND, KOMMUNIKATIONS- UND VERHALTENSTRAINING FUER ERZIEHER, IN: GRUPPENDYNAMIK, 3, 1972, 252-274

FLANDERS, NED A., ANALYZING TEACHING BEHAVIOR, (ADDISON - WESLEY) READING, MASS, 1970

FRAMO, JAMES L, AND ADLERSTEIN, ARTHUR M., A BEHAVIORAL DISTURBANCE INDEX FOR PSYCHIATRIC PATIENT AND WARD DISTURBANCE, IN: JOURNAL OF CLINICAL PSYCHOLOGY, 17, 1961, 260 - 264

FRANK, HELMAR, DERZEITIGE BEMUEHUNGEN UM ERWEITERUNG DES INFORMATIONSPSYCHOLOGISCHEN MODELLS, IN: GRUNDLAGENSTUDIEN AUS KYBERNETIK UND GEISTESWISSENSCHAFT, 18, 1977, 61 - 72

FREEDMAN, RICHARD D., STUMPF, STEPHEN A., STUDENT EVALUATIONS OF COURSES AND FACULTY BASED ON A PERCEIVED LEARNING CRITERION: SCALE CONSTRUCTION, VALIDATION, AND COMPARISON OF RESULTS, IN: APPLIED PSYCHOLOGICAL

MEASUREMENT, 2, 1978, 189-202

DE FREITAS, ANTONIO C., RIBEIRO, GUILHERME H.C., A NEW APPROACH TO DETERMINE THE RELIABILITY OF THE RATER TECHNIQUE BY USE OF SMITH'S RATING SCALES, IN: JOURNAL OF CLINICAL PSYCHOLOGY, 33, 1977, 1032-1035

FRIEDRICHS, JUERGEN, LUEDTKE, HARTMUT, TEILNEHMENDE BEOBACHTUNG, 2. AUFL., (BELTZ) WEINHEIM 1973

FUHRMANN, MANFRED UND EWERT, OTTO, DIE VARIABLE LEHRERVERHALTEN - VERSUCH EINER OPERATIONALEN DEFINITION, IN: BILDUNG UND ERZIEHUNG, 26, 1973, 423 - 432

GERLACH, ANNELIESE UND HOFER, MANFRED, DER EINFLUSS VON STRUKTURIERUNGSHILFEN AUF DAS ERLERNEN EINES STUDIENTEXTES, IN: ZEITSCHRIFT FUER ENTWICKLUNGSPSYCHOLOGIE UND PAEDAGOGISCHE PSYCHOLOGIE, 5, 1973, 91 - 105

GETHMANN, CARL FRIEDRICH, STICHWORT "REALITAET", IN: KRINGS, HERRMANN, BAUMGARTNER, HANS MICHAEL, WILD, CHRISTOPH (HRSG), HANDBUCH DER PHILOSOPHISCHEN GRUNDBEGRIFFE, BD. II, (KOESEL) MUENCHEN 1973, 1168 - 1187

GLUECK, GERHARD, METHODEN DER BEOBACHTUNG, IN: DOHMEN, GUENTHER (HRSG.), FORSCHUNGSTECHNIKEN FUER DIE HOCHSCHULDIDAKTIK, (BECK) MUENCHEN 1971, 57-66

GRAUMANN, CARL FRIEDRICH, GRUNDZUEGE DER VERHALTENSBEOBACHTUNG, IN: DERS. UND HECKHAUSEN, HEINZ (HRSG.), ENTWICKLUNG UND SOZIALISATION, READER ZUM FUNK-KOLLEG PAEDAGOGISCHE PSYCHOLOGIE, BD. 1, (FISCHER) FRANKFURT 1974, 14-41

GRELL, JOCHEN, TECHNIKEN DES LEHRERVERHALTENS, (BELTZ) WEINHEIM, 1974

GRUEMER, KARL-WILHELM, BEOBACHTUNG, (TEUBNER) STUTTGART 1974

HABERMAS, JUERGEN, ERKENNTNIS UND INTERESSE, IN: ALBERT, HANS UND TOPITSCH, ERNST (HRSG), WERTURTEILSSTREIT, (WISS. BUCHGES.) DARMSTADT 1971, 334 - 352

HALL, EDWARD T., A SYSTEM FOR THE NOTATION OF PROXEMIC BEHAVIOR, IN: AMERICAN ANTHROPOLOGIST, 65, 1963, 1003 - 1026

HARARI, OREN AND ZEDECK, SHELDON, DEVELOPMENT OF

BEHAVIORALLY ANCHORED SCALES FOR THE EVALUATION OF FACULTY TEACHING, IN: JOURNAL OF APPLIED PSYCHOLOGY, 58, 1973, 261-265

HARMAN, HARRY H., MODERN FACTOR ANALYSIS, 3RD. ED., (UNIVERSITY OF CHICAGO PRESS) CHICAGO 1976

HARTMANN, DONALD P., CONSIDERATIONS IN THE CHOICE OF INTEROBSERVER RELIABILITY ESTIMATES, IN: JOURNAL OF APPLIED BEHAVIOR ANALYSIS, 10, 1977, 103-116

