

TESTZENTRALE

Die Testzentrale ist eine zentrale Stelle für die Beschaffung von psychodiagnostischen Verfahren und psychologischer Fachliteratur. Aufgrund enger Zusammenarbeit mit in- und ausländischen Verlagen und allen wichtigen Testzentren des Auslandes verfügt sie über ein großes Lager aller gebräuchlicher Testverfahren und der dazugehörigen Literatur. Sie hat sich seit mehr als 20 Jahren auf dieses Gebiet spezialisiert und kann Ihnen deshalb auf schnellstem Wege Ihr Arbeitsmaterial besorgen.

Beispiele aus unserem Lieferprogramm: Göttinger Formreproduktionstest zur Diagnose der Hirnschädigung – The Hunt Minnesota Test for Organic Brain Damage – Hand-Dominanz-Test zur Diagnose der Händigkeit – Hamburger Neurotizismus- und Extraversionsskala für Kinder und Jugendliche – Diagnostikum für Cerebralschädigung – MMPI – Benton-Test – Münchner Alkoholismus Test – TÜLUC-Neuropsychologische Testbatterie – Ishihara Test for Colour Blindness – Mehrdimensionale Schmerzskala – Erlanger Depressions-Skalen

TESTZENTRALE

des Berufsverbandes
Deutscher Psychologen
Daimlerstr. 40 · 7000 Stuttgart 50

APPARATEZENTRUM

Apparate für Psychodiagnostik und Therapie:

Biofeedback-Geräte

zum individuellen Entspannungs- training mit meßtechnischer Erfolgskontrolle (PGR-, EEG-, EMG-, HR-, Atem- und Temperatur- Feedback-Geräte)

Desensibilisierung

Anlage zur Darbietung in Bild und Ton

Enuresis-Alarmgerät

zur Verhaltenstherapie des Bett- nässens

Haptometron

zur Verhaltenstherapie des Stotterns

Pulsmesser

zur Überwachung des Pulsschlags

Audiometer

zur Überprüfung der Hörfähigkeit

Visiontester

zur Überprüfung der Sehfähigkeit

Außerdem umfaßt unser Liefer- programm:

Bildschirm-Testgeräte, Determinationsgeräte, Dynamographen, Flimmergeräte, Mikroprozessor- Steueranlagen, Motorik-Testgeräte, Polygraphen, Reaktionsgeräte, Tachistoskope, Stoppuhren u.v.a.

APPARATEZENTRUM

Dr. C. J. Hogrefe
Postfach 414 · Tel. 0551/5 4044
Rohnsweg 25 · 3400 Göttingen

Systematische Verhaltensbeobachtung von Aufmerksamkeit im Unterricht: Zur Prüfung von Objektivität und Zuverlässigkeit

Klaus Jürgen Ehrhardt, Peter Findeisen,
Gloria Marinello & Hiltrud Reinartz-Wenzel

1. Einführung

Der Mangel an brauchbaren Testverfahren zur Messung von Aufmerksamkeit bei Grundschulkindern (Bartenwerfer 1964; Brickenkamp 1975) war der Anlaß zur Entwicklung einer Methode zur systematischen Verhaltensbeobachtung (SVB) von Aufmerksamkeit während des Unterrichts. Die SVB liefert „L-Daten“ (Cattell 1973, S. 61) wie auch die anderen in der Schulpraxis üblichen Fremdeinstufungen (Übersicht bei Langhorst 1974; Heller und Nickel 1978). Die Güte von L-Daten ist bekanntlich geringer als die von Testdaten. Bei der SVB wird versucht, subjektive Momente bei den Beurteilern möglichst auszuschalten, indem das zu beobachtende Verhalten präzisiert und das Vorgehen bei der Beobachtung genau festgelegt und trainiert wird (Übersicht bei Mees 1977).

