Detlef Sembill

7. Übereinstimmung zwischen Beobachtern – Reliabilitätsprobleme qualitativer Unterrichtsanalysen

I. Einleitung

Übereinstimmungskoeffizienten zwischen Beobachtern werden ermittelt, um über die Zuverlässigkeit der vorgenommenen Kodierungen – in unserem Fall bei Unterrichtsbeobachtungen – Aussagen machen zu können. Begrifflich orientieren sich daher die für diese Maße verwendeten Bezeichnungen in einer über 30jährigen Tradition an der klassischen Testtheorie (an Reliabilitätsmaßen).

Die Möglichkeit, auf der Basis identischer Rohdaten (Beobachterkodierungen) einer eigenen Untersuchung (SEMBILL / WESELOH 1978) mittels dreier unterschiedlich aufgebauter Koeffizienten je nach Belieben eine nicht vorhandene, eine ausreichende oder eine vertrauenerweckende Übereinstimmung zu „erzielen", wobei Konfidenzintervalle gebildet werden konnten, deren Grenzen bei 5 % Irrtumswahrscheinlichkeit insgesamt zwischen – 0,34 und + 0,81 lagen, waren Anlaß, sowohl die der Ermittlung zugrunde liegende statistische Vorgehensweise zu prüfen als auch die Brauchbarkeit der Koeffizienten unter methodologischen Aspekten zu untersuchen.

Die zur Bezeichnung der Koeffizienten verwendeten Begriffe sind nicht einheitlich (vgl. zum Beispiel MERRILL 1946: observer reliability; GUILFORD 1959: interrater consistency; MANDL / HUBER 1979: Beobachterobjektivität; u. a. m.); gleiche statistische Maße verbergen sich auch hinter unterschiedlichen Beziehungen. Der eigentümliche Umgang mit diesen Übereinstimmungsmaßen in der Literatur ist oft durch mangelhafte Information gekennzeichnet bezüglich des Verfahrens der Beobachter / Kodierer, der Stichproben, der Zeiteinheiten, der Intentionen, der Operationalisierungen, der Koeffizientenwahl usw. (vgl. zum Beispiel SCHULZ / TESCHNER / VOIGT 1970; Sp. 759; KERLINGER 1979, 808). Die so relativ isoliert vermittelten Koeffizienten – meist weit größer als .80 – sind schon allein aufgrund ihrer überragenden Höhe

geeignet zu suggerieren, daß vorgeordnete qualitative Probleme zufriedenstellend gelöst seien.
Fehlen Angaben zum Übereinstimmungsproblem überhaupt, wie zum Beispiel in den durchaus wichtigen Berichten von KEIL / PIONTKOWSKI (1973) oder von BROPHY / GOOD (1976), sind unterschiedliche Interpretationen möglich:

a) dererlei Überprüfungen wurden nicht durchgeführt;
b) die ermittelten Koeffizienten waren nicht hoch genug oder
c) diese Koeffizienten erschienen den Verfassern als Zuverlässigkeitsmaße nicht brauchbar.

Bevor wir die statistische Konzeption dieser Koeffizienten diskutieren, soll die das Übereinstimmungsproblem hervorrufende Methode „Unterrichtsbeobachtung" sowohl unter methodologischen Gesichtspunkten als auch auf konkrete Verfahrensweisen hin untersucht werden.

II. Unterrichtsbeobachtung als methodologisches Problem

Die zur Überprüfung von Theorie verwendete Methode Unterrichtsbeobachtung verdeutlicht auf ganz spezifische Weise den schwierigen Übergang von der Beobachtungs- auf die Theorieebene. HEMPEL (1952, 36) hat die Theorie mit einem Netzwerk verglichen:

„Ihre Begriffe sind durch Knoten repräsentiert, wobei die sie verbindenden Fäden zum Teil den Definitionen und zum Teil den fundamentalen und den abgeleiteten Hypothesen, die die Theorie enthält, entsprechen. Das ganze System schwebt über der Ebene der Beobachtung und ist in ihr verankert durch Regeln der Interpretation. Diese lassen sich als Bänder bezeichnen, sie sind keine Bestandteile des Netzwerkes, sondern verbinden bestimmte Punkte hiervon mit bestimmten Stellen der Beobachtungsebene. Mit Hilfe dieser interpretativen Verbindungen kann das Netzwerk als wissenschaftliche Theorie funktionieren. Von gewissen Beobachtungsdaten kann man über ein Band der Interpretation zu einem Punkt des theoretischen Netzwerkes aufsteigen, von dort aus durch Definitionen und Hypothesen zu anderen Punkten gelangen, von denen aus ein anderes Interpretations-Band einen Abstieg auf die Ebene der Beobachtung gestattet" (zitiert nach FRIEDRICHS 1973, 62).

In dieser – recht plastisch dargestellten – Theorieauffassung kann die Gesamtsprache (L) in drei Bereiche gemäß ihren unterschiedlichen Aussagen gegliedert werden (vgl. GROEBEN / WESTMEYER 1975, 62 ff.):

(1) L_B: *Beobachtungssprache*, deren undefinierte nicht-logische Grundbegriffe sich nur auf Beobachtbares beziehen (zum Beispiel Beschreibung der Aktivitäten von Interpretationspartnern);

(2) L_T: *Theoretische Sprache*, die – den Knoten entsprechend – neben logischen Zeichen nur theoretische Begriffe enthält (zum Beispiel Interaktionsmuster, -typen);

(3) Z: *System der Zuordnungsregeln*, die sowohl theoretische Terme als auch Beobachtungsbegriffe enthalten und somit die Verbindung (entsprechend den Interpretationsfäden) der theoretischen Sprache mit der Beobachtungssprache stiften (zum Beispiel Einordnung in bestimmte Interaktionskategorien).

Es scheint ökonomisch und unmittelbar plausibel zugleich zu sein, eine Theorie unter einer möglichst konkreten Zielvorstellung bzw. Fragestellung zu entwickeln.

Wir gehen deshalb bei den weiteren Erörterungen von folgenden Überlegungen aus: Eine von uns langfristig angestrebte Konstruktion wissenschaftlich begründeter Handlungsempfehlungen *für* Unterricht erfordert handlungsrelevantes Wissen *über* Unterricht. Dazu scheint es sinnvoll zu sein – neben der Erhebung von Fragebogen-, Test- oder auch physiologischen Daten –, Unterrichtsgeschehen in natürlichen Situationen über längere Zeit und unter Einbezug von Inhalt, Kontext und Situationsspezifität zu beobachten und zu dokumentieren. Wir suchen also Regelmäßigkeiten in den Charakteristika von Interaktionen im Unterrichtsablauf, unter ausdrücklicher Betonung der Tatsache, daß Unterricht eine intentional auf Verhaltensänderung (Lernziele!) ausgerichtete Veranstaltung ist.

Die Regelmäßigkeiten sind durch den Übergang auf die Theorieebene allerdings nicht mehr „realer", sondern logischer Art (vgl. ULICH 1978, 298 f.), indem man gesetzmäßige Beziehungen zu formulieren versucht, die sich in der Regel an dem in seiner idealen Variante einer deduktivistischen Wissenschaftsauffassung folgenden HEMPEL-OPPENHEIM-Schema orientieren, nach dem gegebene Gesetzesaussagen Bestätigungsversuchen ausgesetzt werden. Da entsprechende Gesetzesaussagen in der Unterrichtswissenschaft aber nicht oder kaum gegeben sind, führt der vielfach gewählte Weg im Sinne eines induktiven Schlußverfahrens mit nachträglicher „Gesetzesbildung" (vgl. ACHTENHAGEN 1979, 271) über die Akkumulation von isolierten und nur selten integrierbaren Einzelbefunden allenfalls zu empirischen Generalisierungen, d. h. zu raumzeitlichen Extrapolationen rein deskriptiv ermittelter, zeitpunktbezogener Zusammenhänge (vgl. GROEBEN / WESTMEYER 1975, 59 und 131).

Die oben angeführten gesetzmäßigen Beziehungen werden zwischen als invariant angenommenen Strukturen und den durch sie bedingten Regelmäßigkeiten formuliert und dabei lediglich als solche konstatiert und nicht bezüglich ihres Zustandekommens aufzuklären versucht (vgl. ULICH 1976, 31). Sie werden in der Regel weder bezüglich der Probleme möglicher Konfundierungen von Gesetzesaussagen mit Ante-

cedensbedingungen reflektiert (vgl. GROEBEN / WESTMEYER 1975, 93; ACHTENHAGEN 1979, 273), noch werden Angaben darüber gemacht, wie die notwendigerweise zu erfüllenden Bedingungen zur Anwendung der Gesetze und damit deren Geltungsbereiche beschaffen sind (vgl. WESTMEYER 1979, 149 f.). Die fehlende Kenntnis über das Zustandekommen der Regelmäßigkeiten ermöglicht es uns nicht, zwischen alternativen Erklärungsketten bezüglich eines Explanandums zu entscheiden, und läßt somit beliebig viele Rekonstruktionsversuche zu (vgl. GROEBEN / WESTMEYER 1975, 96 f.).

Gerade unter unserer Zielsetzung scheint es dringend notwendig zu sein, den Generalisierungsbereich explizit raumzeitlich zu begrenzen und Unterstellungen stabiler Strukturen sowie idealisierende (praxisferne) Bedingungen zu vermeiden (vgl. HAUSSER / KRAPP 1979, 72; WESTMEYER 1979, 150).

Im Rückgriff auf die vorgenommene Aufteilung der Gesamtsprache (L) einer Theorie wollen wir kurz das Transformationsproblem der Beobachtungsbegriffe in quantitative Begriffe der Theoriesprache aufgreifen: Dieses Problem scheint nur zu lösen zu sein, wenn nicht nur die Beobachtungssprache der Theoriesprache über die Zuordnungsregeln eine Bedeutung verleiht, sondern *auch umgekehrt* die Bedeutung von Beobachtungsbegriffen durch die theoretischen Begriffe via Zuordnungsregeln mitbedingt wird (vgl. GROEBEN / WESTMEYER 1975, 64). Die bezeichnete Wechselwirkung verdeutlicht folgendes:

1. Die in den Zuordnungsregeln enthaltenen Kategorien, Kriterien und Methoden (vgl. auch „Überprüfungsverfahren" bei ACHTENHAGEN 1979, 274 f.) befinden sich in einem engen Zusammenhang mit den dahinter stehenden Theorien (vgl. auch HAUSSER / KRAPP 1979, 73), sind also gewissermaßen abhängig von den Problemlösungskonzeptionen der Forscher (vgl. HERRMANN 1979, 30 ff.).
2. Aus der Menge der in der Beobachtungssprache formulierten, generell verfügbaren Rohdaten wird so ein bestimmter und begrenzter Merkmalsraum abgehoben, der von der einzelwissenschaftlichen Methodik (durch das Forscherindividuum) geprägt wird (vgl. GROEBEN / WESTMEYER 1975, 25).
3. Durch die Methode ‚Unterrichtsbeobachtung' wird auf diese Weise Unterrichtsrealität konstruiert. Es ist also zu überprüfen, ob diese Methode – im Sinne des geforderten handlungsrelevanten Wissens über Unterricht – in der Lage ist, eine adäquate Teilmenge von Rohdaten zu selegieren und sie äquivalent in logische Begriffe zu transformieren.

