Zeitschrift für experimentelle und angewandte Psychologie
1980, Band XXVII, Heft 4, S. 670—687

Binäre Kodierung von Sprechen und Blicken: Validität, Reliabilität und ihre Abhängigkeit von der zeitlichen Auflösung*)

Helmut Wagner, Johann H. Ellgring, Andrew H. Clarke

Max-Planck-Institut für Psychiatrie, München
— Abteilung Psychologie —

Validität und Reliabilität der Sprechkodierung einer Beobachtergruppe wurden in Abhängigkeit von der zeitlichen Auflösung untersucht. Die Validität wurde anhand der Übereinstimmung der Beobachter mit einem automatischen Sprachdetektor berechnet. Die Reliabilitätswerte für die Kodierung von Sprechen und Blikken ergaben sich aus der Übereinstimmung der Beobachter untereinander. Im wesentlichen zeigten sich folgende Ergebnisse:

1. Die Validität/Reliabilität der Sprechkodierung ist eine monotone, nichtlineare Funktion der gewählten Auflösung. Die systematischen Fehler, die auf Latenz und Trägheit der menschlichen Beobachter zurückgehen, werden bei einer Auflösung von 400 msec nahezu vollständig unterdrückt.
2. Weder bei der Erfassung des Sprech- noch des Blickverhaltens lassen sich Anzeichen für Observer-Drift feststellen. Trainierte und untrainierte Beobachter unterscheiden sich nicht signifikant.
3. Die Kodierung des Sprechverhaltens ist geringfügig reliabler als die des Blickverhaltens. Dieser Unterschied kann in der Praxis vernachlässigt werden.

1. Einführung

In der vorliegenden Arbeit wird die Genauigkeit menschlicher Beobachter unter zwei Aspekten untersucht.

1. Welche Genauigkeit ist bei der kontinuierlichen Beobachtung einfacher Verhaltensweisen in Echtzeit erreichbar?

2. Wie genau sind die so gewonnenen Beobachtungsdaten im Vergleich mit automatisch erhobenen Daten?

*) Gefördert aus Mitteln der DFG, Antrag Nr. El 67/1.

Diese Probleme stehen in der verzweigten Forschungstradition zur Reliabilität und Validität von Verhaltensbeobachtungen. Üblicherweise werden Reliabilitätsstudien zur Überprüfung einer bestimmten Meßanordnung (v. Cranach & Ellgring, 1973) oder eines bestimmten Kodierungssystems (Vine, 1971) erstellt. Den jeweiligen besonderen Bedingungen entsprechen besondere, nicht verallgemeinerbare Reliabilitätswerte. Das ist bei der Verschiedenheit der Fragestellungen, der verwendeten Methoden und der experimentellen Situationen zwangsläufig.

Die Bemühungen um die Vergleichbarkeit von Ergebnissen sind daher im wesentlichen darauf gerichtet, Rahmenbedingungen und Empfehlungen zu formulieren, deren Befolgung den gängigen Genauigkeitsstandard sichert. In einem zusammenfassenden Artikel nennt Kazdin (1977) als Faktoren, die positiv auf die Meßgenauigkeit einwirken, das Wissen der Beobachter um die Prüfsituation und den Schwierigkeitsgrad des Materials, an dem sie geschult werden. Die Beobachtungsgenauigkeit wird negativ beeinflußt von der Komplexität der Aufgabe, wie etwa der Differenziertheit des verwendeten Kategoriensystems, und durch Observer-Drift, die auf die Veränderung der individuellen Kategoriendefinitionen im Lauf der Beobachtungspraxis zurückgeht.

Wenn man ein hohes Niveau der Reliabilitätswerte anstrebt, sollte man daher die Komplexität der Aufgabe niedrig ansetzen, den Beobachtern mitteilen, daß ihre Leistungen überprüft werden und möglichst schwieriges Übungsmaterial auswählen. Zwar ist eine hohe Reliabilität Voraussetzung für hinreichend valide Beobachtung. Doch kann von der Genauigkeit einer Messung keine strikte Schlußfolgerung auf ihre Validität gezogen werden. Diese prinzipielle Beschränkung von Reliabilitätsstudien entfällt, wenn man einen zugleich reliablen und validen Beobachter zur Verfügung hat, der Bezugsdaten liefert. Die Validität eines Beobachters unterliegt Einschränkungen:

Erstens bezieht sich die Validitätsaussage nicht auf die Gesamtheit des beobachtbaren Verhaltens, sondern nur auf die im Rahmen der Beobachtungsaufgabe definierten Aspekte.

Zweitens gilt die Validitätsaussage ausschließlich für solche Ereignisse, deren Dauer oberhalb der zeitlichen Auflösungsschwelle des Beobachters liegt.

Die vollständige Beschreibung der Beobachter-Validität muß daher auch die Angabe des zeitlichen Auflösungsvermögens enthalten. Wir untersuchten die Validität bzw. Reliabilität für die Kodierung von Sprech- und Blickverhalten auf dem einfachsten Komplexitätsniveau, als On-off-Sequenz. Als idealer Beobachter für das Sprechverhalten fungierte ein automatischer Sprachdetektor. Die Daten dieses Gerätes wurden als Validitätskriterium für die Daten der Beobachter verwendet. Im Fall geringer Übereinstim-

mung zwischen beiden Datensätzen könnte die Kodierungsaufgabe den Fähigkeiten der Beobachter entsprechend modifiziert werden.

Der Gang der Untersuchung wurde durch drei Fragestellungen bestimmt:

1. Wie valide sind menschliche Beobachter bei der Kodierung des Sprechverhaltens in Echtzeit? Im Rahmen dieser Fragestellung wird der Effekt der variierten zeitlichen Auflösung dargestellt.
2. Finden sich Anzeichen für Observer-Drift?
3. Wie reliabel und stabil sind die Ergebnisse der Blickkodierung im Vergleich mit der Kodierung des Sprechverhaltens?