In den ersten Veröffentlichungen der Arbeitsgruppe wurde auf dem Hintergrund ähnlicher Verfahren in der Literatur über die Entwicklung der Methode mit Auswahl der Verhaltenskriterien für Aufmerksamkeit, die Konstruktion von Kategorien und ihre Verknüpfung, das Vorgehen bei der Beobachtung und das Beobachtertraining berichtet (Ehrhardt, Haack, Klich, Marinello, Plassmann, Reinartz-Wenzel & Winzer 1980; Ehrhardt, Findeisen, Marinello & Reinartz-Wenzel 1981, Reinartz 1979; Marinello 1981). Das Ergebnis ist ein System mit drei Kategorien (Blickrichtung, Körperhaltung, Tätigkeit), nach dem das Verhalten einzelner Kinder fortlaufend im 10-Sekunden-Zeittakt eingestuft wird.

Die Fragestellung der vorliegenden Untersuchung ist, ob mit dieser Methode ausreichend objektive und zuverlässige Daten erhoben werden können. Dazu ist zuerst zu fragen, ob die Gütekriterien der klassischen Testtheorie (z.B. Lienert 1969) auf die SVB anwendbar sind. Das ist nach Fasnacht (1979) nicht ganz unproblematisch. Aber solange keine eigenen Gütekriterien für die SVB zur Verfügung stehen, scheint es angemessen, in Analogie vorzugehen und die Beobachterübereinstimmung als Maß für die Objektivität und Testhalbierungs- und Wiederholungszuverlässigkeit als Maße für die Reliabilität zu verwenden.

Bei der Bestimmung der Reliabilität der SVB ergaben sich keine prinzipiellen Schwierigkeiten. Dagegen stellte sich bei der Berechnung der Beobachterübereinstimmung das Problem, welches der verschiedenen in der Literatur angegebenen Verfahren als statistisch aussagekräftig angesehen werden kann. Dieses Problem stellt sich bei

jeder Form der SVB und ist für den mathematisch nicht versierten Leser weder bei M e e s noch in den Originalarbeiten nachvollziehbar. Deshalb wird es in dieser Arbeit ausführlich behandelt, und es wird auch ein Berechnungsverfahren vorgestellt, das bei beliebigen Methoden der SVB angewendet werden kann zur zufallskritischen Prüfung der Beobachterübereinstimmung.

Zur Prüfung der Beobachterübereinstimmung wird folgendermaßen vorgegangen: Zwei Beobachter registrieren simultan und unabhängig voneinander das Verhalten des gleichen Kindes. Aus mehreren solchen Beobachtungen erhält man mehrere „Protokolpaare“. Für die anschließende Berechnung der Übereinstimmung der Protokolle werden verschiedene Verfahren angegeben, die am gleichen Datensatz zu erheblich unterschiedlichen Ergebnissen führen:

(1) Bei der *Summenübereinstimmung* werden die Auftretenshäufigkeiten bzw. Punktwerte in jedem Protokoll summiert und die Summenwerte der Protokolpaare miteinander korreliert. Die Summenübereinstimmung überschätzt die tatsächliche Übereinstimmung, denn die Beobachter können ja zu einer ähnlichen oder sogar der gleichen Zahl von Signierungen kommen, ohne daß die Zeitpunkte der Eintragungen übereinstimmen. Nur die platzweise Überprüfung gewährleistet, daß die Beobachter in der Registrierung tatsächlicher Verhaltensereignisse übereinstimmen.

(2) Die deskriptive *Platz-zu-Platz-Übereinstimmung* (PPO) prüft die Übereinstimmung von Protokolpaaren in jedem einzelnen Beobachtungsintervall (Platz). Sie wurde von B u i j o u , Peterson & A ult (1968) eingeführt und ist in der Verhaltensbeobachtung das am häufigsten verwendete Maß (s. Tab. 2). Eine Übereinstimmung zwischen zwei Beobachtern von 90 % der Plätze wird allgemein als sehr gut bezeichnet (E l s o n 1963, zit. nach F a b n a c h t 1979, S. 29). Diese Bewertung ist nicht allgemein sinnvoll, notwendig ist eine zufallskritische Absicherung der PPO. Auch S c h o u t e n (1980) gibt nur naheliegende Schätzformeln, aber keinen teststatistischen Ansatz für Übereinstimmungswahrscheinlichkeiten an.