III. Beobachtungsverfahren und Beurteilungsproblematik

Es war Ziel der bisherigen Ausführungen (a) aufzuzeigen, daß die Theorie auf die Zuordnungsregeln und damit auf die beobachtbaren Phänomene Einfluß ausübt, und (b) nahezulegen, diesen Einfluß aufgrund der angedeuteten erheblichen Mängel zu reduzieren, mindestens jedoch zu relati-

vieren zugunsten eines größeren, kontextbezogeneren Einflusses des Untersuchungsgegenstandes selbst; im übrigen eine Tendenz, die in allen Teildisziplinen der Sozialwissenschaften zu beobachten ist (vgl. zum Beispiel BÖHME / v. ENGELHARDT 1979; LEMPERT 1979; WESTMEYER 1979; WAHL u. a. 1977; OERTER 1979; SCHANZ 1977).
Wenn wir davon ausgehen, daß im Sinne des oben geforderten handlungsrelevanten Wissens über Unterricht das Unterrichtsgeschehen von Lehrern und Schülern intentional gesteuert wird mit der Absicht der Verhaltensänderung und daher auch variabel sowie situationsspezifisch ist, so ist dieses kontextbezogene, unter Einfluß des Inhaltsaspektes (vgl. ACHTENHAGEN / WIENOLD 1975; HUBER / MANDL in diesem Band) im Zeitablauf zu betrachtende Zusammenspiel zu beschreiben und zu ergründen. Dies wäre Aufgabe einer „educational evaluation", die sich bewußt von einem „educational testing" distanziert (vgl. GLASS 1970, 60 f.). Sie wäre damit gleichzeitig die Bedingung für die notwendige Kontingenzenbildung zwischen anthropogenen / soziokulturellen Voraussetzungen, Unterrichtsaktivitäten und schulischem Leistungsverhalten (vgl. STAKE 1972, 93 f.).
Es scheint fraglich zu sein, ob die oben dargestellten Zielsetzungen mit quantitativen Beobachtungsverfahren allein zu erreichen sind. Bei diesen handelt es sich um vorher entwickelte, geschlossene Erhebungsinstrumente, mit deren Hilfe wahrgenommene „Realität" unmittelbar in sehr spezifische, oft rigide, manchmal sogar äußerst triviale Kategorien „zerteilt" wird (vgl. KERLINGER 1979, 794). Diese erfordern möglicherweise nur noch die zum Beispiel von MEDLEY / MITZEL ausschließlich zugelassene Beantwortung der Frage, ob ein bestimmtes Verhalten aufgetreten ist oder nicht (vgl. MEDLEY / MITZEL 1963 n. KERLINGER 1979, 786, Fußnote 5). Sie eröffnen so (paradoxerweise), von störenden Nebeneffekten der „Alltagsrealität" befreit, die Möglichkeit, kausale Verknüpfungen (in der „Alltagsrealität") nachzuweisen (vgl. auch: SCHÖN 1979, 19 ff.).
Geforderte Prozeßanalysen (vgl. zum Beispiel BROPHY / GOOD 1976; HAUSSER / KRAPP 1979, 74 f.) verlangen darüber hinaus ein flexibles Verfahren, in dem die Erhebungssituation (relativ) offen ist (zum Beispiel offene Interviews, Tonband- und/oder Videobandaufzeichnungen von Unterricht). Die eigentliche Arbeit fällt bei Verwendung eines solchen qualitativen, mittelbaren Beobachtungsinstruments mit dem nachträglichen Kodieren der erhaltenen Rohdaten an. Das bedeutet zwar zunächst nur einen punktuellen Unterschied, da das vorliegende Datenmaterial früher oder später doch quantifiziert werden muß; die Chancen aber, daß der Kontext möglichst lange und möglichst komplex auf dem Weg zur Theorieebene (zum Forscherindividuum) erhalten bleibt, daß unter Umständen die Untersuchungsobjekte (hier Lehrer und

Schüler), die ja in der Unterrichtswissenschaft ebenso „Subjekt" sind wie der Forschende, einen wesentlichen Beitrag zum besseren Verständnis und damit zur besseren Interpretation des Datenmaterials leisten können, sollten gewahrt werden. Auch unter dem Aspekt der „Rückübersetzung" von Theorie in Praxis scheint das arbeitsaufwendigere Verfahren gerechtfertigt.

Reflektieren wir unter unserer Zielsetzung zum Beispiel den Forschungsbericht von SHAVELSON / DEMPSEY-ATWOOD (1976, vgl. insbesondere 553 f.): Dort wurden als mögliche Gründe für die mangelnde Stabilität von Forschungsergebnissen im Sinne angestrebter Generalisierung die systematischen und/oder zufälligen Differenzen von Lehrern und Schülern im Zeitablauf bzw. in unterschiedlichen Situationen (das entspräche in der klassischen Teststatistik der spezifischen Varianz) sowie die verwendeten Meßverfahren genannt. Dieser Trend ist in unserem Verständnis eine Bestätigung für die am symbolischen Interaktionismus und an der Ethnomethodologie orientierte Annahme, daß nicht die Fakten und Daten selbst handlungsrelevant sind, sondern im Sinne der qualitativen Analyse vielmehr deren Bedeutungen (vgl. SCHÖN 1979, 20 f.). Konsequenterweise rückt somit der Bereich, in dem die vermeintliche Fehlervarianz produziert wird, näher in unser forschungsleitendes Interesse.

Wir haben in einem bestimmten Verständnis von Evaluation die Funktion der Zuordnungsregeln aufzuzeigen versucht, die darin besteht, theoriegeleitet Äquivalenzen zwischen den Begriffen der Beobachtungs- und der Theoriesprache herzustellen. Durch die Methode ‚Unterrichtsbeobachtung' wird ein bewußter Wahrnehmungsvorgang in Gang gesetzt, gesteuert durch die vereinbarten Kodierregeln, nach denen die Beschreibungen der Aktivitäten als Bestandteil der Beobachtungssprache im anschließenden Kodiervorgang auch in die entsprechenden Kategorien eingeordnet werden; diese Methode wird von uns in diesem Sinne den Zuordnungsregeln zugerechnet.

Um Mißverständnissen vorzubeugen, sei explizit darauf hingewiesen, daß trotz aller Bemühungen, Unterrichtswissenschaft wertfrei zu betreiben, die Zuordnungsregeln (mehr oder weniger offensichtlich, jedoch unvermeidbar) ein „Bewertungsprogramm" darstellen. Dabei müssen die Wertungen nicht unbedingt auf der Ebene der Zuordnungsregeln direkt vorgenommen werden, sie können auch vor- und/oder nachgelagert sein, zentrieren sich aber dort aufgrund ihrer „Überbrückungsfunktion". So werden Wertungen einmal semantisch fortgeschrieben, wie das auch schon in Begriffen wie zum Beispiel Evaluation, Äquivalenz und Bedeutungszuschreibung zum Ausdruck kommt. Diese Wertungen verschwinden auch nicht durch die (dann nicht mehr hinterfragte) Exterritorialisierung via Experten auf die Metaebene bzw. durch die Überantwortung

an die normative Kraft des (gesellschaftlich-historisch) Faktischen (vgl. GROEBEN 1979, 57 ff.).
Zum anderen werden Wertungen durch Selektionen vorgenommen, die, beginnend mit der Auswahl der Probleme (hier als Ist-Soll-Diskrepanzen verstanden) über die Problemlösungskonzeptionen (s. o.) bis zu den Problemlösungen (Erhebung, Auswertung, Interpretation, Veröffentlichung), unvermeidbar durch implizite Menschenbildannahmen geprägt sind, indem immer Sollvorstellungen, d. h. Ziele der Subjektkonstituierung, mitgeführt werden (vgl. GROEBEN 1979, 55 f.). Aufgrund dieser Überlegungen müssen wir uns fragen, ob es sinnvoll ist, Beobachtungs*systeme* von Beurteilungs*verfahren* nach dem Kriterium der Inferenz abzugrenzen (vgl. SCHWARZER 1975, 754 f.).

IV. Funktion der Beobachter

Akzeptiert man die unvermeidbare und auch notwendige Wertung im Forschungsprozeß (s. auch ‚Vakuumthese' bei GROEBEN 1979, 51 ff.), so ist, wie im letzten Argument schon angedeutet, die Frage nach den beteiligten Personen und ihren Funktionen zu klären.
Die Methode ‚Unterrichtsbeobachtung' erfordert in unmittelbarer oder mittelbarer Form einen oder mehrere Beobachter (mittelbar: zum Beispiel Auswertung von Videobändern), d. h. daß wir es nicht nur auf der Theorie- und auf der Praxisebene mit reflexionsfähigen Individuen zu tun haben, sondern auch im Bereich der Methodenanwendung. Inwiefern dieser Sachverhalt ein Manko ist, oder vielleicht auch Chancen eröffnen kann, muß wohl im Zusammenhang mit Zielsetzungen, Beobachtungsverfahren und Kontrollverfahren diskutiert werden. Wir gehen dabei davon aus, daß Forscher, Beobachter und Untersuchte jeweils unterschiedliche Personen sind. Überlegungen sollen einmal
(a) bezüglich quantitativer Beobachtungsverfahren, zum anderen
(b) bezüglich qualitativer Beobachtungsverfahren (s. o.)
angestellt werden.

Zu (a): Die Beobachter ersetzen hier in ihrer gleichzeitigen Funktion als Kodierer quasi einen Test. Sie sind weder in die Hypothesen noch in die Ziele der Untersuchung eingeweiht und bekommen ein Kategoriensystem mit festen Operationalisierungen (möglichst unter Verzicht von Schätzskalen) vorgegeben. Dieses System ist oft überwiegend unter Reliabilitätsgesichtspunkten konstruiert (zum Beispiel von MEDLEY / MITZEL, vgl. MERKENS / SEILER 1978, 114). Unter diesem Gesichtspunkt wird in der Regel auch das Beobachtungstraining nach Experten-Vorbild durchgeführt (vgl. FIEGUTH 1977, 79 f.). Beobachter, die dabei zu häufig von der Forscher-Norm abweichen, werden „wegen Dummheit oder Bosheit von der weiteren Betrachtung ausgeschlossen" (KRIZ 1978, 90).
Obwohl zum Teil pädagogisch/psychologische Vorkenntnisse bei den Beobachtern verlangt werden, sollen sie nicht wissen, was ein „gutes" Ergebnis oder Verhalten ist

(vgl. FIEGUTH 1977, 80). Die so zur A-Reflexivität Veranlaßten, können unter Umständen in schwierige Entscheidungskonflikte zwischen ihren eigenen Erfahrungen und dem Kategoriensystem bezüglich des aufgetretenen Verhaltens geraten (vgl. auch: WAGNER et al. 1977, 244 ff.). So bleibt – die ungeprüfte – Frage, ob die dem Beobachter vermeintlich entzogenen Wertungen („Exterritorialisierung") nicht doch wieder Eingang in die Zuordnungen finden. Es ist zu bezweifeln, ob die so „funktionierenden" Beobachter bei einer derartigen Zerstückelung des Kontextes auch nur annähernd ihre eigentliche Aufgabe unter unserer Zielsetzung zu leisten vermögen: die äquivalente Umsetzung „realer" in logische Begriffe.
Zu (b): Wegen der ernstzunehmenden Befürchtung, „if observers are brainwashed to the point that they consistently code the same ambiguous behaviors into a certain category, results could be biased" (vgl. FRICK / SEMMEL 1978, 162, in Anlehnung an MEDLEY / NORTON 1971), werden die Beobachter hier als reflexive Individuen betrachtet und eingesetzt. Es wird dabei davon ausgegangen, daß das relativ begrenzte Spektrum der Forscherindividuen über Unterrichtsphänomene durchaus erweitert werden kann. Das ist zum einen möglich durch die Berücksichtigung weiterer, unterschiedlicher Relevanzstrukturen bezüglich der Konstitution lebensweltlicher Situationen und zum anderen dadurch, daß unterschiedliche Beziehungen, hergestellt zwischen individuellen Lernerfahrungen (Wissensvorräten) und aktuellen Erfahrungen bzw. Handlungen (vgl. auch: SCHÜTZ / LUCKMANN 1979), zugelassen werden. Die Beobachter können dadurch – sowohl bei der Instrumentenentwicklung als auch beim Beobachtungs- / Kodiervorgang selbst – als korrigierendes Relativ zum Forscher tätig werden. Dazu müssen sie die Intention (nicht unbedingt einzelne Hypothesen) der Untersuchung kennen (vgl. dazu auch: MEES 1977, 70) sowie den Lerninhalt der zu beobachtenden Unterrichtseinheit. Zur besseren Kontextbeurteilung (zum Beispiel Interaktionsklima, lehrer- oder schülerindividuelle Sichtweisen) scheint ein kombiniertes unmittelbares/mittelbares Beobachtungsverfahren geeignet. Dabei ist vorauszusetzen, daß mindestens ein Teil der Beobachter den Unterricht „live" ohne Kodierzwang erlebt hat, um dann bei den Kodierungen auf der Grundlage von Videoaufzeichnungen, gerade auch für den Beginn der Unterrichtseinheit schon die Relationen zwischen den Interaktionspartnern zu kennen. In diesem Sinne sind Schlußfolgerungen explizit gewollt.
Es scheint evident, daß (a) zugelassene Wertungen auf der Ebene der Zuordnungsregeln, die zwischen den Beobachtern eines Teams diskutiert und entschieden werden, die oben angeführten Konflikte mit den eigenen Lernerfahrungen etc. erheblich reduzieren und (b) daß die Chancen der notwendigen Äquivalenzbildung (s. o.) durch die Nutzbarmachung der menschlichen Reflexions*fähigkeit* steigen.