Die vorliegende Untersuchung war in erster Linie pragmatisch, insofern wir mehr am Betrag von Validität und Reliabilität als an der Analyse ihrer Bedingungen interessiert waren.

Die Beobachtung einer einzigen Variablen und deren binäre Kodierung in zwei Kategorien ist eine Aufgabe von sehr niedrigem Komplexitätsgrad. Dementsprechend erwarteten wir generell hohe Validitäts- und Reliabilitätswerte.

2. Methoden

Material und Beobachter

Das Videomaterial, das für alle Testdurchgänge benutzt wurde, war ein 270 Sekunden langer Ausschnitt aus einem klinischen Interview mit einer agitiert depressiven Patientin. Die Beobachtung dieses Ausschnittes wurde als besonders schwierig eingeschätzt, da das kommunikative Verhalten der Patientin sehr sprunghaft und unvorhersagbar war. Sprechen und Blicken der Patientin wurden von insgesamt zehn Beobachtern, fünf männlichen und fünf weiblichen, kodiert. Die Beobachter waren zwischen 23 und 33 Jahre alt. Fünf von ihnen waren in dieser Form der Beobachtung bereits ausgiebig trainiert, die fünf anderen verfügten über keinerlei Vorerfahrung.

Jeder Beobachter sah den Interviewausschnitt mit normaler Geschwindigkeit; zweimal für die Kodierung des Sprechens und zweimal für die Kodierung des Blickens. Sprechen und Blicken wurden abwechselnd kodiert. Die zehn Beobachter wurden in drei Gruppen mit zweimal vier und einmal zwei Mitgliedern zusammengefaßt, die dieselbe Verhaltensweise simultan kodierten. Jede Gruppe bestand zur Hälfte aus trainierten und untrainierten Beobachtern. In den Durchgängen der Sprechkodierung wurde parallel ein automatischer Sprachdetektor eingesetzt. Das Material wurde über Video-Monitor dargeboten. Die Beobachter waren durch Kopfhörer, mit denen der Ton des Interviews übertragen wurde, akustisch voneinander

abgeschirmt. Sie wurden instruiert, die An- und Abwesenheit von Sprechen zu kodieren bzw. die Bedingungen „Anblicken des Partners" und „Blickabwendung vom Partner" zu diskriminieren. Sie wurden dazu angehalten, so aufmerksam wie möglich zu kodieren. So sollten beispielsweise auch kurze Pausen in einer Äußerung erfaßt werden.

Instrumente

Die kontinuierlichen Kodierungsentscheidungen wurden über prellfreie Drucktasten abgegeben und von einem Prozeßrechner PDP 8 F registriert und gespeichert. Für jeden Beobachter war ein Eingangskanal des Prozessors reserviert. Der Sprachdetektor bildete die Hüllkurve der Sprachamplitude und kodierte anhand eines Schwellenkriteriums den On-off-Verlauf des Sprechens. Die Einstellung des Schwellenkriteriums wurde optisch-akustisch kontrolliert: Ein optisches Signal zeigte den Entscheidungszustand des Sprachdetektors an und wurde vom Versuchsleiter mit der akustischen Aufzeichnung verglichen. Die Schwelle wurde so eingestellt, daß Ton und optisches Signal zur Deckung kamen. Diese Einstellung und die Einstellung der Lautstärke, die den Beobachtern dargeboten wurde, blieb über sämtliche Versuchsdurchgänge hinweg konstant. Der Sprachdetektor und das Datenerfassungssystem werden im technischen Detail von Clarke, Wagner, Rinck und Ellgring (1979) beschrieben.

Sieben Datenkanäle wurden vom Prozessor parallel abgespeichert. Die Kanäle 1 und 2 waren für den Sprachdetektor reserviert, die Kanäle 3—6 für maximal vier Beobachter. Kanal 2, der für die Kodierung eines zweiten Sprechers verfügbar ist, wurde in dieser Untersuchung nicht verwendet. Der siebte Kanal war mit dem Zeitkode belegt, der auf einer Tonspur des Videobandes aufgezeichnet ist. Mit Hilfe des Zeitkodes kann ein zeitlich definierter Interviewausschnitt beliebig oft bildgenau wiederholt werden.

Zeitliche Auflösung

Die zeitliche Auflösung bestimmt, welche Mindestdauer ein kodiertes Ereignis aufweisen muß, um bei der Auswertung noch berücksichtigt zu werden. Bei einer Auflösung von 1000 msec wird ein Zeitfenster dieser Länge über den gesamten Datensatz geschoben. Ereignisse von weniger als einer Sekunde Dauer werden unterdrückt und dem unmittelbar vorhergehenden überschwelligen Ereignis zugeschlagen. Die Untergrenze der zeitlichen Auflösung betrug 40 msec. Das entspricht der Dauer eines Videobildes. Die Auflösung konnte in Schritten von 40 msec bis zu 2000 msec variiert werden.

3. Analyse

Validität der Beobachter

Die Validität der Beobachter wurde als prozentuale Übereinstimmung (abgekürzt PA, von „percentage agreement") der Datensequenz jedes einzelnen Beobachters aus dem ersten Durchgang der Sprechkodierung (S 1) mit der Datensequenz des Sprachdetektors definiert und nach der Formel

$$PA = \frac{\text{Dauer der Übereinstimmung}}{\text{Gesamtdauer}} \times 100$$

berechnet.

Die Dauer der Übereinstimmung ergab sich, indem die Zustände der Beobachter- und der Detektorsequenz in jedem Zeitpunkt verglichen und diejenigen Zeitpunkte aufsummiert wurden, in denen beide Sequenzen den Zustand 0 oder den Zustand 1 aufwiesen. Dieses Verfahren liefert schärfere Ergebnisse als der Vergleich der prozentualen Dauer.