Die zufällig zu erwartende Übereinstimmung ist abhängig von der Zahl der verfügbaren Einstuflmöglichkeiten und deren Auftretenshäufigkeiten. Bei einem einzigen alternativen Kriterium (und gleich häufiger Signierung beider Möglichkeiten) beträgt die zufällige Übereinstimmung 50 %, bei vier Abstufungen wie in dem vorliegenden Kategoriensystem dagegen 25 %. Wird aber eine Kategorie nur selten signiert, dann wird die PPO erheblich überschätzt. Das kritische Verhalten sei z.B. „Sprechen mit Banknachbar“ und komme in einem Zeitraum von 20 Intervallen nur einmal vor. Dieses eine Auftreten kann von den Beobachtern in unterschiedlichen Intervallen (Plätzen) protokolliert werden, und trotzdem ist die Übereinstimmung 90 % aufgrund der 18 anderen Plätze, in denen beide Beobachter übereinstimmend das Fehlen dieses Verhaltens vermerkt haben. Um konservativ zu schätzen, schlägt M e e s (1977, S. 48) in diesen Fällen vor, nur für Plätze mit Eintragungen die Berechnung vorzunehmen („signierte Platzübereinstimmung“), die allerdings auch nicht unabhängig von der Auftretenshäufigkeit ist.

(3) Das voranstehende Beispiel macht deutlich, daß man für die *zufallskritische Überprüfung der Beobachterübereinstimmung* nicht nur vom Konzept der Summen- zu demjenigen der Platz-zu-Platz-Übereinstimmung übergehen muß, sondern darüberhinaus, auf einer dritten logischen Stufe, die durch die *Auftretenshäufigkeit bedingte* PPO als Übereinstimmungsmaß zu definieren hat (vgl. hierzu L i g h t 1971). Man stelle sich einmal vor, daß im Beispiel das „Sprechen mit dem Banknachbarn“ von beiden Beobachtern im gleichen Zeitintervall registriert wurde. Dann wird also eine hundertprozentige PPO angetroffen. Jedoch die Wahrscheinlichkeit dieses Ereignisses betrüge selbst dann, wenn jeder Beobachter das Intervall mit dem kritischen Verhalten völlig willkürlich aus den 20 Zeitintervallen ausgewählt hätte, immerhin noch $20 : 400 = .05$! Man hat also den Wert der PPO von 100 % mit den vorliegenden Informationen über die Auftretenshäufigkeiten (das kritische Verhalten kommt bei beiden Beobachtern genau einmal in 20 Intervallen vor) relativierend zu wichten.

Hier greift nun die klassische Teststatistik ein. Es ist zu entscheiden, ob das angetroffene Ergebnis einer 100%igen PPO die Annahme von „aneinander vorbei urteilenden“ Beobachtern widerlegt, d.h. (in der üblichen verkürzten Formulierung): „Ist die gefundene Übereinstimmung signifikant?“ Da unter der in Frage stehenden Annahme der Wert .05 die Wahrscheinlichkeit des ein-

getretenen Ereignisses ist, muß nun entschieden werden, ob .05 ein sinnvolles „Signifikanzniveau“ ist. Wird diese Frage – gemäß der teststatistischen Tradition – bejaht, so deuten die Daten des Beispiels (aber sozusagen ganz knapp!) auf „übereinstimmende“ Beobachter hin.