V. Reliabilitätsprobleme

Das Gütekriterium Reliabilität (Zuverlässigkeit) als „notwendige, wenn auch nicht hinreichende Bedingung der Gültigkeit" (ANGER 1969, 607) ist ein fundamentaler Begriff der klassischen Testtheorie. Zweck von Reliabilitätsprüfungen ist, durch Angaben bezüglich der Voraussetzungen, der Durchführung, der aufgetretenen Probleme und der erzielten Ergebnisse weitgehende Nachvollziehbarkeit und Vergleichbarkeit, letztlich also Transparenz herzustellen (vgl. LIENERT 1969, 214 ff.). Für KERLINGER (1979) ist die Angabe der Reliabilität eine Voraussetzung, um den Ergebnissen vertrauen zu können (vgl. 683).

Der „wahre Wert" und der „Meßfehler" sind die Grundannahmen der klassischen Testtheorie, die, lediglich an der „Meßgenauigkeit" interessiert, die Definition und Existenz der zu messenden Dimension einfach voraussetzt (vgl. FISCHER 1974, 20). Aber gerade bei der Anwendung auf Verhaltensniveau bedürfen die für „nonliving systems" konstruierten Maße (vgl. FRICK / SEMMEL 1978, 179) einer Theorie des Messens.

„Die Erkenntnis eines allgemeinen Gesetzes muß der Messung vorausgehen. Wird diese Forderung nicht erfüllt, dann handelt es sich nicht um Messung, sondern nur um Schein-Quantifizierung, um Benennung von Ereignissen mit Zahlen anstelle beliebiger anderer Namen" (FISCHER 1974, 23).

Die Zuordnung von Zahlen ist ohnehin durch den Gegenstandsbereich nicht eindeutig vorgegeben, sondern erfolgt willkürlich. Numerische Relationen zwischen den Messungen werden in der Meßtheorie als „sinnlos" bezeichnet, wenn sie keine Entsprechung im empirischen Relativ haben, die abgebildeten beobachtbaren Relationen also nicht auch ohne Messung feststellbar sind (vgl. ebd., 177 f.).
Die Zuverlässigkeit von Messungen beinhaltet als common sense die „Reproduzierbarkeit von Ergebnissen unter den gleichen intersubjektiven Bedingungen" (KRIZ 1978, 84). Das Problem wirklicher äquivalenter Meßwiederholung spielt für die Anwendbarkeit der klassischen Testtheorie eine entscheidende Rolle, da Abweichungen von den laut Testvorgabe konstant zu haltenden Bedingungen unterschiedliche „wahre" Werte und unterschiedliche Fehlervarianzen generieren (vgl. FISCHER 1974, 32). Das bedeutet, daß die zur Ermittlung der Reliabilitätskoeffizienten für die Praxis entwickelten Verfahren wie zum Beispiel Paralleltest, Testwiederholung und Testteilung von der Konstanz der Persönlichkeitsmerkmale und der Konsistenz des Verhaltens ausgehen müssen (vgl. ebd., 45; LIENERT 1969, 210 ff.).
Versuchen wir, diese Überlegungen auf die Methode ‚Unterrichtsbeobachtung' hinsichtlich der vorgestellten extremen Verfahrensweisen und den entsprechenden Funktionen der Beobachter zu beziehen; danach gilt es wohl, zwei Punkten Aufmerksamkeit zu widmen,
(a) Voraussetzungen, die das zu messende Merkmal betreffen, und
(b) Voraussetzungen bezüglich des Verfahrens der Zuverlässigkeitsprüfung.

Zu a: Uns interessierende Untersuchungs*gegenstände* sind „living systems", wie Interaktionsabläufe bzw. Lehr-Lern-Prozesse, sowie die in ihnen aktiv werdenden Individuen Lehrer und Schüler (unter den spezifischen Rahmenbedingungen der Schule). Unterricht als intentional auf Verhaltensänderung (Lernziele!) ausgerichtete Veranstaltung entzieht der oben getroffenen Konstanz/Konsistenzannahme den Boden. Diese Feststellung gilt unseres Erachtens für alle Untersuchungen im Zeitablauf, unabhängig von quantitativen/qualitativen Beobachtungsverfahren.
Zu b: Die „klassischen" Verfahren der Zuverlässigkeitsprüfung (Paralleltest, Testwiederholung und Testteilung) fußen zwar auf der Konstanz/Konsistenz-Annahme,

eine generelle Ablehnung der Brauchbarkeit so begründeter Reliabilitätskoeffizienten ist aber zunächst nur dann zu rechtfertigen, wenn von der unmittelbaren Beobachtung als einziger Möglichkeit ausgegangen wird. Denn die Einmaligkeit jeder Aktivität, die Wichtigkeit ihrer Aufeinanderfolge für Verhaltenskontingenzen verbieten Zerlegungen jeder Art und lassen auch Wiederholungen unsinnig erscheinen, es sei denn, man unterzöge alle am Unterricht Beteiligten einer „Gehirnwäsche" (vgl. FISCHER 1974, 28). So sind im ursprünglichen Sinne auch keine gegeneinanderzuhaltenden Parallelversionen konstruierbar. Denkbar wäre unter Umständen in Anlehnung an die Parellelitätsidee, daß zwei Beobachter (-Teams) mit denselben Kategorien den gleichen Unterrichtsablauf parallel kodieren. Voraussetzung wäre dann allerdings die Äquivalenz der Beobachter (!?).

Der Einsatz der Videotechnik im Sinne eines mittelbaren Beobachtungsverfahrens veranlaßt zu neuen Überlegungen, denn der Unterrichtsablauf kann ohne Beeinträchtigung der am Unterricht Beteiligten durch Lernzuwächse etc. beliebig oft wiederholt werden. Allerdings führt diese Vorgehensweise zu Lernzuwächsen bei den Beobachtern, so daß einer Zuverlässigkeitsprüfung wiederum die Basis genommen wird. Verstärkt wird dieser Sachverhalt hinsichtlich komplexer Kategoriensysteme, in denen man sich die technische Wiederholbarkeit zunutze macht, um beobachtete Aktivitäten unter verschiedenen Aspekten auszuwerten bzw. um den Entscheidungsprozeß zu stabilisieren. Ausweichmöglichkeit bleibt auch hier nur die Mischung der Parallelitäts- und der Wiederholungsidee im „klassischen" Sinne, indem unterschiedliche Beobachter mit dem identischen Kategoriensystem den gleichen Unterrichtsablauf unabhängig voneinander kodieren.
Wir wollen nach dem letztgenannten Verfahren verschiedene statistische Möglichkeiten auf konkrete Daten anwenden (Abschnitt VII), die aus von uns durchgeführten Unterrichtsbeobachtungen (Abschnitt VI) stammen. Im Abschnitt VIII werden wir uns um eine Zusammenfassung der verschiedenen Argumentationsstränge bemühen.

VI. Das Projekt „Kaufvertrag"[1]

Nach Vorarbeiten zu naiv-verhaltenstheoretischen Dispositionskonstrukten im Bereich der „Unterrichtstheorie" von Lehrern und Schülern (ACHTENHAGEN / HEIDENREICH / SEMBILL 1975; ACHTENHAGEN / SEMBILL / STEINHOFF 1979) sollte ein Beitrag zur besseren Beschreibung und – soweit überhaupt möglich – auch zur besseren Erklärung von differentiellen Lernleistungen und individuellen Lernschwierigkeiten aus dem *Zusammenhang* des Unterrichtsablaufes geleistet werden. Dazu wurden Beobachtungen abgeschlossener, zusammenhängender Unterrichtseinheiten zum Thema „Kaufvertrag" im Fach Wirtschaftslehre an kaufmännischen Schulen im Sinne einer oben skizzier-

ten „educational evaluation" durchgeführt. Neben Produktvariablen wurden in einem kombinierten quantitativen/qualitativen Verfahren Unterrichtsablaufvariablen (Kontextvariablen) erfaßt. Beide Variablenkomplexe bezogen sich jeweils auf Lernmaterial (Lernobjekt-)Variablen (zum Beispiel Verständlichkeit oder Lösungsschwierigkeit) sowie personabhängige Variablen im kognitiven und affektiven Bereich (zum Beispiel Intelligenz; Schulangst oder Kongruenzgrad der geforderten Antworten; Art der Rückmeldung).

Es wurden sowohl unmittelbare als auch per Videoaufnahmen mittelbare Beobachtungen durchgeführt. Letztere wurden anhand eines 18 Spalten umfassenden Kodierbogens von einem Kodier-Team, das aus zwei bis drei Beurteilern (ratern) bestand, durchgeführt. Es gab drei rater-teams, die im wöchentlichen Turnus untereinander ausgewechselt wurden, um systematische Fehler weitgehend zu verhindern. Die Mitglieder eines rater-teams mußten sich jeweils immer auf eine Kodierentscheidung einigen. Es waren somit mehrere Kodierentscheidungen pro beobachteter Aktivität notwendig, je nachdem durch welche Kategorien sie zu kennzeichnen war. Zentrale Kodiereinheit war also die Aktivität, die durch bestimmte Verhaltenselemente der Beteiligten geprägt war. Der Unterrichtsablauf wurde dabei am ‚kognitiven Gesprächsstrang' der Unterrichtsteilnehmer verfolgt und mit Hilfe einer modifizierten Regelkreistechnik (LOUIS 1974; HEYMANN 1977) abgebildet. Der jeweilige Sollwert war dabei an den Lernobjekten orientiert. Für das Verkoden einer Unterrichtsstunde mußten aufgrund der Vielzahl der Variablen und den entsprechend mehrfachen Wiederholungen einzelner Aktivitäten sowie den Entscheidungsfindungen mit einem Aufwand von ca. 12 Zeitstunden gerechnet werden.