HARTMANN, HANS-PETER, DIMENSIONSANALYSEN SOZIALER EINSTELLUNGEN UND PERSOENLICHKEITSVARIABLEN BEI LEHRERN, (UNVEROEFFENTL. DIPL.-ARBEIT) UNIVERSITAET GIESSEN 1975

HASEMANN, KLAUS, VERHALTENSBEOBACHTUNG, IN: HEISS, ROBERT (HRSG.), 6.BAND DES HANDBUCHS DER PSYCHOLOGIE, PSYCHOLOGISCHE DIAGNOSTIK, 3. AUFL., (HOGREFE) GOETTINGEN 1971, 807-836

HEIDENREICH, WOLF-DIETER UND HEYMANN, HANS WERNER, LEHR-LERN-FORSCHUNG (NEUERE UNTERRICHTSWISSENSCHAFTLICHE LITERATUR IM SPIEGEL EINES NEUEN FORSCHUNGSANSATZES), IN: ZEITSCHRIFT FUER PAEDAGOGIK, 22, 1976, 225 - 251

HENNIS, WILHELM, LEGITIMITAET - ZU EINER KATEGORIE DER BUERGERLICHEN GESELLSCHAFT, IN: GRAF KIELMANSEGG, PETER (HRSG.), LEGITIMATIONSPROBLEME POLITISCHER SYSTEME, SONDERHEFT 7,1976 DER POLITISCHEN VIERTELJAHRESSCHRIFT, (WESTDEUTSCHER VERLAG) OPLADEN 1976, 9-38

HERRMANN, THEO, PERSOENLICHKEITSMERKMALE, (KOHLHAMMER) STUTTGART 1973

HERRMANN, THEO, LEHRBUCH DER EMPIRISCHEN PERSOENLICHKEITSFORSCHUNG, 3. AUFLAGE, (HOGREFE) GOETTINGEN 1976

HERRMANN, THEO, DEUTSCH, WERNER, PSYCHOLOGIE DER OBJEKTBENENNUNG, (HUBER) BERN 1976

HOEDER, JUERGEN, JOOST, HARTMUT, KLYNE, PETER, ZUSAMMENHAENGE ZWISCHEN HAUPTDIMENSIONEN DES LEHRERVERHALTENS UND MERKMALEN DES ERLEBENS VON SCHUELERN IM UNTERRICHT, 22, 1975, 88-96

HOFER, MANFRED, DIE SCHUELERPERSOENLICHKEIT IM URTEIL DES LEHRERS, (BELTZ) WEINHEIM 1969

HOFSTAETTER, PETER R., EINFUEHRUNG IN DIE SOZIALPSYCHOLOGIE, (KROENER) STUTTGART 1973

HUTT, C. AND HUTT, S.J. (HRSG), BEHAVIORAL RESEARCH IN PSYCHIATRY, OXFORD 1969

HYMANN, RAY, THE NATURE OF PSYCHOLOGICAL INQUIRY, (PRENTICE HALL) ENGLEWOOD CLIFFS, N. J. 1964

IRLE, MARTIN, LEHRBUCH DER SOZIALPSYCHOLOGIE, (HOGREFE) GOETTINGEN 1975

JAHODA, MARIE, DEUTSCH, MORTON, COOK, STUART W., RESEARCH METHODS IN SOCIAL RELATIONS, PART ONE, BASIC PROCESSES, (DRYDEN) NEW YORK 1951

JANSSEN, MANFRED, DER PROZESS DER SELBSTSTEUERUNG, (UNVEROEFFENTLICHTE ARBEIT) MUENSTER 1977

JOERG, SABINE, ASPEKTE DER AUFMERKSAMKEIT, IN: FERNSEHEN UND BILDUNG, 11, 1977, 7-25

KEIL, WOLFGANG UND PIONTKOWSKI, URSULA, STRUKTUREN UND PROZESSE IM HOCHSCHULUNTERRICHT, (BELTZ) WEINHEIM 1973

KEMPF, WILHELM F., LEHRKE, M., SUBJECT MATTER DIRECTED MOTIVATION AND ITS EVALUATION BY MEANS OF QUESTIONAIRES, IN: STUDIES IN EDUCATIONAL EVALUATION 1975,2

KENNEDY, JOHN J. ET AL., ADDITIONAL INVESTIGATIONS INTO THE NATURE OF TEACHER CLARITY, IN: JOURNAL OF EDUCATIONAL RESEARCH, 1978, 3-10

KLAFKI, WOLFGANG, ASPEKTE KRITISCH-KONSTRUKTIVER ERZIEHUNGSWISSENSCHAFT, (BELTZ) WEINHEIM 1976

KLEINE, DIETMAR, MERKENS, HANS, UEBERPRUEFUNG EINES FRAGEBOGENS ZUR BEURTEILUNG VON LEHRVERANSTALTUNGEN, IN: PSYCHOLOGIE IN ERZIEHUNG UND UNTERRICHT, 26, 1979, 149-153

KLEITER, EKKEHARD F., UEBER BEDINGUNGEN DER INTERNEN VERKNUEPFUNGSSTRUKTUR VON URTEILSBEGRIFFEN BEI DER PERSONENBEURTEILUNG, IN: UNTERICHTSWISSENSCHAFT 3, 1975, 52 - 83