Damit ist jedoch nur gemeint – und diese Anmerkung ist wichtig –, daß die folgende statistische Nullhypothese zurückgewiesen wird: Bei dem durch die Auftretenshäufigkeiten von 1 : 20 festgelegten „bedingten Zufallsexperiment“ (das in der Auswahl je eines kritischen Zeitintervalls durch die zwei Beobachter besteht) ist jeder der beiden Beobachter ganz ohne Bezug zum anderen vorgegangen. Die logische Verneinung dieser Nullhypothese der „Nicht-Übereinstimmung“ reicht aber möglicherweise nicht aus, um ein psychologisches Konzept der „Übereinstimmung“ abzudecken. Vor einer Überbewertung eines „statistisch signifikanten Ausmaßes an Übereinstimmung“ zwischen zwei Beobachtern muß also gewarnt werden.

Bei der Auswahl eines speziellen als „Statistik“ verwendbaren Maßes gibt es schon darum, weil die voranstehenden Argumente in Einzelheiten modifizierbar sind, mehrere Möglichkeiten. Wir haben hier die von uns mit κ_0 bezeichnete standardisierte Version des Index κ , der von L i g h t (1971) auch als deskriptiver Skore gefertigt wird, verwendet. κ_0 kann als z-Wert aufgefaßt werden, sein Signifikanzniveau ist in der entsprechenden Tafel abzulesen. (Hier geht eine Voraussetzung ein, die für das Beispiel folgendes bedeuten würde: Nicht als „Auftretensquote“, sondern als Wahrscheinlichkeit dafür, daß das „Sprechen mit dem Banknachbarn“ in einem betrachteten Beobachtungsintervall registriert wird, wäre der Wert 1 : 20 ein gegebener Parameter des Beobachterverhaltens.)

2. Methodik

Die Untersuchung fand Anfang 1979 an zwei öffentlichen Grundschulen mit städtischem Einzugsgebiet statt, wo Schulleitung und betroffene Lehrer an der Fragestellung interessiert und zur Mitarbeit bereit waren. 60 Kinder des zweiten Schuljahrs wurden, teils mehrfach, beobachtet. Die ersten 4 Wochen dienten dem Beobachtertraining während des üblichen Unterrichts; in den letzten zwei Wochen (Testphase) war der Unterricht standardisiert.

Vorgehen bei der Beobachtung. An der Untersuchung nahmen zwei Beobachter teil. Sie beobachteten anfangs ein Kind, später ohne Schwierigkeiten gleichzeitig zwei Kinder. Die Beobachter setzten sich etwa zwei Meter voneinander entfernt so an den Rand des Klassenraums, daß sie die beiden Kinder bequem im Auge hatten. Über Ohrclip erhielten sie vom Tonband alle 10 Sek. ein Zeichen, bei dem sie sofort auf dem Protokollvordruck ihre Einstufung vorzunehmen hatten. Den 10-Sek.-Abschnitt bezeichnen wir mit *Beobachtungsintervall*, die Gesamtdauer der Beobachtung eines Kindes als *Beobachtungszeitraum*.

Grundlage für die Einstufung war das Kategoriensystem in Tabelle 1. Das Merkmal „Aufmerksamkeit im Unterricht“ war durch die Kategorien Blickrichtung, Körperhaltung und -ausdruck und Tätigkeit definiert. Welche konkreten Verhaltensweisen vorliegen müssen, damit eine Kategorie erfüllt ist, erläutert eine Liste von Beispielen und Regelungen für Sonderfälle (siehe E h r h a r d t et al. 1981). Für jedes Beobachtungsintervall erhielt das Kind entsprechend der Anzahl der erfüllten Kategorien zwischen 0 und 3 Punkten. Es kam vor, daß ein Kind während eines Beobachtungsintervalls

nicht eingestuft werden konnte (Kind verdeckt, Beobachter unaufmerksam). Dieses (seltene) Ereignis wurde ebenso wie kurze Unterbrechungen im Unterricht gesondert protokolliert.