Alle an der Auswertung maßgeblich beteiligten Kodierer (Wirtschaftspädagogik-Studenten) haben von Planungsbeginn an bis zum Abschluß der Kodierarbeiten (ca. 1 1/2 Jahre) intensiv mitgearbeitet. Sie kannten die juristische Materie des Kaufvertrages ebenso wie das Lehrmaterial und das den Klassen zur Verfügung stehende Lernmaterial und haben die unmittelbaren Unterrichtsbeobachtungen mit durchgeführt (Kamera, Ton und Protokolle), wobei ihnen die Schüler vom Klassenspiegel und von den Probeaufnahmen her, die auch zur gegenseitigen Gewöhnung durchgeführt wurden, bekannt waren. Sie waren ebenfalls maßgeblich an den Operationalisierungen des Analyseschemas beteiligt, deren Rohfassungen anhand der durchgeführten Filmaufnahmen und zum Teil erstellter Wortprotokolle über mehrere Wochen hinweg verbessert wurden.

Bei der Zuverlässigkeitsprüfung wurden drei 10-Minuten Stichproben jeweils von den drei rater-teams unabhängig voneinander kodiert. Jede Stichprobe entstammte einer anderen Klasse; zwei Lehrer wurden erfaßt.

Die ersten Überprüfungen wurden durchgeführt, nachdem die Hälfte aller Filme kodiert war, die zweiten nach etwa Zweidrittel aller Kodierungen und die dritten in der Schlußphase der Kodierungen.
Das konkrete Zahlenmaterial, das nachfolgend der Veranschaulichung der abgeleiteten Formeln dient, resultiert aus Kodierungen der nominalskalierten Variablen „(zusätzliche) Stimulanz". Diese Variable gehört zu den lehrerspezifischen, affektiven Unterrichtsablaufvariablen. Die Kodierungen umfassen die nonverbalen und paralinguistischen Kanäle der Interaktion. Beurteilt wird, ob der Lehrer von seinem sonst üblichen individuellen Verhalten im Unterricht intentional abweicht.

Kodierweise: 1 = übliches individuelles Verhalten
2 = intentional unübliches, individuelles Verhalten (zusätzliche Stimulanz)

VII. Übereinstimmungsgrad zwischen Beobachtern

Wir wollen in mehreren Schritten die Wirkungsweisen und Interpretationsmöglichkeiten unterschiedlich aufgebauter Übereinstimmungsmaße aufzeigen, zunächst anhand des am meistverwendeten Maßes zur Feststellung der ‚Intercoderreliabilität' (vgl. HERKNER 1974, 177), das wir einfach als Grundmuster eines Übereinstimmungsmaßes bezeichnen.[2]

1. Grundmuster eines Übereinstimmungsmaßes

Die im Zeitablauf beobachteten Aktivitäten werden unter verschiedenen Aspekten kategorisiert. Jede dieser Kategorien hat mehrere Ausprägungen. Geprüft wird, ob die rater bei den Kodierungen jeweils dieselben Ausprägungen der betreffenden Kategorie gewählt haben, oder ob es (wenn ja, welche) Abweichungen gegeben hat.
Das Grundmuster eines Zuverlässigkeitsmaßes für die Übereinstimmung von Kodierungen unterschiedlicher Beobachter können wir folgendermaßen beschreiben: Zwei rater(-teams) nehmen unabhängig voneinander n . . verschiedene Kodierungen vor. *Beide rater ordnen jede einzelne Kodierung einer von K Ausprägungen einer Kategorie zu. Es sei:*

$n_{k\kappa}$ = Anzahl der Kodierungen, die von rater A der Ausprägung k und von rater B der Ausprägung κ zugeordnet werden ($k, \kappa = 1, \ldots, K$).

Dann ist die Gesamtzahl der Kategorisierungen:

$$n\,.\,. = \sum_{k=1}^{K} n_{k.} = \sum_{\kappa=1}^{K} n_{.\kappa} = \sum_{k=1}^{K}\sum_{\kappa=1}^{K} n_{k\kappa}$$

Im Hinblick auf die Messung der Übereinstimmung bei der Kodierung interessieren in erster Linie die Größen n_{kk}:

n_{kk} = Anzahl der Kodierungen, die sowohl vom rater A als auch vom rater B der Ausprägung k zugeordnet werden (k = 1, . . ., K).

Damit erhalten wir die Gesamtzahl der übereinstimmenden Kodierungen:

$$\sum_{k=1}^{K} n_{kk}$$

Wir wollen den Anteil der übereinstimmenden Kodierungen an der Gesamtzahl der Kodierungen das „Grundmuster eines Übereinstimmungsmaßes" ($\hat{P}$) nennen:

$$(1)\quad \hat{P} = \frac{1}{n..} \sum_{k=1}^{K} n_{kk}$$

Die Maßzahl $\hat{P}$ kann nur Werte zwischen 0 und 1 annehmen: Je dichter $\hat{P}$ bei 1 liegt, um so größer ist der Grad der Übereinstimmung zwischen den beiden ratern.[3]
Falls bei der Kodierung nur k = 2 Ausprägungen zugrunde gelegt werden (Tab. 1), erhalten wir für $\hat{P}$:

$$\hat{P} = \frac{1}{n..} (n_{11} + n_{22})$$

Tabelle 1

Team B \ Team A	Ausprägungen 1	2	Σ
Ausprägungen 1	n_{11}	n_{21}	$n_{.1}$
Ausprägungen 2	n_{12}	n_{22}	$n_{.2}$
Σ	$n_{1.}$	$n_{2.}$	$n..$

Zwei wichtige Probleme sollen hier kurz angesprochen werden:

1. Die Frage nach der Identität der vergleichbaren Ereignisse;
2. Die Frage nach den inhaltlichen Implikationen des Begriffes ‚Übereinstimmung'.

Zu 1: Die Übereinstimmung „parallelkodierter" Unterrichtsbeobachtungen festzustellen scheint nur dann sinnvoll, wenn es sich jeweils um dieselben Aktivitäten handelt. Ein extremes fiktives Beispiel soll das verdeutlichen (vgl. Tabelle 2).
Durch Verschiebungen im Zeittakt eines entsprechenden Beobachtungsverfahrens sind nicht identische Ereignisse verglichen worden. Das Team A hat immer 2 kodiert, wenn Team B 1 kodiert hat und umgekehrt. Summativ betrachtet hat jedes Team bei jeder Ausprägung 20 Einheiten kodiert; man könnte eine totale Übereinstimmung vermuten, der Blick in die Zellen der Matrix (insbesondere der Haupt-

Tabelle 2

A / B	1	2	Σ
1	–	20	20
2	20	–	20
Σ	20	20	40

diagonalen) zeigt uns jedoch, daß dies mitnichten der Fall ist: $\hat{P}$ *nach Formel (1) ist 0.*

Ein vergleichbares Vorgehen hat FLANDERS 1960 (vgl. auch: SCHULZ / TESCHNER / VOGT 1970, Sp. 705 f.) gewählt. Die einzige Übereinstimmung, die man hierbei feststellen kann, ist die bezüglich der Gesamteinheiten bzw. die bezüglich der Wahl von einzelnen Ausprägungen (damit könnte man näherungsweise Aussagen über das Einhalten eines vorgegebenen Zeittaktes machen).

Wenn beide Beobachterteams unterschiedlich viele Kodierungen vornehmen, zum Beispiel Σ A = 44 und Σ B = 38, verschärft sich das Problem, da die Erstellung einer Matrix nicht mehr möglich ist. Es ergeben sich Probleme nicht definierter und nicht interpretierbarer Ereignisse (vgl. KRIZ 1978, 93 ff.). Versucht man dieses Problem wie zum Beispiel KOWATRAKUL (1959) mit relativen Daten zu umgehen, handelt man sich nur eine zusätzliche Fehlerquelle ein (vgl. Tabelle 3):

Tabelle 3

	1	2	Σ
A	25 %	75 %	100 %
B	25 %	75 %	100 %

Ohne Kenntnis der absoluten Gesamtzahl der jeweiligen Kodierungen kann man sich keine Vorstellungen über die Größenordnungen der möglicherweise erheblichen Abweichungen machen, ganz abgesehen von der Frage, welche Ereignisse überhaupt miteinander verglichen worden sind.[4]

Zu 2: Schaltet man – wie oben geschehen – die Aktivitäten gleich, bleibt dennoch das Problem ungeklärt, wann eigentlich Übereinstimmung herrscht.[5] So ist zu fragen:

(a) Müssen bei der Feststellung von Übereinstimmung nicht auch die übereinstimmenden Nicht-Wahlen anderer Kategorien (Variablen) berücksichtigt werden?

(b) Müssen nicht auch die übereinstimmenden Nicht-Wahlen anderer Ausprägungen der benutzten Kategorien berücksichtigt werden?

Für die folgenden Ausführungen sei Übereinstimmung wie folgt definiert: *Verschiedene Beobachter kodieren dieselbe Aktivität in dieselbe Ausprägung derselben Kategorie.*

2. *Verfeinerungen des Grundmusters*

Diese Problematik gilt auch für die „Verfeinerungen" des Grundmusters eines Übereinstimmungsgrades, den „Konsistenzindex" und den „Inter-

coderreliabilitätskoeffizienten". da sie sich im wesentlichen nur in dem unterschiedlichen Ausmaß des Einbezugs der Zufallswahrscheinlichkeit unterscheiden.

a) Konsistenzindex

Die von uns eingeführte Maßzahl $\hat{P}$ läßt die Möglichkeit zufälliger Übereinstimmungen bei den Kodierungen beider rater unberücksichtigt. Um dies zu verdeutlichen, wollen wir von folgender Situation ausgehen:
Rater A hat jeder der K Ausprägungen die gleiche Anzahl der Kodierungen – nämlich $\frac{1}{k}$ n . . Kodierungen – zugeordnet. Ebenso hat rater B jeder Ausprägung $\frac{1}{k}$ n . . Kodierungen zugeordnet.
Falls beide rater ihre Zuordnungen unabhängig voneinander getroffen haben[6], werden wir – bei einer sehr großen Anzahl n . . von Kodierungen – als Ergebnis $\frac{1}{K^2}$ n . . Kodierungen erhalten, die beide rater der Ausprägung k zugeordnet haben. *Insgesamt liegen damit*

$$\sum_{k=1}^{K} \left(\frac{1}{K^2} n\,..\right) = \frac{1}{k} n\,..$$

Kodierungen vor, die von beiden ratern jeweils den gleichen Ausprägungen zugeordnet worden sind. Diese Übereinstimmungen sind allein auf den Zufall zurückzuführen und können nicht als Indiz für eine systematische Übereinstimmung bei der Verkodung gewertet werden. Wir fassen zusammen:

Wenn als Ergebnis der Kodierung $\sum_{k=1}^{K} n_{kk}$ *Übereinstimmungen ermittelt werden, so sind davon* $\frac{1}{K}$ n . . *Übereinstimmungen auf Zufallseinflüsse zurückzuführen.* Die Anzahl der nichtzufälligen Übereinstimmungen bei der Kodierung beträgt damit:

$$\sum_{k=1}^{K} n_{kk} - \frac{1}{K} n\,..$$

(Beobachtete Anzahl der Übereinstimmungen – zufällige Anzahl der Übereinstimmungen).