KOLAKOWSKI, LESZEK, DIE PHILOSOPHIE DES POSITIVISMUS (1966), (PIPER) MUENCHEN 1971

KOLB, GUENTER, INTERVIEW UND FRAGEBOGEN, IN: DOHMEN, G. (HRSG.), FORSCHUNGSTECHNIKEN FUER DIE HOCHSCHULDIDAKTIK, (BECK) MUENCHEN 1971, 67-78

KOSKENNIEMI, MATTI, ELEMENTE DER UNTERRICHTSTHEORIE,

(EHRENWIRTH) MUENCHEN 1971

KUEHN, WOLFGANG, EINFUEHRUNG IN DIE MULTIDIMENSIONALE SKALIERUNG, (REINHARDT) MUENCHEN 1976

LANGER, INGHARD, SCHULZ VON THUN, FRIEDEMANN, MESSUNG KOMPLEXER MERKMALE IN PSYCHOLOGIE UND PAEDAGOGIK, HEFT 68 DER BEIHEFTE DER ZEITSCHRIFT: PSYCHOLOGIE IN ERZIEHUNG UND UNTERRICHT, (REINHARDT) MUENCHEN 1974

LAZARSFELD, PAUL FELIX, EVIDENCE AND INFERENCE IN SOCIAL RESEARCH, IN: LERNER, DANIEL (ED.), EVIDENCE AND INFERENCE, NEW YORK 1959

LEHRKE, M., KEMPF, WILHELM F., SACHBEZOGENE MOTIVATION IM SCHULUNTERRICHT UND DIE MESSUNG IHRER VERAENDERUNG, IN: BERICHT UEBER DEN 29. KONGRESS DER DT. GESELLSCHAFT F. PSYCHOLOGIE, (HOGREFE) GOETTINGEN 1975

LENNEBERG, ERIC H., BIOLOGISCHE GRUNDLAGEN DER SPRACHE, (SUHRKAMP) FRANKFURT 1977

LEPSIUS, RAINER M., GESELLSCHAFTSANALYSE UND SINNGEBUNGSZWANG, IN: ALBRECHT, GUENTER U.A. (HRSG.), SOZIOLOGIE (FESTSCHRIFT FUER RENE KOENIG), (WESTDTSCH. VERL.) OPLADEN 1973

LIENERT, GUSTAV A., TESTAUFBAU UND TESTANALYSE, 3. AUFL., (BELTZ) WEINHEIM 1969

LIENERT, GUSTAV A., VERTEILUNGSFREIE METHODEN IN DER BIOSTATISTIK, BD. 1, 2. AUFL., (HAIN) MEISENHEIM AM GLAN 1973

LIENERT, GUSTAV A., ORLIK, P., KINDLICHE VERHALTENSSTOERUNGEN IM SPIEGEL EINES ELTERNFRAGEBOGENS, IN: FOERSTER, ECKARD, WEWETZER, KARL-HERMANN (HRSG.), JUGENDPSYCHIATRISCHE UND PSYCHOLOGISCHE DIAGNOSTIK, (HUBER) BERN 1966, 59-66

VAN LIESHOUT, KEES F.M., ENTWICKLUNG VON BEURTEILUNGSSKALEN FUER DAS SOZIALE VERHALTEN VON KINDERN, IN: REINERT, GUENTHER, (HRSG), BERICHT UEBER DEN 27. KONGRESS DER DT. GESELLSCHAFT FUER PSYCHOLOGIE IN KIEL 1970, (HOGREFE) GOETTINGEN 1973

LIGHT, RICHARD J., ISSUES IN THE ANALYSIS OF QUALITATIVE DATA, IN: TRAVERS, ROBERT M. W. (ED.), SECOND HANDBOOK OF RESEARCH ON TEACHING, (RAND MCNALLY) CHICAGO 1973, 318-381

LIND, GUNTER, SACHBEZOGENE MOTIVATION IM

NATURWISSENSCHAFTLICHEN UNTERRICHT, (BELTZ) WEINHEIM 1975

LORD, FREDERIC M., NOVICK, MELVIN R., STATISTICAL THEORIES OF MENTAL TEST SCORES, (ADDISON-WESLEY) READING, MASS, 1968

LOSER, FRITZ, LEHRTHEORIE ALS THEORIE DER LEHR- UND LERNSITUATION, IN: BILDUNG UND ERZIEHUNG, 30, 1977, 419 - 425

LYONS, JOHN, SEMANTICS, VOL. 1, (UNIVERSITY PRESS), CAMBRIDGE 1977

MACH, ERNST, ERKENNTNIS UND IRRTUM, 6. AUFL., (WISS. BUCHGES.; REPOGRAPH. NACHDRUCK DER 5. AUFL.), DARMSTADT 1968

MANNING, SUSAN KARP, RATINGS OF THE AUDITORY AND VISUAL SIMILARITY OF CONSONANTS: IMPLICATIONS FOR RESEARCH, IN: BEHAVIOR RESEARCH METHODS AND INSTRUMENTATION, 10, 1978, 1-7