Außer der prozentualen Übereinstimmung (PA) wurde der Index Kappa (k) nach der Formel

$$k = \frac{PA - c}{100 - c}$$

berechnet.

c ist der Prozentbetrag der nach Zufall zu erwartenden prozentualen Übereinstimmung und ergibt sich aus dem Produkt der Randhäufigkeiten.

$$c = (\text{S off von D} \times \text{S off von O} + \text{S on von D} \times \text{S on von O}) \times 100$$

Hierbei ist „S off von D" die relative Zeitdauer, für die der Detektor die Abwesenheit von Sprechen kodiert, „S off von O" die relative Zeitdauer, für die der Beobachter (Observer), die Abwesenheit von Sprechen kodiert. Entsprechendes gilt für S on-Zustände.

In einschlägigen Artikeln (Hartmann, 1977) wird darauf hingewiesen, daß Kappa ein besseres Maß für die Inter-Rater-Reliabilität darstellt als die prozentuale Übereinstimmung und zwar hauptsächlich deshalb, weil es den Zufallsanteil c berücksichtigt. Dem Betrag nach gleiche PAs werden durch Kappa hinsichtlich der gegebenen Randhäufigkeiten relativiert.

Allgemein gilt $k < PA$,

$$\text{weil } \frac{PA - c}{100 - c} < \frac{PA}{100}$$

Daher wird allgemein (vgl. Hartmann, 1977) das Kriterium für k mit > 0.6 niedriger angesetzt als das Kriterium für PA mit $> 80\%$. Die Über-

einstimmung zwischen Detektor und Beobachtern (abgekürzt DOA, von 'detektor-observer agreement'), ausgedrückt in Prozent und als k, wurde mit Auflösungen von 40, 120, 200, 280, 400, 600 und 1000 msec berechnet.

Observer-Drift

Die Observer-Drift bei der Sprechkodierung wurde kontrolliert, indem die Detektor-Beobachter-Übereinstimmung (DOA) des zweiten Durchgangs der Sprechkodierung (S2) gegen die DOA von S1 einseitig auf Verschlechterung untersucht wurde. Die Auflösung wurde für beide Durchgänge auf 400 msec festgesetzt.

Mit einer zweiten Methode wurden Sprech- und Blickkodierung auf Anzeichen von Observer-Drift untersucht. Bei einer Auflösung von 400 msec wurde in beiden Durchgängen der Sprech- (S1 und S2) und der Blickkodierung (B1 und B2) die Übereinstimmung zwischen den Beobachtern (abgekürzt IOA, von 'inter-observer agreement') bestimmt. Die Berechnung der IOA entspricht genau dem oben beschriebenen Algorithmus für die Berechnung der DOA.

Aus der Menge der möglichen Beobachterpaare wurden diejenigen ausgewählt, die aus einem untrainierten und einem trainierten Beobachter zusammengesetzt sind. Die IOAs dieser Paare wurden einseitig gegen die Hypothesen

$$IOA_{S1} > IOA_{S2} \text{ und } IOA_{B1} > IOA_{B2} \text{ geprüft.}$$

Reliabilitätsvergleich zwischen Sprechen und Blicken

Die Kodierungen der beiden Variablen „Sprechen" und „Blicken" sind nur hinsichtlich ihrer Reliabilität vergleichbar, da für die Erfassung des Blickverhaltens kein valider Beobachter vergleichbar dem Sprachdetektor zur Verfügung stand.

Um zu klären, ob beide Variablen gleich zuverlässig kodiert werden, wurden die IOAs der Durchgänge S1 und B1 intraindividuell verglichen. Die zeitliche Auflösung betrug 400 msec.

Da die Überprüfung der Drift lediglich Verschlechterungen der Übereinstimmung erfaßt, wurden zur Bestimmung der Retest-Reliabilität die IOAs beider Durchgänge innerhalb der Variablen hinsichtlich ihrer Niveaukonstanz verglichen. Zusätzlich wurden die Korrelationen der IOAs zwischen S1/S2 und B1/B2 berechnet.

4. Ergebnisse

Validität der Beobachter

Die Validität der Beobachter, ausgedrückt als Übereinstimmung zwischen Beobachter und Detektor (DOA), steigt mit zunehmender Vergrößerung der zeitlichen Auflösung monoton und nichtlinear an. In Tabelle 1 sind die Übereinstimmungen jedes Beobachters mit dem Detektor (in Prozent und als Kappa) für die verschiedenen Auflösungen im Durchgang S1 zusammengestellt. In den beiden letzten Zeilen stehen die Durchschnittswerte und die durchschnittliche nach Zufall zu erwartende Übereinstimmung $\bar{c}$. Die Zunahme der Übereinstimmung mit gröber werdender Auflösung läßt sich in jedem Einzelfall verfolgen. Bemerkenswert ist, daß $\bar{c}$ bis 400 msec nahezu konstant bleibt.

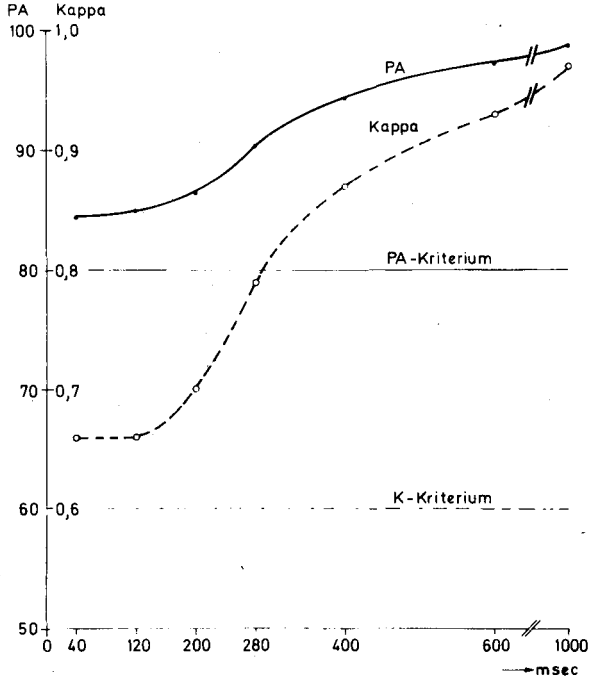

Abb. 1
Abhängigkeit der durchschnittlichen prozentualen Übereinstimmung (PA) und des durchschnittlichen Kappa von der zeitlichen Auflösung.