Wenn wir weiter berücksichtigen, daß bei n . . Kodierungen die Maximalzahl der Übereinstimmungen natürlich n . . beträgt, können wir schließlich die folgende Maßzahl bilden:

$$= \frac{\Sigma n_{kk} - \frac{1}{K} n..}{n.. - \frac{1}{K} n..} = \frac{\frac{\Sigma n_{kk}}{n..} - \frac{1}{K}}{1 - \frac{1}{K}} = \left(\frac{K}{K-1}\right)\left(\hat{P} - \frac{1}{K}\right)$$

Diese Kennziffer ist in der Literatur als Konsistenzindex S bekannt (vgl. BENNETT / ALPERT / GOLDSTEIN, nach: HERKNER 1974, 178):

$$(2) \quad \hat{S} = \left(\frac{K}{K-1}\right)\left(\hat{P} - \frac{1}{K}\right)$$

$\hat{S}$ ist eine Maßzahl für die nicht-zufällige Übereinstimmung zwischen zwei ratern. $\hat{S}$ ist gleich Null, wenn die beobachtete und die zufällige Anzahl der Übereinstimmungen gleich ist. Positive Werte von $\hat{S}$ weisen auf eine systematische Übereinstimmung zwischen den beiden ratern hin. *Diese Übereinstimmung ist um so größer, je dichter $\hat{S}$ bei Eins liegt. $\hat{S}$ kann nie größer als $\hat{P}$ werden, wie man leicht aus der folgenden Umformung von (2) ersieht:*

$$\hat{S} = \hat{P} - \frac{1}{K-1}(1 - \hat{P}) < \hat{P}$$

Allerdings wird mit wachsender Anzahl K von Ausprägungen der Unterschied zwischen $\hat{S}$ und $\hat{P}$ immer geringer:

$$\lim_{K \to \infty} (\hat{S} - \hat{P}) = 0$$

Im Gegensatz zu $\hat{P}$ kann $\hat{S}$ auch negative Werte annehmen: In einem solchen Fall ist die beobachtete Anzahl von Übereinstimmungen noch geringer als die auf Zufallseinflüsse zurückzuführende Anzahl von Übereinstimmungen. *Es gilt*:

$$-\frac{1}{K-1} < \hat{S} < +1$$

Falls der Verkodung nur K = 2 Ausprägungen zugrunde liegen, erhalten wir:

$$\hat{S} = 2\hat{P} - 1$$

$$-1 < \hat{S} < +1$$

b) Intercoderreliabilitätskoeffizient

Bei der Ermittlung des Konsistenzindex $\hat{S}$ wird die Anzahl zufälliger Übereinstimmungen unter der Voraussetzung bestimmt, daß beide rater jeder Ausprägung jeweils die gleiche Anzahl Kodierungen $(\frac{1}{K} n..)$ zu-

ordnen. Der Konsistenzindex ist somit streng genommen nur eine sinnvolle Maßzahl für die Übereinstimmung zwischen den ratern, wenn diese Voraussetzung – zumindest näherungsweise – erfüllt ist. Wir wollen deshalb im folgenden eine weitere Maßzahl herleiten, die ohne diese Voraussetzung auskommt und auch in Situationen angewandt werden kann, die eine Berechnung von $\hat{S}$ fragwürdig erscheinen lassen. Wir werden anschließend zeigen, daß $\hat{S}$ als ein Spezialfall dieser Maßzahl angesehen werden kann.

Werden der Ausprägung k (k = 1, . . ., K) von rater A $n_{k.}$ und von rater B $n_{.k}$ Kodierungen zugeordnet, *so würden sich bei Unabhängigkeit der Zuordnungen*

$$\frac{1}{n..} n_{k.}\, n_{.k}$$

Kodierungen ergeben, die von beiden ratern zufallsbedingt gleichzeitig der Ausprägung k zugeordnet werden. *Wenn wir die Gesamtzahl zufälliger Übereinstimmungen über alle Ausprägungen $n..P_e$ nennen, erhalten wir:*

$$(3) \quad n..P_e = \frac{1}{n..} \sum_{k=1}^{K} n_{k.}\, n_{.k}$$

Das weitere Vorgehen erfolgt analog zur Herleitung von $\hat{S}$, so daß wir schließlich zu der folgenden Maßzahl gelangen:

$$\frac{\text{Beobachtete – zufällige Anzahl Übereinstimmungen}}{\text{Maximale – zufällige Anzahl Übereinstimmungen}}$$

$$= \frac{\sum n_{kk} - \frac{1}{n..} \sum n_{k.}\, n_{.k}}{n.. - \frac{1}{n..} \sum n_{k.}\, n_{.k}} = \frac{\hat{P} - P_e}{1 - P_e}$$

Diese Maßzahl ist in der Literatur als Intercoderreliabilitätskoeffizient $\hat{\Pi}$ bekannt (vgl. Herkner 1974, 178)[7]*:*

$$(4) \quad \hat{\Pi} = \frac{\hat{P} - P_e}{1 - P_e}$$

Ähnlich wie $\hat{S}$ ist auch die Maßzahl $\hat{\Pi}$ gleich Null, wenn die beobachtete Anzahl der Übereinstimmungen gleich der zufälligen Anzahl der Übereinstimmungen ist. Positive Werte von $\hat{\Pi}$ sind Indiz für eine systematische Übereinstimmung zwischen den beiden ratern. *Aus (4) folgt:*

$$\hat{\Pi} = \hat{P} - \frac{P_e}{1 - P_e}(1 - \hat{P}),$$

so daß wir erhalten:

$\hat{\Pi} < \hat{P}$

und

$$-\frac{P_e}{1-P_e} < \hat{\Pi} < +1$$

$\hat{\Pi}$ ist wie $\hat{S}$ eine Maßzahl für die nicht-zufällige Übereinstimmung zwischen zwei ratern. Allerdings erfolgt – wie bereits erwähnt – die Berechnung von $\hat{S}$ unter der Voraussetzung, daß beide rater einer jeden Ausprägung die gleiche Anzahl ($\frac{1}{k}$ n . .) von Kodierungen zuordnen. Ist diese Voraussetzung erfüllt, so stimmen $\hat{\Pi}$ und $\hat{S}$ überein, *denn jetzt gilt:*

$$n_{k.} = n_{.k} = \frac{1}{K} n\,.\,. \quad (K = 1, \ldots, K);$$

eingesetzt in die Herleitungsformel von $\hat{\Pi}$ *ergibt sich:*

$$\hat{\Pi} = \frac{\Sigma\, n_{kk} - \frac{1}{n..} \Sigma\, n_{k.} n_{.k}}{n.. - \frac{1}{n..} \Sigma\, n_{k.} n_{.k}}$$

$$= \frac{n..\hat{P} - \frac{1}{n..} \Sigma\, (\frac{1}{k} n..)(\frac{1}{k} n..)}{n.. - \frac{1}{n..} \Sigma\, (\frac{1}{k} n..)(\frac{1}{k} n..)}$$

$$= \frac{n..\hat{P} - n..\frac{1}{K}}{n.. - n..\frac{1}{K}}$$

$$= \frac{\hat{P} - \frac{1}{K}}{1 - \frac{1}{K}} = (\frac{K}{K-1})(\hat{P} - \frac{1}{K}) = \hat{S}$$

$\hat{S}$ kann somit als ein Spezialfall von $\hat{\Pi}$ *angesehen werden.*

3. Praktische Anwendung der Übereinstimmungsmaße

Wir wollen die Anwendung der eingeführten Übereinstimmungsmaße an einem Beispiel veranschaulichen. Dazu greifen wir auf die in Tabelle 4 zusammengestellten Daten eines Gruppenvergleichs (rater-teams A und B) zurück.

Tabelle 4

Team B \ Team A	Ausprägungen 1	2	Σ
Ausprägungen 1	36	13	49
2	4	2	6
Σ	40	15	55

Wir bestimmen:

$$\hat{P} = \frac{1}{n..} \sum_{k=1}^{K} n_{kk} = \frac{1}{55}(36 + 2) = 0.691$$

$$\hat{S} = \left(\frac{K}{K-1}\right)\left(\hat{P} - \frac{1}{K}\right) = 2\,(0.691 - 0.500) = 0.382$$

und mit

$$P_e = \frac{1}{n..^2} \sum_{k=1}^{K} n_{k.}\, n_{.k} = \frac{1}{55^2}\,(49 \cdot 40 = 6 \cdot 15) = 0.678$$

erhalten wir:

$$\hat{\Pi} = \frac{\hat{P} - P_e}{1 - P_e} = \frac{0.691 - 0.678}{1 - 0.678} = 0.040$$

Dieses Beispiel läßt die Problematik der Maßzahlen $\hat{P}$ und $\hat{S}$ sehr deutlich werden. Mit der Maßzahl $\hat{P}$ werden unterschiedslos zufällige und nichtzufällige Übereinstimmungen erfaßt: $\hat{P}$ fällt somit vergleichsweise groß aus und ist zur Kennzeichnung der Übereinstimmung zwischen den beiden ratern wenig geeignet. Ähnliches gilt für die Maßzahl $\hat{S}$: Zwar wird hier auf die nicht-zufälligen Übereinstimmungen abgestellt, doch läßt sich $\hat{S}$ nur dann sinnvoll anwenden (s. o.), wenn beide rater jeder der beiden Ausprägungen (zumindest näherungsweise) jeweils die gleiche Anzahl von Kodierungen ($\frac{1}{K}\, n\, .\, . = \frac{1}{2}55$) zugeordnet haben. Das ist im vorliegenden Beispiel ganz offensichtlich nicht der Fall.
In welchem Maße $\hat{S}$ dann zu einer Fehleinschätzung des Übereinstimmungsgrades führt, wenn die Voraussetzungen für die Anwendung von $\hat{S}$ nicht erfüllt sind, zeigt sich an dem für $\hat{\Pi}$ errechneten Wert: Dieser ist deutlich kleiner als $\hat{S}$ und stellt im vorliegenden Beispiel eine angemessene Kennziffer zur Abschätzung des Übereinstimmungsgrades dar.
Die Genauigkeit dieser Abschätzung stellt ein weiteres Problem dar, auf das wir bislang noch nicht eingegangen sind. $\hat{P}$ ebenso wie $\hat{S}$ und $\hat{\Pi}$ sind

aufgrund von Stichprobenbeobachtungen ermittelt worden. Diese Stichprobe besteht im vorliegenden Beispiel aus $n\,..\,= 55$ Kodierungen. Es wird bezüglich jeder Kodierung festgestellt, ob beide rater sie der gleichen oder zwei unterschiedlichen Ausprägung(en) zugeordnet haben. $\hat{P}$, $\hat{S}$ und $\hat{\Pi}$ stellen so gesehen Zufallsvariablen dar, die als Schätzungsfunktion für die jeweils zugrunde liegenden Parameter P, S und Π nach der Methode ‚maximum likelihood' (vgl. LARSON 1969, 225) herangezogen werden. Wir werden im folgenden Konfidenzintervalle für P, S und Π bestimmen, um so zu Aussagen über die Genauigkeit der Schätzungen dieser Parameter durch die Schätzfunktionen $\hat{P}$, $\hat{S}$ bzw. $\hat{\Pi}$ zu gelangen. *Die Zufallsvariable X = Anzahl der Übereinstimmungen in einer Stichprobe vom Umfang n . ., also*

$$X = n\,..\,\hat{P},$$

ist binominalverteilt mit dem Parameter n . . und P. Es gilt:

$$E(X) = n\,..\,P$$
$$Var\,(X) = n\,..\,P\,(a - P)$$

wobei die unbekannte Varianz von X geschätzt wird durch

$$\hat{Var}(X) = n\,..\,\hat{P}\,(1 - \hat{P}).$$

Unter der Bedingung $n\,.. > \frac{9}{P(1-P)}$ *(CLAUSS / EBNER 1977, 160 f.) ist die Zufallsvariable*

$$\frac{X - n\,..\,P}{\sqrt{n\,..\,P(1-P)}} = \frac{n\,..\,P - n\,..\,P}{\sqrt{n\,..\,P\,(1-P)}}$$

näherungsweise standardnormalverteilt, und es gilt:

$$prob\ (-1.96 \leq \frac{n\,..\,\hat{P} - n\,..\,P}{\sqrt{n\,..\,\hat{P}\,(1-\hat{P})}} \leq +1.96) = 0.95$$

Daraus erhalten wir:

$$prob\ (\hat{P} - 1.96\sqrt{\frac{1}{n\,..}\,\hat{P}\,(1-\hat{P})} \leq P \leq \hat{P} + 1.96\,\sqrt{\frac{1}{n\,..}\,\hat{P}\,(1-\hat{P})}\ ) = 0.95$$

Das bedeutet: Mit einer Wahrscheinlichkeit von 0.95 wird das Konfidenzintervall

$$[\,\hat{P} - 1.96\,\sqrt{\frac{1}{n\,..}\,\hat{P}\,(1-\hat{P})}; \hat{P} + 1.96\,\sqrt{\frac{1}{n\,..}\,\hat{P}\,(1-\hat{P})}\,]$$

den unbekannten Parameter P überdecken. Oder anders formuliert: Wenn wir sehr viele Stichproben vom Umfang $n\,..\, = 55$ ziehen, wird sich in 95 % aller Fälle ein Konfidenzintervall ergeben, das den Parameter P überdeckt.