MANSTETTEN, RUDOLF, DIE IMPULSGEBUNG DES LEHRERS UND IHRE AUSWIRKUNGEN AUF DAS SCHUELERVERHALTEN - EIN UNTERRICHTSEXPERIMENT, IN: DEUTSCHE BERUFS- UND FACHSCHULE, 73, 1977, 523-531

MANZ, WOLFGANG, DIE BEOBACHTUNG VERBALER KOMMUNIKATION IM LABORATORIUM, IN: KOOLWIJK, JUERGEN VAN, WIEKEN-MAYSER, MARIA, TECHNIKEN DER EMPIRISCHEN SOZIALFORSCHUNG, BD.3, (OLDENBOURG), MUENCHEN 1974, 27-65

MEDLEY, DONALD M. AND MITZEL, HAROLD E., APPLICATION OF ANALYSIS OF VARIANCE TO THE ESTIMATION OF THE RELIABILITY OF OBSERVATIONS OF TEACHERS' CLASSROOM BEHAVIOR, IN: JOURNAL OF EXPERIMENTAL EDUCATION, 27, 1958, 23 - 35

MERKENS, HANS, PROBLEME UND SCHWIERIGKEITEN BEI DER BEOBACHTUNG ALS EINER EMPIRISCHEN METHODE, IN: PL, 9, 1972, 75-82

MERKENS, HANS, UEBERLEGUNGEN ZU EINER THEORIE DER BEOBACHTUNG, IN: UNTERRICHTSWISSENSCHAFT 1974, 2, 14-20

MESCHKOWSKI, HERBERT, ELEMENTE DER MODERNEN MATHEMATIK, IN: DERS. (HRSG.), MEYERS HANDBUCH UEBER DIE MATHEMATIK, 2. AUFL., MANNHEIM 1972, 13-40

MICHEL, LOTHAR, ALLGEMEINE GRUNDLAGEN PSYCHOMETRISCHER

TESTS, IN: HEISS, ROBERT (HRSG.), PSYCHOLOGISCHE DIAGNOSTIK, 6. BAND DES HANDBUCHS DER PSYCHOLOGIE, 3. AUFLAGE, (HOGREFE) GOETTINGEN 1971, 19-70

MIELKE, ROSEMARIE, EINSTELLUNGEN UND VERHALTEN BEI LEHRERN UNTER BERUECKSICHTIGUNG VON INTERNER/EXTERNER KONTROLLE UND MERKMALEN DER SCHULUMWELT, BIELEFELDER ARBEITEN ZUR SOZIALPSYCHOLOGIE , NR. 28, JANUAR 1978

MILLER, GEORGE A., JOHNSON-LAIRD, PHILIP N., LANGUAGE AND PERCEPTION, (UNIVERSITY PRESS), CAMBRIDGE 1976

MITTENECKER, ERICH, SUBJEKTIVE TESTS ZUR MESSUNG DER PERSOENKICHKEIT, IN: HEISS, ROBERT (HRSG.), PSYCHOLOGISCHE DIAGNOSTIK, BD. 6 DES HANDBUCHS DER PSYCHOLOGIE, 3. AUFL., (HOGREFE), GOETTINGEN 1971, 461-487

MOLLENHAUER, KLAUS, RITTELMEYER, CHRISTIAN, METHODEN DER ERZIEHUNGSWISSENSCHAFT, (JUVENTA) MUENCHEN 1977

MOSER, HEINZ, AKTIONSFORSCHUNG ALS KRITISCHE THEORIE DER SOZIALWISSENSCHAFTEN, (KOESEL) MUENCHEN 1975

MUELLER-WOLF, HANS-MARTIN, LEHRVERHALTEN AN DER HOCHSCHULE, (DOKUMENTATION) MUENCHEN 1977

MUELLER-WOLF, HANS-MARTIN UND FITTKAU, BERND, LEHRVERHALTEN VON HOCHSCHULLEHRERN UND SEINE BEDEUTUNG FUER EINSTELLUNGEN UND VERHALTEN VON STUDENTEN, IN: ZEITSCHRIFT FUER ENTWICKLUNGSPSYCHOLOGIE UND PAEDAGOGISCHE PSYCHOLOGIE, 3, 1971, 165-180

MUMMENDEY, AMELIE UND MUMMENDEY, HANS DIETER, BEGRIFF, MESSUNG UND VERHALTENSRELEVANZ SOZIALER EINSTELLUNGEN, BIELEFELDER ARBEIEN ZUR SOZIALPSYCHOLOGIE, BERICHT NR. 21, AUGUST 1977

MUMMENDEY, HANS DIETER U.A., UNTERSUCHUNGEN MIT EINEM MEHRDIMENSIONALEN SELBSTEINSCHAETZUNGSVERFAHREN, BIELEFELDER ARBEITEN ZUR SOZIALPSYCHOLOGIE, BERICHT NR. 14, JANUAR 1977

MUMMENDEY, HANS DIETER, EINSTELLUNGEN (SETS) BEI DER ERFORSCHUNG DER BEZIEHUNG ZWISCHEN EINSTELLUNGEN (ATTITUDES) UND OFFENEM VERHALTEN, BIELEFELDER ARBEITEN ZUR SOZIALPSYCHOLOGIE, BERICHT NR. 17, MAERZ 1977