Tabelle 1

Individuelle und durchschnittliche Werte der DOA im Durchgang „Sprechen 1" mit verschiedenen zeitlichen Auflösungen

Auflös. (msec)	40		120		200		280		400		600		1000	
DOA	%	K	%	K	%	K	%	K	%	K	%	K	%	K
Beobacht.														
1	84,5	.66	84,6	.66	86,2	.70	91,9	.82	94,8	.88	97,8	.95	99,3	.98
2	82,6	.62	82,7	.62	84,2	.65	89,0	.76	92,4	.83	94,8	.88	96,8	.92
3	84,3	.66	85,4	.68	87,3	.72	93,1	.85	96,9	.93	98,2	.96	99,2	.98
4	84,7	.67	85,5	.68	87,8	.73	92,8	.84	94,8	.88	96,6	.92	98,2	.96
5	82,9	.61	82,5	.59	84,5	.64	86,9	.70	90,7	.77	95,0	.88	96,2	.91
6	84,9	.65	85,3	.66	86,3	.68	89,5	.76	94,3	.86	98,0	.95	100,0	1.0
7	84,3	.65	84,2	.65	86,2	.69	90,2	.78	95,0	.88	97,4	.94	97,5	.94
8	84,6	.66	84,8	.66	85,6	.67	89,0	.75	94,3	.87	98,2	.96	100,0	1.0
9	84,9	.66	85,6	.69	87,6	.73	89,9	.78	95,4	.90	97,6	.95	99,3	.98
10	87,7	.67	87,9	.74	90,7	.79	92,6	.84	96,6	.92	98,2	.96	100,0	1.0
$\bar{x}$	84,5	.66	84,9	.66	86,6	.70	90,5	.79	94,5	.87	97,2	.93	98,7	.97
$\bar{c}$	54,9		55,3		55,2		55,4		56,6		58,2		60,6	

Die Validität nimmt mit gröber werdender Auflösung zu. Sie beginnt bei 40 msec mit etwa 85% (Kappa = 0.66) und steigt zwischen 120 und 280 msec beschleunigt an. Bei Auflösungen von mehr als 400 msec geht sie in eine Asymptote über, die bei 1000 msec etwa 99% (Kappa = 0.97) erreicht. Die Beschleunigung der Kappawerte ist stärker ausgeprägt als die der Prozentwerte.

Observer-Drift

Der Vergleich der DOAs aus beiden Kodierungsdurchgängen, in denen der Detektor valide, retest-stabile Bezugswerte liefert, erbringt keine Anhaltspunkte für Observer-Drift. Das gilt auch für die entsprechenden IOAs. In diesem und den folgenden Auswertungsschriften wurden ausschließlich die Daten verwendet, die sich bei einer Auflösung von 400 msec ergeben. In Tabelle 2a werden die Prozentwerte der DOA aus beiden Durchgängen einander gegenübergestellt. Die Hypothese, daß vom ersten zum zweiten Durchgang eine Verschlechterung stattfindet, konnte nicht bestätigt werden (Wilcoxon-Test, $p > 0.05$).

In einem weiteren Schritt wurden nur die Paare berücksichtigt, die aus einem trainierten und aus einem untrainierten Beobachter zusammengesetzt sind. Die IOAs dieser Paare sind in Tabelle 2b zusammengestellt. Weder für die Variable „Sprechen" noch für die Variable „Blicken" konnte eine Verschlechterung der IOA von der ersten zur zweiten Kodierung festgestellt werden (Wilcoxon-Test, $p > 0.05$).

Reliabilitätsvergleich zwischen Sprechen und Blicken

Der Reliabilitätsvergleich zwischen Sprechen und Blicken wurde in drei Stufen vollzogen. Der Vergleich der IOAs aus den ersten beiden Kodierungsvorgängen (Zeile 1 und Zeile 3 von Tabelle 3) ergab einen signifikanten Unterschied auf dem 5% − Niveau zugunsten von Sprechen (Wilcoxon-Test). Diese Variable wird demnach zuverlässiger kodiert.

Die Vergleiche der ersten und zweiten Durchgänge innerhalb jeder Variablen (Tabelle 3: Zeile 1 versus Zeile 2 und Zeile 3 versus Zeile 4) ergaben keine signifikanten Unterschiede (Wilcoxon-Test).

Das bedeutet zugleich, daß auch hier keine Observer-Drifts festzustellen sind.

Die Korrelationen zwischen den IOAs der ersten und zweiten Durchgänge betragen − 0.02 bei Sprechen und + 0.36 bei Blicken. Damit liegen beide Koeffizienten unterhalb der Konfidenzschranke von + 0.55. In beiden Variablen besteht kein statistisch signifikanter Zusammenhang zwischen den IOAs des ersten und des zweiten Kodierungsdurchgangs.