Wir haben in der vorliegenden Stichprobe den Wert P = 0.691 erhalten. Daraus ergibt sich das folgende *95 %-Konfidenzintervall für P:*

[0.569; 0.813].

In der gleichen Weise bestimmen wir – zunächst in allgemeiner Form – die entsprechenden *Konfidenzintervalle für S:*

$$\left[\hat{S} - 1.96 \sqrt{\frac{1}{n..}\left(\hat{S} - \frac{1}{K-1}\right)(1-\hat{S})};\right.$$

$$\left.\hat{S} + 1.96 \sqrt{\frac{1}{n..}\left(\hat{S} + \frac{1}{K-1}\right)(1-\hat{S})}\right]$$

und für Π^{s}:

$$\left[\hat{\Pi} - 1.96 \sqrt{\frac{1}{n..}\left(\hat{\Pi} + \frac{P_e}{1-P_e}\right)(1-\hat{\Pi})};\right.$$

$$\left.\hat{\Pi} + 1.96 \sqrt{\frac{1}{n..}\hat{\Pi} + \frac{P_e}{1-P_e}}\right)(1-\hat{\Pi})}\ \Bigg]$$

Aufgrund unserer Stichprobenbeobachtungen erhalten wir daraus für S das *95 %-Konfidenzintervall:*

[0.139; 0,625]

und für Π *das Konfidenzintervall:*

[– 0.339; 0.419].

Für unser konkretes Beispiel ergeben sich für P, S und Π *(zusammenfassend) folgende Schätzwerte und Konfidenzintervalle:*

$\hat{P}$ = 0.69 (0.57; 0.81)
$\hat{S}$ = 0.38 (0.14; 0.63)
$\hat{\Pi}$ = 0.04 (– 0.34; 0.42)

Die Berücksichtigung der zufälligen Übereinstimmungen unter zunehmend spezifizierten Voraussetzungen führt zu zunehmend breiten Konfidenzintervallen, d. h. daß die Genauigkeit der Schätzungen, inwieweit beim Kodier-Stichprobenumfang $n\ .\ . = 55$ der jeweils zugrunde liegende Parameter bei 5 % Irrtumswahrscheinlichkeit überdeckt wird, sinkt.
Sehen wir uns die Schätzwerte der drei ermittelten Kennziffern aller neun möglichen Gruppenvergleiche (vgl. Abschn. VI., Design der Zuverlässigkeitsprüfung) bezüglich unserer Beispielvariablen an:

$0.65 < \hat{P}_{1-9} < 0.87$
$0.30 < \hat{S}_{1-9} < 0.74$
$-\ 0.06 < \hat{\Pi}_{1-9} < 0.25$

Offensichtlich haben wir oben einen Gruppenvergleich herausgegriffen, dessen Koeffizienten sich eher im Bereich der unteren Grenzen befinden. Das Entscheidende für unsere Argumentation kommt aber unvermindert deutlich zum Vorschein: Die beträchtliche statistische „Bandbreite", mit der es annähernd gleichgut möglich ist, einen zufriedenstellenden oder einen nicht-vorhandenen Übereinstimmungsgrad zu konstatieren.
Dabei scheint noch ein Interpretationshinweis notwendig zu sein. Sicherlich sind die einzelnen Elemente der Koeffizienten $\hat{S}$ und $\hat{\Pi}$ Prozentwerte, somit kann man die Koeffizienten selbst auch als Prozentwerte auffassen. Allerdings ist es außerordentlich schwierig, sich Klarheit über die Bezugsgröße zu verschaffen. So läßt sich im Gegensatz zum $\hat{P}$ auch nicht sagen, $\hat{\Pi} = 0.5$ ist doppelt so hoch wie $\hat{\Pi} = 0.25$. Eine isolierte Betrachtung von $\hat{\Pi}$ scheint ebenso wenig sinnvoll, wie ein direkter Vergleich mehrerer $\hat{\Pi}$ bezüglich unterschiedlicher Variablen (und erschwert daher auch die Vergleichbarkeit zwischen unterschiedlichen Untersuchungen!).
Die exponierte Herausstellung der Koeffizienten als Prozentwerte (vgl. zum Beispiel RITSERT 1972, 63; MANDL / HUBER 1979, 68) erscheint aufgrund unserer Überlegungen nicht sehr sinnvoll. Man kann sich (inzwischen für die Probleme sensibilisiert) des Eindrucks nicht ganz erwehren, daß hier eine implizite Aufwertung des $\hat{\Pi}$ vorgenommen werden soll, etwa: $\sqrt{\hat{\Pi}} \triangleq \sqrt{r^2}$, einem Zusammenhangsmaß intervallskalierter Daten. Es sei zumindest darauf hingewiesen, daß es für diese Entsprechung keine Grundlage gibt.[9]

VIII. Schlußbetrachtung

Knüpfen wir noch einmal unmittelbar an die Reliabilitätsprobleme des 5. Abschnitts an: Als Voraussetzung der allgemein dargestellten und konkret durchgerechneten „Ausweichmöglichkeiten" eines „Übereinstimmungsgrades zwischen Beobachtern" wurde die Äquivalenz der unterschiedlichen Beobachter genannt, ein Widerspruch in sich selbst. Dennoch bemühen sich quantitativ orientierte Verfahren durch intensive Schulung, eine Normierung der Beobachter/Kodierer nach Vorgabe des (der!) Experten zu erreichen (vgl. FIEGUTH 1977). In diesem Zusammenhang sind auch das Nichtinformiertsein bezüglich der Untersuchungsziele, die rigide, kontext-(schätz-)freien Kategorien und eben auch der eigentümliche Umgang mit den Übereinstimmungkoeffizienten zu verstehen. Die Angabe der Reliabilität im Sinne oben skizzierter Transparenz ist durch einen Koeffizienten allein nicht zu leisten, zu vielfältig (genutzt) sind die „Make ups", von denen wir hier nur einige benannt haben (Erhöhung der Anzahl der Ausprägungen, prozentuale und

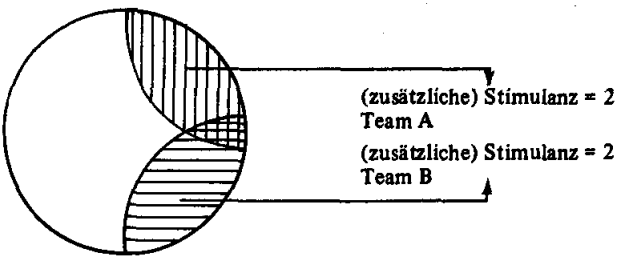
summative Verarbeitung der Daten, die Wahl des „richtigen" Koeffizienten).
Sicherlich sind die Übereinstimmungen in derartigen reliabilitätsgerechten Beobachtungssystemen numerisch höher als bei qualitativen Verfahrensweisen (dabei sollen schon die üblichen Übereinstimmungskoeffizienten für Schätzskalen bei nur 0.50 – 0.60 liegen; vgl. GUILFORD 1971, 149). Aber auch quantitative Verfahren können keine Äquivalenz der Beobachter gewährleisten, abgesehen davon, daß der Kontext außerhalb der Betrachtungsweise bleibt. Man sollte doch im Auge behalten, daß bei einer Unterrichtsanalyse die Zuverlässigkeit im Sinne von Reproduzierbarkeit nicht jene von Beobachterurteilen meinen kann, sondern nur die Reproduzierbarkeit von Ergebnissen beobachteter Unterrichtsaktivitäten in anderen äquivalenten Unterrichtssituationen (vgl. auch: KRIZ 1978, 88 et passim).
Hält man sich das angestrebte Ziel, nämlich Regelmäßigkeiten im Unterrichtskontext zu entdecken und zu beschreiben, vor Augen, muß man sich angesichts der konkreten Vorgehensweise nicht über die wenig ergiebigen Ergebnisse wundern:
Auf der Theorieebene werden Aussagen formuliert, die von nicht nachvollziehbaren Invarianzstrukturen und idealisierenden Bedingungen geprägt sind und für die es auch keine Theorie des Messens gibt. Dementsprechend werden sowohl im Bereich der Zuordnungsregeln als auch bei den dazugehörigen Gütekriterien (hier Reliabilität) diese Gleichförmigkeitsannahmen vorausgesetzt und damit festgeschrieben, siehe kontextfreie Kategorienbildungen, Beobachternormierungen (Äquivalenzannahmen) und Konstanz/Konsistenz-Annahmen der zu untersuchenden Merkmale. Schon aufgrund der Wechselwirkungen zwischen den Ebenen ist es wahrscheinlich, daß sich in den Ergebnissen Regelmäßigkeiten finden werden; ob es allerdings die gesuchten Regelmäßigkeiten sind (sein können), die dann, kausal verknüpft, gesetzmäßig vom Netzwerk der Theorie aufgenommen werden, erscheint sehr fraglich. Fraglich ist unseres Erachtens dann auch, ob man im Sinne KERLINGERs den Ergebnissen wirklich sehr vertrauen kann, wenn sie durch einen sehr hohen Übereinstimmungskoeffizienten gekennzeichnet sind.
Unter Zuhilfenahme qualitativer Untersuchungsverfahren, die vom Anspruch her im wesentlichen auf Hypothesengenerierung ausgelegt sind, können wir uns (leider) auch nicht aller Mängel entledigen. Im theoretischen Bereich stützen sie sich fast zwangsläufig auf Ergebnisse quantitativer Verfahren, eine zugrunde liegende Theorie des Messens wird man auch hier vergeblich suchen. Mit der Beteiligung der Beobachter in ihrer Eigenschaft als reflexionsfähige Individuen wird aber zumindest eine wichtige Maßnahme getroffen, die nicht von vornherein dem Untersuchungsziel zuwiderläuft. Analog dazu werden die Äquivalenz der Be-

obachter ebenso wie die Konstanz/Konsistenz-Annahme bezüglich der zu messenden Merkmale abgelehnt und damit auch die Relevanz oben dargestellter Reliabilitätsmaße im Sinne eines Gütekriteriums. Das ist auch der Grund dafür, daß die getroffene Begriffswahl nur den einfachen Sachverhalt des Prüfungsvorganges benennt und auf eine Überhöhung im Sinne der Konsistenz, Objektivität oder Reliabilität verzichtet (sprachlich entsprechend dem „interjudge agreement"; vgl. MEUX / SMITH 1964).
Unter einem anderen Gesichtspunkt können die Übereinstimmungskoeffizienten recht nützlich sein: Bei der sorgfältigen Analyse der Rohdaten. Sie können wie bei dem dargestellten Beispiel der „(zusätzlichen) Stimulanz" in der zunächst augenscheinlichen Übereinstimmung differenzierte Sichtweisen der Beobachter signalisieren.
In Tabelle 4 können wir sehen, daß 36 der 38 Übereinstimmungen aus den Kodierungen 1/1 resultieren, und die relative Häufigkeit für die Ausprägung 2 lediglich 19,1 % beträgt. Der letztgenannte Wert paßt gut in das Gesamtbild aller Kodierungen bezüglich „(zusätzlicher) Stimulanz": Tabelle 5 zeigt, daß jede fünfte bis sechste Aktivität über alle Lehrer hinweg als intentional unübliches, individuelles Verhalten eingestuft wurde.