NATSOULAS, THOMAS, INTERPRETING PERCEPTUAL REPORTS, IN: PSYCHOLOGICAL BULLETIN, 70, 1968, 575-591

NEUBERGER, OSWALD, EXPERIMENTELLE UNTERSUCHUNG VON

FUEHRUNGSSTILEN, IN: GRUPPENDYNAMIK 1972, HEFT 2, S. 192 - 219

NICKEL, HORST, DIE LEHRER-SCHUELER-BEZIEHUNG AUS DER SICHT NEUERER FORSCHUNGSERGEBNISSE, IN: PSYCHOLOGIE IN ERZIEHUNG UND UNTERRICHT, 23, 1976, 153-172

NICKLIS, WERNER S., GLANZ UND ELEND DES NEO-EMPIRISMUS IN DER UNTERRICHTSFORSCHUNG, ANMERKUNGEN ZU HELLMUTH WALTERS "EINFUEHRUNG IN DIE UNTERRICHTSFORSCHUNG", IN: ZEITSCHRIFT FUER PAEDAGOGIK, 24, 1978, 571-579

NICKLIS, WERNER S., ZUR SELBSTVERTEIDIGUNG DER "EXPERIMENTELLEN UNTERRICHTSFORSCHUNG", EINE REPLIK AUF H. WALTERS ERWIDERUNG, IN: ZEITSCHRIFT FUER PAEDAGOGIK, 25, 1979, 313-318

NIETEN, HEINZ-DIETER, OBJEKTIVIERUNG DER UNTERRICHTSBEURTEILUNG DURCH KRITERIENGRUPPEN, IN: WESTERMANNNS PAEDAGOGISCHE BEITRAEGE, 28, 1976, 644 - 645

NUTHALL, GRAHAM, SNOOK, IVAN, CONTEMPORARY MODELS OF TEACHING, IN: TRAVERS, ROBERT M.W. (ED.), SECOND HANDBOOK OF RESEARCH ON TEACHING, (RAND MCNALLY) CHICAGO 1973, 1-46

OERTER, ROLF, MODERNE ENTWICKLUNGSPSYCHOLOGIE, 12. AUFL., (AUER), DONAUWOERTH 1973

OEVERMANN, ULRICH U.A., BEOBACHTUNGEN ZUR STRUKTUR DER SOZIALISATORISCHEN INTERAKTION, IN: AUWAERTER, M., KIRSCH, E., SCHROETER, M. (HRSG.), SEMINAR: KOMMUNIKATION, INTERAKTION, IDENTITAET, FRANKFURT 1976, 371 - 403

OPP, KARL DIETER, METHODOLOGIE DER SOZIALWISSENSCHAFTEN, (ROWOHLT) REINBEK 1978

PAUSE, GERHARD, MERKMALE DER LEHRERPERSOENLICHKEIT, IN: INGENKAMP, KARLHEINZ UND PAREY, EVELORE (HRSG.), HANDBUCH DER UNTERRICHTSFORSCHUNG, TEIL II, (BELTZ) WEINHEIM 1970, SP. 1352 - 1526

PHILLIPS, BERNARD S., EMPIRISCHE SOZIALFORSCHUNG, (SPRINGER), WIEN 1970

PINTHER, ARNOLD, BEOBACHTUNG, IN: FRIEDRICH, WALTER, HENNING, WERNER (HRSG.), DER SOZIALWISSENSCHAFTLICHE FORSCHUNGSPROZESS, (VEB DEUTSCHER VERLAG DER WISSENSCHAFTEN), BERLIN-O 1975, 497-518

POPP, MANFRED, MERKMALE UND ZUSAMMENHAENGE VON ERZIEHUNGSVERHALTEN UND GESAMTVERHALTEN VON LEHRERSTUDENTEN IN DER SELBSEINSCHAETZUNG, IN: PSYCHOLOGIE IN ERZIEHUNG UND UNTERRICHT, 21, 1974, 281-285

PORTELE, GERHARD, ANSAETZE ZU EINER THEORIE INTRINSISCH MOTIVIERTEN LERNENS, (DISS.) MANNHEIM 1972

PRIOR, HARM, KRITISCHE BIBLIOGRAPHIE ZUR HOCHSCHULDIDAKTIK, BLICKPUNKT HOCHSCHULDIDAKTIK HEFT 17, HAMBURG 1971

PUTNAM, HILARY, MIND, LANGUAGE AND REALITY, (CAMBRIDGE UNIVERSITY PRESS), CAMBRIDGE 1975

REISS JR., ALBERT J., SYSTEMATIC OBSERVATION OF NATURAL SOCIAL PHENOMENA, IN: COSTNER, HERBERT L. (ED.), SOCIOLOGICAL METHODOLOGY, (JOSSEY-BASS), SAN FRANCISCO 1971, 3-33

REMMERS, H. H., RATING METHODS IN RESEARCH ON TEACHING, DTSCH. BEARBEITUNG VON: TENT, LOTHAR (1971)