Tabelle 2a
DOA(%) für beide Durchgänge mit 400 msec zeitlicher Auflösung

Beobachter	1	2	3	4	5	6	7	8	9	10
Durchgang S1	94,8	92,4	96,9	94,8	90,7	94,3	95,0	94,3	95,4	96,6
Durchgang S2	94,1	94,6	95,3	94,6	95,7	96,9	92,5	96,1	92,3	94,4

Tabelle 2b
IOAs(%) für S1, S2, B1 und B2 der gemischten Beobachterpaare mit 400 msec zeitlicher Auflösung

Beobachterpaare	1/3	1/4	2/3	2/4	5/7	5/8	6/7	6/8	9/10
Durchgang S1	95,0	96,5	94,4	95,9	94,0	93,6	96,3	97,4	95,8
Durchgang S2	96,8	94,7	93,9	92,2	93,3	95,6	91,9	96,7	94,3
Durchgang B1	95,9	94,6	95,3	92,9	90,5	95,9	93,2	96,7	93,8
Durchgang B2	95,6	97,9	97,5	96,7	92,6	95,6	94,7	94,6	89,5

Tabelle 3
IOAs(%) beider Durchgänge in beiden Variablen mit 400 msec zeitlicher Auflösung

Beobachterpaare	1/2	1/3	1/4	2/3	2/4	3/4	5/6	5/7	5/8	6/7	6/8	7/8	9/10
Durchgang S1	96,2	95,0	96,5	94,4	95,9	96,4	95,7	94,0	93,6	96,3	97,4	96,6	95,8
Durchgang S2	92,9	96,8	94,7	93,9	92,2	97,3	96,8	93,3	95,6	91,9	96,7	92,2	94,3
Durchgang B1	95,1	95,9	94,6	95,3	92,9	93,3	95,3	90,5	95,9	93,2	96,7	93,1	93,8
Durchgang B2	95,0	95,6	97,9	97,5	96,7	96,9	95,6	92,6	95,6	94,7	94,6	92,4	89,5

5. Diskussion

Validität der Beobachter

Der Sprachdetektor ist ein Beobachter, der praktisch ohne zeitliche Verzögerung kodiert, kürzeste Sprechpausen und Sprechakte noch erfaßt und keine Aufmerksamkeitsschwankungen kennt. Seine Daten können daher als valide gelten, sorgfältige Einstellung der Schwelle vorausgesetzt. Demgegenüber ist die Kodierung der Beobachter fehlerhaft. Durch den Vergleich valider mit fehlerhaften Daten wird der Validitätsgrad dieser Daten direkt bestimmt. Zugleich wird die Reliabilität der fehlerhaften Daten konservativ geschätzt. Denn da die Reliabilität mindestens so hoch ist wie die Validität, kann der tatsächliche Reliabilitätswert keinesfalls niedriger ausfallen als die empirische Abweichung der fehlerhaften von einem idealen Beobachter. Die Kodierung entspricht einer Reaktionszeitaufgabe: Auf den plötzlichen Sprung einer unabhängigen Variablen soll mit möglichst geringer Verzögerung richtig reagiert werden. Der Betrag der Verzögerung hängt von verschiedenen Begleitumständen ab. Folgen die Reizsprünge wie hier in Serie und ist die Entscheidung binär, beträgt die Reaktionszeit in der Population der 20—40jährigen etwa 300 msec. Das gilt für ein durchschnittliches Inter-Stimulus-Intervall von zwei Sekunden. Wird der Reizabstand verkürzt, steigt die Reaktionszeit (RT) bis auf 400 msec (Welford, 1977).

Selbst unter Reizbedingungen, bei denen das Auftreten des bevorstehenden Reizsprungs näherungsweise vorhersagbar ist, wird die RT nicht weiter als bis auf etwa 150 msec verkürzt (Naeaetaenen & Merisalo, 1977). Außerdem schwankt auch unter konstanten Reizbedingungen die RT in Abhängigkeit von Fluktuationen der Aufmerksamkeit und der Reaktionsbereitschaft (Argyle & Cook, 1976, S. 52f.). Die Kodierungsentscheidungen folgen also dem Ablauf der tatsächlichen Ereignisse mit einer Latenz, die zwischen 150 und 400 msec streut. Diese Latenz ist einer der systematischen Fehler in den Beobachterdaten.

Wenn ein relativ stabiler Zustand durch kurze Fluktuationen unterbrochen wird, äußert sich die Latenz als „Trägheit". Das Ereignis, das eigentlich kodiert werden sollte, ist bereits vorüber, ehe die durch die Reaktionszeit verzögerte Kodierung realisiert werden konnte. Die Kodierung wird dann meistens nicht mehr abgegeben. Kurze Zustandsänderungen, wie Stockungen innerhalb einer Äußerung, die vom Sprachdetektor einwandfrei erfaßt werden, gehen somit verloren.

Eine weitere Fehlerquelle ergibt sich aus der Tatsache, daß die Übergänge zwischen den Sprechen-on- und Sprechen-off-Zuständen keine momentanen Ereignisse sind, sondern eine gewisse zeitliche Erstreckung haben. Es

ist ein Beobachtungsproblem, die Trennung zwischen den Zuständen jeweils am selben Zeitpunkt eines solchen Übergangs zu setzen.

Der Detektor, der die beobachtete Variable auf ihre physikalische Intensität, den Schallpegel, reduziert, trennt die beiden Zustände jeweils exakt an dem Punkt des Übergangs, der durch die Einstellung des Schallpegels vorgegeben ist. Den Beobachtern fehlt ein solches Kriterium, ihre Entscheidungen fallen daher vergleichsweise unscharf aus. Die durch die Latenz verzögerte Entscheidung wird also durch die Unschärfe des Kriteriums zusätzlich verrauscht.

Die Beobachtercharakteristika „Latenz" und „Trägheit" erzeugen einen systematischen Fehler, der in der Umgebung der Zustandsänderungen lokalisiert ist. Unschärfe des Entscheidungskriteriums, Aufmerksamkeitsschwankungen, Fehlentscheidungen und Blockaden der motorischen Reaktion kommen als unsystematische Fehler hinzu. Beide Fehlerarten sind in den Reliabilitätsmaßen prozentuale Übereinstimmung (PA) und Kappa konfundiert (Hartmann, 1977).

Üblicherweise versucht man den Fehleranteil durch Beobachtertraining zu reduzieren, wobei besonders die Fokussierung der Entscheidungsschwellen angestrebt wird (Baer, 1977). Erhöhung der Validität/Reliabilität wird dabei durch Verbesserung der Kodierung angestrebt. Dieses Verfahren hat den Nachteil, daß es mit einem hohen Arbeitsaufwand verbunden ist.