Tabelle 5

Lehrer / (zusätzliche) Stimulanz	1	2	3	4	Σ
1	1218	776	332	182	2498
2	275	146	64	45	530
	(18,4 %)	(15,8 %)	(16,6 %)	(19,8 %)	(17,5 %)
Σ	1493	922	386	227	3028

Jedoch scheinen, wie die nachstehende Abbildung zeigen soll, unterschiedliche Auffassungen bezüglich der inhaltlichen Beschaffenheit zusätzlich stimulierender Anregungen zu bestehen:
Nur ein kleiner Anteil (doppelt schraffierte Fläche) der jeweils von Team A und B empfundenen Norm-Abweichungen des beobachteten Gesamtverhaltens des Lehrers (Kreis) bezieht sich auf identische Ereignisse. Wir führen diesen Tatbestand auf das Zusammenfassen zweier kompremierter Interaktionskanäle, des non-verbalen und des paralinguistischen, in einer Variablen unter dem Aspekt der zusätzlichen Anregung zurück. Dadurch können unterschiedliche Wahrnehmungspräferenzen auditiver und visueller Aspekte zum Tragen gekommen sein. Betrachtet man die Beobachter als eine Stichprobe aus der Population



der Wahrnehmungspräferenzträger, der auch die Schüler zuzurechnen sind (vgl. auch: KRIZ 1978, 89), so kann man annehmen, daß bei der Kodierung „2" zumindest für einen Teil der Schüler eine zusätzliche Stimulanz vorgelegen hat. Unter diesem inhaltlichen Gesichtspunkt scheint es nicht aussichtslos, sich die betreffenden Aktivitäten auf ihre anderen Verhaltenselemente hin anzuschauen.
Weitergehende Arbeiten werden sich unseres Erachtens Gedanken machen müssen, wie die beobachtbaren Phänomene besser über die Zuordnungsregeln auf die Theoriebildung Einfluß nehmen können, und wie gegebenenfalls gefundene praxis-(alltags-)relevante Ergebnisse auf ihre Reproduzierbarkeit in äquivalenten Unterrichtssituationen geprüft werden können.
Das Zulassen von Situationalität und unterschiedlichen Relevanzstrukturen führt zu einer Vermischung von Stichproben-, Reliabilitäts- und Validitätsaspekten (vgl. KRIZ 1978, 88). Die Beteiligung der Beobachter zum einen auf seiten des Forschers (Intentionen, Kategoriensystem), zum anderen auf seiten der Untersuchten (Einschätzungen aus der Sicht des Lehrers bzw. der Schüler) verdeutlichen das ebenso wie Ansätze, bei denen teilweise eine Identität zwischen Untersuchten und Beobachtern besteht (vgl. zum Beispiel WAHL u. a. 1977; WAGNER 1977). Die Fragen, wie dabei handlungsleitende von handlungsrechtfertigenden Kognitionen zu unterscheiden sind, und welche Aussagen bei dererlei Konfundierungen von Forschern, „Instrumenten" und „Untersuchungsgegenständen" möglich bzw. welche Interpretationen nötig sind, sind sicherlich längerfristig aktuell. Ohne eine synchrone Weiterentwicklung von wissenschaftstheoretischen Überlegungen und äquivalenten Meßtheorien wird die Unterrichtswissenschaft allerdings zumindest in dem hier diskutierten Bereich zirkulär bleiben.

Anmerkungen

1 Ein erster Bericht über Vorgehensweise und Analyseschema findet sich in SEMBILL / WESELOH 1978.

2 LIENERT 1978 diskutiert das Problem der subjektiven Schätzung von qualitativen oder quantitativ abgestuften qualitativen Merkmalen durch Kundige unter dem Begriff „Bonitur" (vgl. 1 ff. und 483). Dabei ist „Schätzung" hier der „Messung" gegenübergestellt, während im Sinne einer möglichst genauen Bestimmung eines unbekannten Populationskennwertes „Schätzung" (engl.: estimation) einen Meßvorgang voraussetzt (vgl. 68).

3 Eine Durchsicht der Forschungsberichte des Journal of Applied Behavior Analysis (JABA) der Jahrgänge 1968 bis 1975 (KELLY 1977) dokumentiert den hohen Anwendungsgrad des ‚percentage agreement' (Bei ca. der Hälfte der Studien lagen die Werte über 90 %).
Eine im Anschluß daran zum Teil scharf geführte Diskussion im JABA (1977 bis 1979) ist sehr kontrovers. Sie ist auf der einen Seite geprägt von einem Mißtrauen bis zur Ablehnung von „correlational-like measures" (wie zum Beispiel der in 7.2.2 diskutierte ‚Intercoderreliabilitätskoeffizient') bei gleichzeitigem Bemühen in einer „I've got a better agreement measure – Mentalität" (CONE 1979) um vereinfachte, genauere Prozentkoeffizienten und deren Darstellungsweise (vgl. BAER 1977; HOPKINS / HERMANN 1977; YELTON / WILDMAN / ERICKSON 1977; HARRIS / LAHEY 1978; BIRKIMER / BROWN 1979 a und b). Auf der anderen Seite steht der Versuch, spezielle Korrelationsstatistiken trotz „Neuartigkeit" und notwendigen „verfeinerten Statistikkenntnissen" dem Forscher *und* dem Forschungskonsumenten näherzubringen (vgl. HARTMANN 1977; KRATOCHWILL / WETZEL 1977), unter energischen Hinweisen auf Fehlerquellen und mögliche Fehlschlüsse bei der Verwendung der vorgeschlagenen Prozentkoeffizienten (vgl. HARTMANN / GARDNER 1979; HOPKINS 1979).

4 Der in der Einleitung vorgenommene Verweis auf KERLINGER (1979, 808) bezieht sich auf die „überragenden" Ergebnisse KOWATRAKULs (Koeffizienten bis .98), dessen spezifische Vorgehensweise jedoch nicht dargestellt wird.

5 KRIZ (1978, 95) dokumentiert für den Bereich der Inhaltsanalyse dieses Problem recht anschaulich.

6 „Unabhängige Zuordnung" bedeutet hier: Die bedingte Wahrscheinlichkeit, daß rater B eine Kodierung der Ausprägung k zuordnet, unter der Bedingung, daß rater A diese Kodierung der Ausprägung κ zugeordnet hat, ist gleich der unbedingten Wahrscheinlichkeit, daß rater B diese Kodierung der Ausprägung k zuordnet.

7 Einem sehr spezifischen „eigentümlichen Umgang mit Übereinstimmungsmaßen" sind wir hier zunächst aufgesessen: HERKNER (1974, 178) gibt in seiner Formel [4] unter Berufung auf SCOTT (1955) einen *falschen* Ausdruck für P_e wieder:

HERKNER: $P_e = \sum_{l=1}^{k} P_{il} P_{jl}$, wobei P_{il} der relative Anteil der Urteile in Kategorie l beim Koder i ist (P_{jl} analog).

SCOTT (1955, 324): $P_e = \sum_{i=1}^{k} P_i^2$, wobei P_i die relative Häufigkeit der Kodierungen in Kategorie i (relativ zur Gesamtzahl der Kodierungen) ist.

SCOTT geht also von der Annahme aus, daß der Anteil der Urteile in einer Kategorie an der Gesamtzahl der Kodierungen für jeden Kodierer hinreichend gleich ist.
HERKNERs Angabe entspricht pikanterweise dem P_e aus der Kappa-Formel von COHEN (1960), die formal wie das SCOTTsche Π aufgebaut ist. FRIEDE (1981, 2) macht auf diesen Fehler ebenfalls aufmerksam. Die Kappa-Formel hat insbesondere im Psychological Bulletin eine breite Diskussion entfacht (vgl. zum Beispiel COHEN 1968; FLEISS / COHEN / EVERITT 1969; LIGHT 1971; FLEISS 1971; HUBERT 1977 und 1978; FLEISS / NEE / LANDIS 1979; vgl. auch: LIENERT 1978, 636 ff.). Die dort angesprochenen überwiegend speziellen statistischen Probleme (zum Beispiel Gewichtungen nach unterschiedlichen Konventionen, ‚rater-sets', Singulär-Kappa, exakte Varianzangaben, fixierte und variable Randsummen, Generalisierungen) sollen unter unserer mehr methodologisch akzentuierten Fragestellung nicht ausgebreitet werden. Sofern Verweise notwendig erscheinen, werden wir sie im weiteren Verlauf unseres Beitrages geben, nun wohl wissend, daß unser $\hat{\Pi}$ und κ identisch sind.

8 $X = n\,..\,\Pi$, $EX = n\,..\,\hat{\Pi}$, $Var\,(X) = n^2\,..\,Var(\Pi)$.
Da Π unbekannt ist, wird die $Var(\Pi)$ geschätzt durch

$$Var(\hat{\Pi}) = \frac{1}{(1-P_e)^2} \cdot Var(\hat{P}).$$

$$Var\,(\hat{P}) = \frac{P\,(1-P)}{n\,..}$$ muß geschätzt werden durch

$$\hat{Var}(\hat{P}) = \frac{\hat{P}\,(1-\hat{P})}{n\,..}.$$ So ergibt sich letztlich für die geschätzte Varianz von

$$\hat{\Pi}: \hat{Var}(\hat{\Pi}) = \frac{1}{(1-P_e)^2} \cdot \frac{\hat{P}\,(1-\hat{P})}{n\,..}.$$

Dieser Ausdruck entspricht den Angaben zum Beispiel bei SCOTT (1955, 325), dort als erwartungstreue Schätzung mit $(n\,..\,-\,1)$, und COHEN (1960, Formel 7; 1968, Formel 2).
Es ergibt sich somit die näherungsweise standardnormalverteilte Zufallsvariable

$$\frac{n\,..\,(\Pi - \hat{\Pi})}{n\,..\sqrt{\frac{1}{1-P_e)^2} \cdot \frac{\hat{P}\,(1-\hat{P})}{n\,..}}}$$

deren Wurzelausdruck sich in

$$\sqrt{\frac{1}{n\,..}\,(\hat{\Pi} + \frac{P_e}{1-P_e})\,(1 - \hat{\Pi})}$$

umformen läßt.
FLEISS / COHEN / EVERITT (1969, 325, Formel 13; s. auch: HUBERT 1977, 295) zeigen, daß die angegebene Varianz nicht korrekt ist, zu Überschätzungen und damit zu konservativen Signifikanztests und Konfidenzintervallen führt (d. h. man verzichtet eher auf den Nachweis eines bestehenden Unterschiedes, als einen nicht bestehenden Unterschied als bestehend anzugeben; vgl. LIENERT 1973, 150).
Die angegebene korrekte Varianz

$$\hat{Var}(\hat{\kappa}) = \frac{1}{N\,(1-p_c)^4}\,\{\,\sum_{i=1}^{k} p_{ii}\,X\,[(1-p_c) - (p_{.i} + p_{i.})\,(1-p_o)]^2 + (1-p_o)^2 \sum_{\substack{i=1 \\ i\neq j}}^{k}\sum_{j=1}^{k} p_{ij}\,(p_{.i} + p_{j.})^2 - (p_o\,p_c - 2\,p_c + p_o)^2\,\}$$

bestätigt für unser Beispiel die Überschätzungstendenz:
$\hat{Var}(\hat{\kappa})$ = .03223
$\hat{Var}(\hat{\Pi})$ = .03745,
allerdings scheint der Aufwand unter Anwendungsgesichtspunkten fragwürdig.