REVENSTORF, DIRK, LEHRBUCH DER FAKTORENANALYSE, (KOHLHAMMER), STUTTGART 1976

RICKER, GUENTER, DIDAKTISCHE THEORIE DER UNTERRICHTSFORSCHUNG, (DISS.) GIESSEN 1976

ROGHMANN, KLAUS, DOGMATISMUS UND AUTORITARISMUS, KRITIK DER THEORETISCHEN ANSAETZE UND ERGEBNISSE DREIER WESTDEUTSCHER UNTERSUCHNUGNEN, (HAIN) MEISENHEIM AM GLAN 1966, 224-241

ROHRMANN, B., EMPIRISCHE STUDIEN ZUR ENTWICKLUNG VON ANTWORTSKALEN FUER DIE SOZIALWISSENSCHAFTLICHE FORSCHUNG, IN: ZEITSCHRIFT FUER SOZIALPSYCHOLOGIE, 1978, 222-245

ROSENSHINE, BARAK, INTERACTION ANALYSIS: A TARDY COMMENT, IN: PHI DELTA KAPPAN, 51, 1970, 445-446

ROSENSHINE, BARAK, TEACHING BEHAVIOURS AND STUDENT ACHIEVEMENT, (NAT. FOUND. FOR ED. RES. IN ENGL. AND WALES) LONDON 1971

ROSENSHINE, BARAK, DIE BEOBACHTUNG DES UNTERRICHTS IN DER KLASSE, IN: HOFER, MANFRED, WEINERT, FRANZ E. (HRSG.), LERNEN UND INSTRUKTION, READER ZUM FUNKKOLLEG PAEDAGOGISCHE PSYCHOLOGIE, BD. 2, (FISCHER) FRANKFURT 1974, 200-217

ROSENSHINE, BARAK, FURST, NORMA, THE USE OF DIRECT OBSERVATION TO STUDY TEACHING, IN: TRAVERS, ROBERT M. W. (ED.), SECOND HANDBOOK OF RESEARCH ON TEACHING, (RAND MCNALLY), CHICAGO 1973, 122-183

ROSENTHAL, ROBERT, EXPERIMENTER EFFECTS IN BEHAVIORAL RESEARCH, (ACC), NEW YORK 1966

RYANS, DAVID G., CHARACTERISTICS OF TEACHERS, (AMERICAN COUNCIL ON EDUCATION) WASHINGTON D.C. 1960

RYANS, DAVID G., SOME RELATIONSHIPS BETWEEN PUPIL BEHAVIOR AND CERTAIN TEACHER CHARACTERISTICS, IN: JOURNAL OF EDUCATIONAL PSYCHOLOGY, 52, 1961, 82-90

SALZMANN, CHRISTIAN GOTTHILF, KREBSBUECHLEIN, HRSG. V. M. HARTMANN, (JAEGER), LEIPZIG O.J. (1780)

SALZMANN, CHRISTIAN GOTTHILF, AMEISENBUECHLEIN, HRSG. V. TH. FRITZSCH, (FRIEDRICH BRANDSTETTER) LEIPZIG 1915 (1806)

SCHERER, KLAUS R., BEOBACHTUNGSVERFAHREN ZUR MIKROANALYSE NON-VERBALER VERHALTENSWEISEN, IN: KOOLWIJK, JUERGEN VAN, WIEKEN-MAYSER, MARIA, TECHNIKEN DER EMPIRISCHEN SOZIALFORSCHUNG, BD. 3, (OLDENBOURG), MUENCHEN 1974, 66-109

SCHMIDTKE, ARMIN, GROFFMANN, KARL JOSEF, SCHALLER, SYLVIA, SELBST- UND IDEALBILDER VON STUDENTEN DES ERSTEN UND ZWEITEN BILDUNGSWEGES UND IHRE VERMUTETE UND REALE BEURTEILUNG DURCH HOCHSCHULLEHRER, IN: PSYCHOLOGIE IN ERZIEHUNG UND UNTERRICHT, 25, 1978, 90-100

SCHMITZ, REINHARD, UEBERSICHT UEBER AUSGEWAEHLTE MESSINSTRUMENTE ZUR HOCHSCHULSOZIALISATIONSFORSCHUNG, D.I.P.- METHODOLOGISCHE ARBEITSBERICHTE NR. 1, MUENSTER 1975

SCHNEIDER, H., STICHWORT "ESSENTIALISMUS", IN: RITTER, JOACHIM (HRSG.), HISTORISCHES WOERTERBUCH DER PHILOSOPHIE, BD.2, (WISS. BUCHGES.) DARMSTADT 1972, SP.751-753

SCHNELLE, HELMUT, SPRACHPHILOSOPHIE UND LINGUSISTIK, (ROWOHLT) REINBEK 1973

SCHOMAKER, HELLA UND LEHMANN, ACHIM, DATEN ZUR STUDIENSITUATION UND MOTIVATION VON STUDENTEN, SCHRIFTENREIHE DER HOCHSCHULE FUER WIRTSCHAFT, BAND 2, BREMEN 1976

SCHOTT, ERICH, ZUR EMPIRISCHEN UND THEORETISCHEN GRUNDLEGUNG EINES BEWERTUNGSINSTRUMENTES FUER VORLESEUNGEN, BLICKPUNKT HOCHSCHULDIDAKTIK HEFT 28, 1973