Wir variierten stattdessen die zeitliche Auflösung. Die Filterung des Datensatzes mit einem Zeitfenster induziert nachträglich eine Art von Trägheit, indem Zustandsänderungen subliminaler Dauer unterdrückt, werden. Das Zeitfenster wirkt auch zwischen den Kanälen und löscht deshalb Zeitdifferenzen zwischen den Beobachtern, die kürzer als das Zeitfenster sind. Es wird also durch die Filterung der systematische Fehler reduziert, der durch Trägheit und Latenz der Beobachter entsteht und damit steigt der Betrag der Übereinstimmung zwischen Beobachtern und Detektor (DOA), wie aus Abbildung 1 zu ersehen ist. Mit der Vergröberung der Auflösung steigt freilich auch der Verlust an Feinstruktur, also der Verlust „echter" Ereignisse. Es ist daher notwendig, einen Kompromißwert für die Auflösung zu bestimmen. Die Validität des Detektors wird dabei natürlich jeweils auf Ereignisse oberhalb der Fensterlänge beschränkt.

Die Veränderung der DOA, wie sie in Abbildung 1 dargestellt ist, gibt Aufschluß über Umfang und Verteilung der Fehler. Von 40 bis 80 msec ändert sich am Betrag der Übereinstimmung nahezu nichts. Das bedeutet, daß kaum Fehler in diesem Zeitbereich liegen. Der beschleunigte Anstieg der Kurve zwischen 120 und 400 msec zeigt, daß hier die Hauptmasse der Fehler lokalisiert ist. Es ist anzunehmen, daß die systematischen Fehler durch Latenz und Trägheit bei 400 msec annähernd eliminiert sind und die

verbleibende Inkongruenz zwischen Detektor- und Beobachterdaten zu Lasten gröberer unsystematischer Fehler geht. Fehler, die länger als eine Sekunde dauern, machen etwa ein Prozent der Gesamtzeit aus. Für Ereignisse oberhalb von 400 msec können die Beobachter zu 95% als valide gelten. Kappa liegt bereits bei feinster Auflösung, in der PA mit sämtlichen systematischen Fehlern belastet ist, bei allen Beobachtern oberhalb des gängigen Kriteriums von 0.6. Bei sukzessiv vergröberter Auflösung zeigt Kappa eine ähnliche Verlaufsform wie die Kurve des PA, nämlich anfängliche Beschleunigung und nachfolgende Stabilisierung. Der Abstand zwischen den beiden Kurven verringert sich absolut: Der Zufallsanteil $\bar{c}$ wächst um sechs Prozenteinheiten, während PA wegen der Unterdrückung der systematischen Fehler um vierzehn Prozenteinheiten ansteigt. Der Zähler des Quotienten Kappa nimmt also zu, während der Nenner gleichzeitig abnimmt. Daher steigt der Wert des ganzen Terms überproportional, verglichen mit dem Anstieg der prozentualen Übereinstimmung (PA), die nur von der Verringerung der systematischen Fehler profitiert.

Der durchschnittliche Zufallsanteil $\bar{c}$ verändert sich relativ wenig und erst bei höheren Auflösungen. Das bedeutet, daß bis zur Auflösung von etwa 400 msec das Gleichgewicht der Randverteilung zwischen on- und off-Zuständen nicht verändet wird. Die Wahl des Zeitfensters ist ein Spezialfall der Abwägung von Fehlerrisiken der 1. und 2. Art: Bei der Festsetzung von Testkriterien wird versucht, die Summe der Fehlerrisiken zu minimieren. Bei der Wahl des Zeitfensters ließen wir uns von folgender Überlegung leiten: Da im Kommunikationsprozeß relevante Ereignisse von relativ kurzer Dauer, wie Stockungen im Redefluß und „back-channel"-Verhalten auftreten, sollte die Auflösung nicht zu grob gewählt werden. In früheren Untersuchungen wurde meist mit Auflösungsgenauigkeiten zwischen 200 und 300 msec gearbeitet: Jaffe und Feldstein (1970) verwendeten eine Auflösung von 280 msec zur Erfassung von Sprechen, Vine (1971) wählte Einheiten von 250 msec für Blicken. Kendon (1967) arbeitete mit 500 msec für Sprechen und Blicken. Da die Kodierungslatenz bis 400 msec streut und der Zuwachs an Übereinstimmung von 280 bis 400 msec immerhin vier Prozent beträgt, wählten wir für die weitere statistische Analyse unserer Beobachtungsergebnisse eine Auflösung von 400 msec. Wir nehmen an, daß damit die systematischen Fehleranteile nahezu vollständig ausgetilgt werden und halten diesen Genauigkeitsgewinn für wichtiger als Verluste der Feinstruktur in diesem Zeitbereich.

Der Nachweis der monotonen Beziehung zwischen zeitlicher Auflösung und Validität/Reliabilität erlaubt die Wahl der Auflösung nach inhaltsadäquaten Gesichtspunkten. Mit welcher Auflösung schließlich gearbeitet wird, hängt in erster Linie davon ab, welche Auflösung die geplante Analyse verlangt und welchen Fehlerbetrag man in Kauf nehmen will.

Die hohe Übereinstimmung zwischen automatischem Detektor und menschlichen Beobachtern zeigt, daß die Kodierungsprozedur den Fähigkeiten der Beobachter hinreichend angepaßt ist.