9 Dieser Interpretationshinweis verdeutlicht auch noch einmal die Brisanz, die in der JABA-Diskussion (s. Anm. 3) um die Anwendungsmöglichkeiten, den Anwendungswillen und den Anwendungsnutzen steckt.

Literatur

Achtenhagen, F.: Einige Überlegungen zum gegenwärtigen Stand der Unterrichtswissenschaft, in: Unterrichtswissenschaft 1979 (Nr. 3), 269 – 282

Achtenhagen, F. / Heidenreich, W.-D. / Sembill, D.: Überlegungen zur „Unterrichtstheorie" von Handelslehrerstudenten und Referendaren des Handelslehramtes, in: DtBFsch (1975) 71, 578 – 601

Achtenhagen, F. / Wienold, G., u. a.: Lehren und Lernen im Fremdsprachenunterricht, 2 Bände, München 1975

Achtenhagen, F. / Sembill, D. / Steinhoff, E.: Die Lehrerpersönlichkeit im Urteil von Schülern, in: Z. f. Päd. (1979) 25, Nr. 2, 191 – 208

Anger, H.: Befragung und Erhebung, in: Graumann, C. F. (Hrsg.), Handbuch der Psychologie, Bd. 7/1, Göttingen 1969

Baer, D. M.: Reviewer's comment: Just because it's reliable doesn't mean that you can use it, in: Journal of Applied Behavior Analysis (JABA) (1977) 10, 117 bis 119

Bennett, E. M. / Alpert, M. R. / Goldstein, A. C.: Communications through Limited Response Questioning, Public Opinion Quarterly (1954) 18, 303 – 308

Birkimer, J. C. / Brown, J. H.: A graphical judgemental aid which summarizes obtained and chance reliability data and helps assess the believevability of experimental effects, in: JABA (1979 a) 12, 523 – 533

– Back to basics: Percentage agreement measures are adequate, but there are easier ways, in: JABA (1979 b) 12, 535 – 543

Böhme, G. / v. Engelhardt, M.: Zur Kritik des Lebensweltbegriffes, in: Böhme, G. / v. Engelhardt, M. (Hrsg.), Entfremdete Wissenschaft, Frankfurt am Main 1979, 7 – 25

Brophy, J. E. / Good, T. L.: Die Lehrer-Schüler-Interaktion, München et al. 1976

Clauß, G. / Ebner, H.: Grundlagen der Statistik, 2. Aufl. Thun et al. 1977

Cohen, J.: A coefficient of agreement for nominal scales, in: Educational and Psychological Measurement (1960) 20, 37 – 46

– Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit, in: Psychological Bulletin (Psych. Bull.) (1968) 70, 213 – 220

Cone, J. D.: Why the „I've got better agreement measure" literature continues to grow: A commentary on two articles by Birkimer and Brown, in: JABA (1979) 12, 571

Fieguth, G.: Beobachtertraining, in: Mees, U. / Selg, H. (Hrsg.), Verhaltensbeobachtung und Verhaltensmodifikation, Stuttgart 1977, 78 – 87

Fischer, G. H.: Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen, Bern / Stuttgart / Wien 1974

Flanders, N. A.: Interaction Analysis in the Classroom. A Manual for Observers. College of Education, University of Minnesota, Minneapolis 1960

Fleiss, J. L. / Cohen, J. / Everitt, B. S.: Large sample standard errors of Kappa and Weighted Kappa, in: Psych. Bull. (1969) 72, 323 – 327
Fleiss, J. L.: Measuring nominal scale agreement among many raters, in: Psych. Bull. (1971) 76, 378 – 382
Fleiss, J. L. / Nee, J. C. M. / Landis, H. R.: Large sample variance of Kappa in the case of different sets of raters, in: Psych. Bull. (1979) 86, 974 – 977
Frick, T. / Semmel, M. I.: Observer Agreement and Reliabilities of Classroom Observational Measures, in: Rev. of Educ. Res. (1978) 48, 157 – 184
Friede, Chr. K.: Verfahren zur Bestimmung der Intercoderreliabilität für nominalskalierte Daten, in: ZEP 1981 (Nr. 1), 1 – 25
Friedrichs, J.: Methoden empirischer Sozialforschung, Reinbek bei Hamburg 1973
Glass, G. V.: Comments on Professor Bloom's Paper, in: Wittrock, M. C. / Wiley, D. E., The Evaluation of Instruction, New York et al. 1970
Groeben, N. / Westmeyer, H.: Kriterien psychologischer Forschung, München 1975
– Normkritik und Normbegründung als Aufgabe der Pädagogischen Psychologie, in: Brandtstädter, J. / Reinert, G. / Schneewind, K. A. (Hrsg.), Pädagogische Psychologie: Probleme und Perspektiven, Stuttgart 1979, 51 – 77
Guilford, J. P.: Persönlichkeit, 5. Aufl. Weinheim et al. 1971
– Personality, New York et al. 1959
Hausser, K. / Krapp, A.: Wissenschaftstheoretische und methodologische Implikationen einer pädagogischen Theorie des Interesses, in: Z. f. Päd. (1979) 25, Nr. 1, 61 – 79
Kerlinger, F. N.: Grundlagen der Sozialisationswissenschaften, Bd. 2, Weinheim et al. 1979
Kowatrakul, S.: Some Behaviors of Elementary School Children Related to Classroom Activities and Subject Areas, in: J. of Educ. Psych. (1959) 50, 121 – 128
Kratochwill, T. R. / Wetzel, R.: Observer agreement, credibility, and judgement: Some considerations in presenting observer agreement data, in: JABA (1977) 10, 133 – 139
Kriz, J.: Zuverlässigkeit und Gültigkeit, in: Lisch, R. / Kriz, J., Grundlagen und Modelle der Inhaltsanalyse, Reinbek 1978, 84 – 104
Larson, H. J.: Introduction to Probability Theory and Statistical Inference, Monterey 1969
Lempert, W. / Hoff, E.-H. / Lappe, L.: Konzeptionen zur Analyse der Sozialisation durch Arbeit. Theoretische Vorstudien für eine empirische Untersuchung. Max-Planck-Institut für Bildungsforschung. Materialien aus der Bildungsforschung Nr. 14, Berlin 1979
Lienert, G. A.: Testaufbau und Testanalyse, 3. Aufl. Weinheim et al. 1969
– Verteilungsfreie Methoden in der Biostatistik, Bd. I und II, Meisenheim am Glan, 1973 und 1978
Light, R. J.: Measures of response agreement for qualitative data: Some generalizations and alternatives, in: Psych. Bull. (1971) 76, 365 – 377
Louis, B.: Unterrichtliche Steuerung und Selbständigkeit des Denkens, München 1974
Mandl, H. / Huber, G. L.: Komplexität schulischer Interaktionsprozesse, in: Unterrichtswissenschaft (1979) 7, 63 – 77
Medley, D. M. / Norton, D. P.: The Concept of Reliability As it Applies to Behavior Records. Paper presented at the meeting of the American Psychological Association, Washington, D. C., 1971
Medley, D. M. / Mitzel, H. E.: Measuring Classroom Behavior by Systematic Observation, 1963, in: Schulz, W. / Teschner, W. P. / Voigt, J., Verhalten im Unterricht. Seine Erfassung durch Beobachtungsverfahren, 1969, in: Handbuch der Unterrichtsforschung, Teil I, Weinheim et al. 1970

Mees, U.: Methodologische Probleme der Verhaltensbeobachtung in der natürlichen Umgebung: II. Beobachter und Beobachtete als mögliche Fehlerquellen von Beobachtungsdaten, in: Mees, U. / Selg, H. (Hrsg.), Verhaltensbeobachtung und Verhaltensmodifikation, Stuttgart 1977, 66 – 77

Merkens, H. / Seiler, H.: Interaktionsanalyse, Stuttgart et al. 1978

Merrill, B.: A Measurement of Mother-Child Interaction, in: J. of Abnormal and Soc. Psych. (1946) XLI, 37 – 49

Meux, M. / Smith, B. O.: Logical Dimensions of Teaching Behavior, in: Biddle, B. J. / Ellena, W. J. (Hrsg.), Contemporary Research on Teacher Effectiveness, New York et al. 1964

Oerter, R.: Welche Realität erfaßt Unterrichtsforschung? in: Unterrichtswissenschaft 1979 (Nr. 1), 24 – 43

Ritsert, J.: Inhaltsanalyse und Ideologiekritik, Frankfurt am Main 1972

Scott, W. A.: Reliability of Content Analysis: The Case of Nominal Scale Coding. Public Opinion Quarterly (1955) 19, 321 – 325

Sembill, D. / Weseloh, G.: Untersuchungen zur Fachdidaktik des Wirtschaftslehreunterrichts – am Beispiel „Kaufvertrag", in: DtBFsch (1978) 74, 587 – 610

Shavelson, R. / Dempsey-Atwood, N.: Generalizability of Measures of Teaching Behavior, in: Review of Educational Research (46) 1976, 553 – 611

Schanz, G.: Grundlagen der verhaltensorientierten Betriebswirtschaftslehre, Tübingen 1977

Schön, B.: Quantitative und qualitative Verfahren in der Schulforschung, in: Schön, B. / Hurrelmann, K. (Hrsg.), Schulalltag und Empirie, Weinheim / Basel 1979, 17 – 29

Schütz, A. / Luckmann, T.: Strukturen der Lebenswelt, Bd. 1, Frankfurt am Main 1979

Schulz, W. / Teschner, W. P. / Voigt, J.: Verhalten im Unterricht, in: Ingenkamp, K. / Parey, E. (Hrsg.), Handbuch der Unterrichtsforschung, Teil I, Weinheim et al. 1970, Sp. 633 – 852

Schwarzer, R.: Instrumente der empirischen Curriculumevaluation, in: Frey, K. (Hrsg.), Curriculum-Handbuch, Bd. II, München 1975, 748 – 766

Stake, R. E.: Verschiedene Aspekte pädagogischer Evaluation, in: Wulf, Christoph, Evaluation, München 1972, 92 – 112

Ulich, D.: Pädagogische Interaktion, Weinheim / Basel 1976

– Erziehungswissenschaft: Erfahrungswissenschaftliche Methoden, in: Hierdeis, H. (Hrsg.), Taschenbuch der Pädagogik, Teil 1, Baltmannsweiler 1978, 296 – 312

Wagner, A. C. / Uttendorfer-Marek, I. / Weidle, R.: Die Analyse von Unterrichtsstrategien mit der Methode des „Nachträglichen Lauten Denkens" von Lehrern und Schülern zu ihrem unterrichtlichen Handeln, in: Unterrichtswissenschaft 1977 (Nr. 3), 244 – 253

Wahl, D. / Schlee, J. / Lutz, M. / Reinhard, W.: Naive Verhaltenstheorie von Lehrern, Weingarten 1977

Westmeyer, H.: Zur Handlungsrelevanz der Verhaltenstheorie. Einige kritische Aspekte, in: Krumm, V. (Hrsg.), Zur Handlungsrelevanz der Verhaltenstheorie, München / Wien / Baltimore 1979, 146 – 155

Yelton, A. R. / Wildmann, B. G. / Erickson, M. T.: A probability-based formula for calculating interobserver agreement, in: JABA (1977) 10, 127 – 131