SCHULZ, WOLFGANG, TESCHNER, WOLFGANG P., VOIGT, JUTTA, VERHALTEN IM UNTERRICHT, SEINE ERFASSUNG DURCH BEOBACHTUNGSVERFAHREN, IN: INGENKAMP, KARLHEINZ UND PAREY, EVELORE (HRSG.), HANDBUCH DER UNTERRICHTSFORSCHUNG, TEIL I, 2. AUFL., (BELTZ) WEINHEIM 1971, SP. 633 - 851

SCHULZ, WOLFGANG U.A., BEOBACHTUNG UND ANALYSE VON UNTERRICHT, (BELTZ) WEINHEIM 1973

SEARLE, JOHN R., SPEECH ACTS, (CAMBRIDGE UNIVERSITY PRESS), LONDON 1969

SEITZ, WILLI, PERSOENLICHKEITSBEURTEILUNG DURCH FRAGEBOGEN, (WESTERMANN) BRAUNSCHWEIG 1977

SIEBERT, HORST, ANSAETZE UND ERGEBNISSE DER UNTERRICHTSFORSCHUNG IN DER ERWACHSENENBILDUNG, IN: ZEITSCHRIFT FUER PAEDAGOGIK, 23, 1977, 663 - 679

SIMONS, HERIBERT, FORSCHUNG IM BEREICH DES COLLEGE UND DER UNIVERSITAET, IN: INGENKAMP, KARLHEINZ UND PAREY, EVELORE (HRSG.), HANDBUCH DER UNTERRICHTSFORSCHUNG, TEIL III, (BELTZ) WEINHEIM 1970, SP. 3341 - 3484

SIMPSON, RAY H., THE CASE OF SELF-EVALUATION PROCEDURES BY LECTURES IN EDUCATIONAL PSYCHOLOGY, IN: EDUCATIONAL REVIEW, 18, 1965, 25 - 33

SINGLETON, W. T., TECHNIQUES FOR DETERMINING THE CAUSES OF ERROR, IN: APPLIED ERGONOMICS, 3, 1972, 126-131

SIX, BERND, DIE RELATION VON EINSTELLUNG UND VERHALTEN, IN: ZEITSCHRIFT FUER SOZIALPSYCHOLOGIE, 6, 1975, 270-296

SMITH, JUDITH M., A NEW RATER SELECTION TECHNIQUE FOR USE WITH BEHAVIORAL RATING SCALES, IN: JOURNAL OF CLINICAL PSYCHOLOGY, 30, 1974, 40-43

SMITS, GERARD J., A FORTRAN IV PROGRAM TO COMPUTE RELIABILITY OF RATINGS FOR TWO OR MORE JUDGES, IN: BEHAVIOR RESEARCH METHODS & INSTRUMENTATION, 10, 1978, 864

SODEUR, WOLFGANG, EMPIRISCHE VERFAHREN ZUR KLASSIFIKATION, (TEUBNER STUDIENSKRIPTEN) STUTTGART 1974

SOMMER, ROBERT, THE DISTANCE FOR COMFORTABLE CONVERSATION; A FURTHER STUDY, IN; SOCIOMETRY, 25, 1962, 111 - 116

STEGEMANN, GUDRUN, PH-STUDENTEN MIT UND OHNE ABITUR - EIN VERGLEICH AUSGEWAEHLTER EINSTELLUNGSDIMENSIONEN, (UNVEROEFFENTL. ZULASSUNGSARBEIT) PH HEIDELBERG 1975

STEGMUELLER, WOLFGANG, DAS WAHRHEITSPROBLEM UND DIE IDEE DER SEMANTIK, 2. AUFL., (SPRINGER) WIEN 1977

STEINZOR, BERNARD, THE DEVELOPMENT AND EVALUATION OF A MEASURE OF SOCIAL INTERACTION, IN; HUMAN RELATIONS, 2, 1949, 103 - 121

STEVENS, STANLEY S., MATHEMATICS, MEASUREMENT AND PSYCHOPHYSICS, IN; DERS. (ED.), HANDBOOK OF EXPERIMENTAL PSYCHOLOGY, (WILEY) NEW YORK 1951, 1-49

SUMASKI, WERNER, DIMENSIONEN DES LEHRERVERHALTENS UND IHRE PHAENOMENALE REPRAESENTIERTHEIT, (DISS.) PH NIEDERSACHSEN ABT. HILDESHEIM, 1977

TAUSCH, REINHARD, TAUSCH, ANNE-MARIE, ERZIEHUNGSPSYCHOLOGIE, 5. AUFL., (HOGREFE), GOETTINGEN 1970

TAUSCH, REINHARD UND TAUSCH, ANNE-MARIE, ERZIEHUNGSPSYCHOLOGIE, 7. AUFL. (HOGREFE) GOETTINGEN 1973

TENT, LOTHAR, UNTERSUCHUNGEN ZUR ERFASSUNG DES VERHAELTNISSES VON ANPASSUNG UND LEISTUNG BEI VORWIEGEND PSYCHISCH BEANSRUCHENDEN TAETIGKEITEN, IN; ARCHIV FUER DIE GESAMTE PSYCHOLOGIE, 115, 1963, 105 - 172