Observer-Drift

In der Praxis interessiert an der Observer-Drift vor allem, ob sie zur Verschlechterung des Datenmaterials führt. Der Vergleich von DOA1 mit DOA2 (vgl. Tabelle 2a) weist die Unterschiede zwischen beiden Durchgängen als Zufallsschwankungen aus. Wenn Observer-Drift auftritt, muß sie besonders deutlich bei untrainierten Beobachtern zum Vorschein kommen, weil Beobachter mit längerer Praxis über stabilisierte Entscheidungskriterien verfügen. Auch der Vergleich zwischen erfahrenen und untrainierten Beobachtern (vgl. Tabelle 2b) ergab für beide beobachteten Variablen keine signifikanten Unterschiede. Offenbar wurde die Kriteriendefinition von untrainierten wie von trainierten Beobachtern konsistent angewendet. Das mag auf die Einfachheit der Kodierungsaufgabe zurückzuführen sein. Eine Konsequenz dieses Resultats besteht darin, daß sich unter den gegebenen Umständen ein längeres Beobachtertraining erübrigt.

Es ist bekannt, daß die Bestimmung des On-off-Übergangs des Blickverhaltens um so schwieriger wird, je genauer die Definition des „on-Zustandes" vorgegeben wird. Wir gaben den Beobachtern eine relativ unscharfe Definition — Blicken-on liegt vor, wenn der Beobachter dem Partner ins Gesicht blickt — und überließen somit die Präzisierung bewußt jedem einzelnen von ihnen. Damit gingen wir das Risiko verminderter Reliabilitäten und erhöhter Drift ein, meßbar als Absinken der IOAs, falls die Beobachter ihren Definitionsspielraum unterschiedlich und inkonsistent benutzt hätten.

Daß dies nicht der Fall war, werten wir als Bestätigung des Versuchs, beim Blickverhalten die Festlegung der Zustandskriterien dem Vorverständnis der Beobachter zu überlassen. Daraus darf natürlich nicht auf die Zulässigkeit dieser Vorgehensweise bei der Kodierung komplexerer Verhaltensweisen geschlossen werden.

Reliabilitätsvergleich zwischen Sprechen und Blicken

Niveau der Beobachterübereinstimmung (IOA). Die Übereinstimmung zwischen zwei Beobachtern mit extrem unterschiedlichen Kodierungsstilen kann durchaus niedriger ausfallen als die Übereinstimmung jedes dieser beiden Beobachter mit dem Detektor. Aus diesem Grund und weil für die Blickkodierung kein valider Vergleichswert verfügbar war, wurden neben der DOA auch die IOAs berechnet.

Die Beobachter ähneln sich in ihrem Kodierungsverhalten insofern, als ihnen die Fehler von Latenz und Trägheit gemeinsam sind. Dadurch wird die Übereinstimmung zwischen ihnen tendenziell erhöht. Andererseits begeht jeder Beobachter unsystematische Fehler, die mit hoher Wahrscheinlichkeit ungleichzeitig mit den unsystematischen Fehlern anderer Beobachter auftreten und sich somit als Nichtübereinstimmung addieren. Welcher dieser beiden Faktoren das Endresultat der IOA stärker beeinflußt, ist nicht von vornherein entscheidbar. Wie aus Tabelle 3 zu entnehmen ist, liegt die durchschnittliche IOA in beiden Durchgängen der Sprechkodierung bei 95% (Kappa = 0,89). Das entspricht ziemlich genau den DOA-Werten. Anscheinend heben sich die Wirkungen beider Faktoren auf.

Der Vergleich der IOA-Werte für Sprechen und Blicken gibt Aufschluß über den relativen Schwierigkeitsgrad beider Variablen. Es stellte sich heraus, daß die IOAs bei der Variablen „Sprechen" signifikant höher liegen als bei der Variablen „Blicken". Die relativ höhere Schwierigkeit der Variablen „Blicken" rührt vermutlich vom Fehlen einiger Hinweise her, die bei der Sprechbeobachtung gegeben sind: Der Beginn eines Sprechakts ist zeitlich weitgehend durch die Gesprächsstruktur determiniert und leichter antizipierbar als das Anblicken des Gesprächspartners. Anhand grammatikalischer Regeln ist auch das Ende des Sprechaktes in etwa vorhersagbar. Für die Darbietung des Beobachtungsmaterials hat dieses Ergebnis jedoch keine Konsequenz: Sowohl die durchschnittliche IOA bei B1 mit 94,3% als auch die einzelnen IOAs, die durchwegs über 90% liegen, sind akzeptabel. Die Bemerkungen über die Fehlerquellen wie Latenz und Trägheit treffen auch hier zu. Die Reliabilitätswerte entsprechen den in der Literatur angegebenen und liegen auf dem Niveau der Sprechkodierung (vgl. Tabelle 3). Offensichtlich eliminiert die Auflösung von 400 msec auch die systematischen Fehleranteile der Blickkodierung, während die groben unsystematischen Fehler nicht stärker als beim Sprechen ins Gewicht fallen.

Es erscheint daher nicht notwendig, für die Erfassung der theoretisch schwierigeren Variablen „Blicken" praktisch die Darbietungsgeschwindigkeit zu vermindern oder ein besonderes Beobachtertraining durchzuführen. Die Unterschiede zwischen den Beobachtern, die aus Tabelle 3 ersichtlich sind, können allerdings benutzt werden, um die einzelnen Beobachter zur Kodierung der Variablen einzuteilen, bei der sie die relativ besseren Leistungen erbracht haben.

Der Vergleich zwischen erstem und zweitem Durchgang innerhalb der Variablen bestätigt das Ergebnis der Drift-Untersuchung. Das Niveau der durch IOAs ausgedrückten Reliabilität ist stabil.

Korrelation der IOAs. Üblicherweise wird die Korrelation zwischen den Werten zweier Testdurchgänge als Index für die Retest-Reliabilität herangezogen. Am Messinstrument „Test" wird untersucht, ob es die Population

bei wiederholter Messung in gleicher Anordnung auf der Meßdimension abbildet. Für konsistente Niveauverschiebungen der gesamten Population, wie sie z.B. durch systematische Fehler verursacht werden können, ist dieser Index unempfindlich. Die Niveaustabilität der Meßwerte muß demnach gesondert überprüft werden, wenn sie in die Reliabilitätsaussage einbezogen wird. Das wurde bereits im vorherigen Abschnitt demonstriert.