Tabelle 1:
Kategorien für Aufmerksamkeit im Unterricht

Kategorie	erfüllt	nicht erfüllt
(a) Blickrichtung	blickt zum Unterrichtsmittelpunkt (UM)	blickt woanders hin
(b) Körperhaltung u. Körpераusdruck	ausgerichtet auf UM und angespannt	abgewandt erschlafft
(c) Tätigkeit	übt die für die Aufgabe notwendige Tätigkeit aus	tut nebenher etwas anderes

Kodierung

3 Punkte: alle 3 Kategorien sind über das Beobachtungsintervall hinweg (volle 10 Sek. abzüglich ca. 1 Sek. für Protokollierung) erfüllt

2 Punkte: 2 Kategorien über ganzes Beobachtungsintervall erfüllt

1 Punkt: 1 Kategorie über das ganze Beobachtungsintervall erfüllt

0 Punkte: keine Kategorie erfüllt oder tut etwas anderes oder „träumt“

Sonderfälle

- vom Verhalten nicht zu entscheiden, ob Kategorie erfüllt oder nicht
- Kind momentan nicht zu beobachten
- Beobachter momentan unaufmerksam
- Pause im Unterricht (Zeiten, wo Aufmerksamkeit nicht gefordert)

Das *Beobachtertraining* dauerte vier Wochen und fand nicht am Videogerät, sondern live im Unterricht statt. In den ersten zwei Wochen machten sich die Beobachter mit der Beobachtungssituation, der Anwendung des Kategoriensystems und der Kodierung im Zeittakt vertraut. Beide beobachteten stets simultan das gleiche Kind. Im Anschluß an jeden Beobachtungszeitraum haben sie die beiden Protokolle miteinander verglichen und unterschiedliche Einstufungen diskutiert. Ab der dritten Woche wurden gleichzeitig zwei Kinder beobachtet und die Beobachterübereinstimmung fortlaufend berechnet.

Die fünfte und sechste Woche waren die *Testphase*. Es wurden weiterhin zwei Kinder simultan beobachtet, die Protokolle aber nicht mehr verglichen. Für die Trainingszeit und die Prüfung der Beobachterübereinstimmung sind unterschiedliche Unterrichtssituationen kein Nachteil, denn das Spektrum der zu erwartenden kindlichen Verhaltensweisen ist dann größer. In der Testphase haben wir dagegen versucht, die *Rahmenbedingungen weitgehend zu standardisieren*. Die Beobachtung fand nur dienstags, mittwochs und donnerstags in der zweiten und vierten Schulstunde statt. Während dieser Zeit hat der Lehrer über 14 Tage stets den gleichen Unterricht gehalten, nämlich gemeinsames Vorlesen, wobei abwechselnd ein Kind vorlas, und die anderen Kinder den Text mit dem Finger zu verfolgen hatten. Alle 22 Kinder einer Klasse (15 Jungen, 7 Mädchen) wurden in zufällig festgelegter Reihenfolge im Abstand von genau einer Woche zweimal beobachtet. Die Stichprobe ist nicht besonders groß. Es kam uns aber darauf an, die Situation möglichst konstant zu halten und nicht durch Hinzunahme einer weiteren Klasse die Bedingungen heterogener zu gestalten.

Maße für Beobachterübereinstimmung. Zur Ermöglichung der formalen Definition solcher Maße stellen wir zunächst die von Light (1971) benutzte Terminologie bereit: Sei C die Anzahl möglicher Einstufungen bzw. Punktzahlen, das ist die um 1 vermehrte Anzahl der Kategorien (also hier $C = 4$): Für $1 < i < C$ bezeichnen wir als i -te Einstufung den Punktwert $i-1$, der ausdrückt, daß ($i-1$) Kategorien als erfüllt beurteilt wurden. Sei n die Anzahl der Beobachtungsintervalle (Plätze), die nach erfolgter Beobachtung eines Kindes für die Berechnung eines Übereinstimmungswertes herangezogen werden.¹⁾

Mit n_{ij} soll die Anzahl derjenigen Beobachtungsintervalle bezeichnet werden, in denen beide Beobachter (übereinstimmend) die i -te Einstufung gewählt haben, mit n_{i+} (bzw. n_{++}) wird die Anzahl solcher Intervalle bezeichnet, in denen der erste (bzw. zweite) Beobachter die i -te Stufe signiert hat (gleichgültig, wie der jeweils andere Beobachter sich entschieden hat).