TENT, LOTHAR, SCHAETZVERFAHREN IN DER UNTERRICHTSFORSCHUNG, IN; INGENKAMP, KARLHEINZ UND PAREY, EVELORE (HRSG.), HANDBUCH DER UNTERRICHTSFORSCHUNG, TEIL 1, 2. AUFL., WEINHEIM 1971, SP. 853-1.000

TERHART, EWALD, DIE LOGIK DES LEHRENS, IN; BILDUNG UND ERZIEHUNG, 30, 1977, S. 441 - 456

TRAVERS, ROBERT M.W. (ED.), SECOND HANDBOOK ON TEACHING (RAND MC NALLY) CHICAGO 1973

TSCHERNER, KLAUS, MASENDORF, FRIERICH, ANALYSE VON SCHUELERBEURTEILUNGEN UND ZEUGNISNOTEN BEI EINZELNEN LEHRERN, IN; PSYCHOLOGIE IN ERZIEHUNG UND UNTERRICHT,

21, 1974, 135-149

UEBERLA, KARL, FAKTORENANALYSE, 2. AUFL., (SPRINGER) BERLIN 1971

WAGNER, HANS, STICHWORT "BEGRIFF", IN: KRINGS, HERRMANN, BAUMGARTNER, HANS MICHAEL, WILD, CHRISTOPH (HRSG.), HANDBUCH PHILOSOPHISCHER GRUNDBEGRIFFE, BD. 1, (KOESEL) MUENCHEN 1973, 191-209

WALBERG, HERBERT J., MODELS FOR OPTIMIZING AND INDIVIDUALIZING SCHOOL LEARNING, IN: INTERCHANGE, 2, 1971, 15 - 17

WALTER, HELLMUTH, EINFUEHRUNG IN DIE UNTERRICHTSFORSCHUNG, (WISS. BUCHGES.) DARMSTADT 1977

WALTER, HELLMUTH, EXPERIMENTELLE UNTERRICHTSFORSCHUNG - EINE ESOTERISCHE SEZESSIONSWISSENSCHAFT? EINE ERWIDERUNG AUF WERNER S. NICKLIS, IN: ZEITSCHRIFT FUER PAEDAGOGIK, 25, 1979, 307-312

WATSON, JEANNE, POTTER, ROBERT J., AN ANALYTIC UNIT FOR THE STUDY OF INTERACTION, IN: HUMAN RELATIONS, 15, 1962, 245-263

WATSON, JOHN B., BEHAVIORISMUS (HRSG. V. C.F. GRAUMANN), (KIEPENHEUR UND WITSCH) BERLIN 1968

WEBB, EUGENE J. ET AL., NICHTREAKTIVE MESSVERFAHREN, (BELTZ), WEINHEIM 1975

WEICK, KARL E., SYSTEMATIC OBSERVATIONAL METHODS, IN: LINDZEY, GARDNER, ARONSON, ELLIOT, THE HANDBOOK OF SOCIAL PSYCHOLOGY, VOL. TWO, 2ND ED., (ADDISON-WELEY) READING, MASS. 1968, 357-451

WIECHARDT, DOERTE, ZUR ERFASSUNG DES SELBSTKONZEPTS, IN: PSYCHOLOGISCHE RUNDSCHAU, 28, 1977, 294 - 304

WIENS, A. ET AL., CAN INTERVIEWER INTERACTION MEASURES BE TAKEN FROM TAPE RECORDINGS?, IN: THE JOURNAL OF PSYCHOLOGY, 63, 1966

WILCKE, BERND-ACHIM, STUDIENMOTIVATION UND STUDIENVERHALTEN, (HOGREFE) GOETTINGEN 1976

WIRTH, SIEGFIRED, DIE EINSTELLUNG DER STUDENTEN AN DER WIRTSCHAFTSUNIVERSITAET WIEN ZUR BETRIEBSWIRTSCHAFTLICHEN AUSBILDUNG, (UNVEROEFFENTL. DIPL.-ARBEIT) WIRTSCHAFTSUNIVERSITAET WIEN 1977

WITTE, WILHELM, DAS PROBLEM DER BEZUGSSYSTEME, IN: METZGER, WOLFGANG (HRSG.), BAND 1 DES HANDBUCHS DER PSYCHOLOGIE, ALLGMEINE PSYCHOLOGIE , I. DER AUFBAU DES ERKENNENS, 1. HALBBAND: WAHRNEHMUNG UND BEWUSSTSEIN, (HOGREFE) GOETTINGEN 1966, 1003-1027

ZABECK, JUERGEN, PARADIGMAPLURALISMUS ALS WISSENSCHAFTSTHEORETISCHES PROGRAMM, IN: BRAND, WILLI, BRINKMANN, DOERTE (HRSG.), TRADITION UND NEUORIENTIERUNG IN DER BERUFS- UND WIRTSCHAFTSPAEDAGOGIK (FESTSCHRIFT FUER LUDWIG KIEHN), (WIECHELMANN) HAMBURG 1978, 291-332

ZECHA, GERHARD, WIE LAUTET DAS "PRINZIP DER WERTFREIHEIT"?, IN: ZEITSCHRIFT FUER SOZIOLOGIE UND SOZIALPSYCHOLOGIE, 28, 1976, 609-648