Die gebräuchliche Logik der Korrelation als Reliabilitätsindex kann jedoch im vorliegenden Fall nicht übernommen werden. Die Fragestellung zielt nämlich nicht auf die Zuverlässigkeit des Tests — hier: des Beobachtungsmaterials — als Meßinstrument für die Population der Beobachter, sondern umgekehrt auf die Zuverlässigkeit oder Konstanz der Beobachter als Meßinstrumente für das konstante Material (vgl. Baer, 1977). Falls die Beobachter reliabel sind, müssen ihre Testwerte intraindividuell konstant bleiben. Sind sie darüber hinaus auch valide, liefern sie interindividuell identische Werte.

Werden die Meßwerte von unsystematischen, nicht beobachterspezifischen Fehlern überlagert, geht die Retest-Korrelation gegen Null. Konstantes Niveau und niedrige Korrelation zusammen belegen dann gerade die Güte der Meßinstrumente „Beobachter".

Die Niveaukonstanz über die einzelnen Durchgänge wurde bereits nachgewiesen. Die Korrelation der Sprechkodierung zwischen beiden Durchgängen beträgt − 0.02. Das bedeutet, daß Unterschiede in den Kodierungsleistungen der Beobachterpaare im wesentlichen auf unsystematische und damit von Durchgang zu Durchgang variierende Fehler zurückgehen. Diese Fehler können in ihrem Absolutbetrag vernachlässigt werden. Auch die Blickdaten beider Durchgänge korrelieren mit + 0.36 nicht signifikant. Die Unterschiede zwischen den Beobachterpaaren sind also auch bei der Blickerfassung über die Zeit hinweg nicht stabil und können daher dem Einfluß unsystematischer Fehler zugeschrieben werden.

Aufgrund des absolut hohen Niveaus der IOAs und wegen der korrelativ nachgewiesenen Abwesenheit systematischer Verzerrungen können die Kodierungen der beiden Variablen „Sprechen" und „Blicken" für praktische Zwecke als gleichwertig angesehen und verarbeitet werden.

Summary

The reliability of the binary coding of speech and gaze was examined as a function of the temporal resolution of measurement. The reliability was calculated as the agreement between human observers. In addition, the validity of the speech coding could be determined from the agreement between the human observers and an electronic speech detector.

The following results were obtained:

1. The validity/reliability of the speech coding is a monotonic nonlinear function of the temporal resolution. Those systematic errors due to observers' latency and inertia are effectively suppressed with a resolution of 400 msec.
2. No evidence of observer-drift was found for either speech or gaze coding. No significant difference in observation accuracy was found between trained and untrained observers.
3. Speech coding was found to be more reliable than gaze coding. This may be interpreted as a result of the qualitative difference between the variables. For most practical purposes the accuracy of coding of the two variables can be regarded as equivalent.

Résumé

La validité et la fidélité des codifications des activités de parole ont été examinées, pour un groupe d'observateurs, en fonction de la durée minimale des observations retenues. La validité a été estimée par la concordance entre observateurs d'une part et détecteur électronique d'autre part. La fidélité des codifications de la parole et des regards a été estimée par l'accord des observateurs entre eux. Les principaux résultats sont les suivants:

1. La validité/fidélité des codifications de la parole est une fonction monotone et non-linéaire de la durée minimale retenue. Les fautes systématiques résultent, pour l'essentiel, de la latence et de l'inertie des observateurs peuvent être presque entièrement supprimée en fixant la durée minimale des activités codifiées à 400 msec.
2. Ni la codification de la parole, ni celle du regard ne révèle une instabilité de jugement des observateurs (observerdrift). Les observateurs entraînés et non-entraînés ne diffèrent pas de façon significative.
3. La codification de la parole a une fidélité un peu plus élevée que celle du comportement visuel. La raison de cette différence est à chercher dans la nature des deux variables. Dans la pratique, on peut cependant se permettre de négliger cette différence.

Literatur

Argyle, M. & Cook, M.: Gaze and mutual gaze. Cambridge, London, New York, Melbourne: Cambridge University Press, 1976.

Baer, D. M.: Reviewer's comment: just because it's reliable doesn't mean that you can use it. Journal of Applied Behavior Analysis, 1977, 10, 117—119.

Clarke, A., Wagner, H., Rinck, P. & Ellgring, J. H.: A system for computer aided observation and recording of social behaviour. Unveröffentlichtes Manuskript, 1979.
v. Cranach, M. & Ellgring, J. H.: The perception of looking behavior. In: M. von Cranach & I. Vine (Hrsg.), Social communication and movement. London: Academic Press, 1973.
Hartmann, D. P.: Considerations in the choice of interobserver reliability estimates. Journal of Applied Behaviour Analysis, 1977, **10**, 103—116.
Jaffe, J. & Feldstein, S.: Rhythms of Dialogue. New York: Academic Press, 1970.
Kendon, A.: Some functions of gaze direction in social interaction. Acta Psychologica, 1967, **26**, 1—47.
Naeaetaenen, R. & Merisalo, A.: Expectancy and preparation in simple reaction time. In: S. Dornic (Hrsg.), Attention and performance VI. New Jersey: Lawrence Erlbaum, 1977.
Vine, I.: Judgement of direction of gaze - an interpretation of discrepant results. British Journal of Social and Clinical Psychology, 1971, **10**, 320—331.
Welford, A. T.: Serial reaction times, continuity of task, single channel effects, and age. In: S. Dornic (Hrsg.), Attention and performance VI. New Jersey: Lawrence Erlbaum, 1977.

Anschrift des Verfassers:

Dipl.-Psych. H. Wagner
Max-Planck-Institut für Psychiatrie
Kraepelinstraße 10
8 München 40