Die Summe $S: = \sum_{i=1}^C n_{ij}$ ist die absolute Zahl von Übereinstimmungen. Dabei ist die Zählweise insofern „konservativ“, als Abweichungen von einem Punkt genauso als Nicht-Übereinstimmung gewertet werden wie Abweichungen von drei Punkten.

¹⁾ Um der statistischen Abhängigkeit der Meßwerte untereinander entgegenzuwirken, wurde nur jeder vierte Beobachtungswert für die Berechnung verwendet.

Ausgehend vom „Elementarmaß“ S, können jetzt anspruchsvollere Maße angegeben werden:

(1) Platz-zu-Platz-Übereinstimmung in Prozent nach der Formel von Bijou et al. (1968, nach Mees 1977, S. 47):

$$\text{PPU\%} = \frac{\text{Übereinstimmungen}}{\text{Übereinstimmungen} + \text{Nicht-Übereinstimmungen}} \cdot 100 = \frac{100}{n} S$$

(2) Für die zufallskritische Überprüfung der Übereinstimmung von zwei Beobachtern gingen wir gemäß dem Ansatz von Light (1971) vor, anders als Fricke (1972) und Lindner (1980), die mit dem U-Koeffizienten ein nicht-bedingtes Übereinstimmungsmaß bereitstellen.

Bei Light (1971, S. 367) wird der auf Cohen zurückgehende Skore vorge-

$$\kappa = \frac{\frac{C}{n} - P}{n(1-P)}$$

schlagen. Wir haben hier Light's Schreibweise abgeändert und

$$\frac{\sum_{i=1}^C n_{ii} - nP}{n^2} \quad \text{gleich } P \text{ gesetzt.}$$

Unter der in der Einführung genau erläuterten statistischen Nullhypothese der Nicht-Übereinstimmung (die Beobachter urteilen „stochastisch unabhängig“ voneinander) kann P als die (aus den Daten geschätzte) Wahrscheinlichkeit dafür gedeutet werden, daß zu einem festen Zeitpunkt beide Beobachter das gleiche Urteil abgeben: dann ist

$$E(S) = nP$$

der Erwartungswert und

$$\sigma(S) = \sqrt{nP(1-P)}$$

die Standardabweichung von S. Somit kann

$$\kappa = \sqrt{\frac{P}{n(1-P)}} \cdot \frac{S - E(S)}{\sigma(S)}$$

geschrieben werden. Gegenüber κ bevorzugen wir nun den modifizierten Skore

$$\kappa_O = \sqrt{n} \cdot \sqrt{\frac{1-P}{P}} \cdot \kappa = \frac{S - E(S)}{\sigma(S)}$$

der als standardisierte Form sowohl von S (bzw. von PPU%) als auch von κ aufgefaßt werden kann:

$$E(\kappa_O) = 0; \sigma(\kappa_O) = 1.$$

Die Größe S ist eine mit den Parametern n und P binomialverteilte, also κ_O eine (annäherungsweise) standard-normalverteilte Zufallsvariable. Diese Aussage gilt, hieran sei noch einmal erinnert, unter einer „bedingten“ Nullhypothese, die die Werte n, n_{ij} und n_{+i} ($1 < i < C$) als „gegeben“ voraussetzt.

In Tabelle 2 sind die verschiedenen Maße am Beispiel einer zweiminütigen Beobachtung eines Kindes zusammengestellt und die Berechnung von κ_O erläutert.

Tabelle 2:
Beispiel einer zweiminütigen Beobachtung eines Kindes mit Berechnung der verschiedenen Maße.

Kind: y x		Beobachtungsintervalle	Beobachtungswerte		Übereinstimmung
Beobachtung			Beobachter 1	Beobachter 2	
1. Minute	1		0	0	+
	2		0	1	
	3		3	3	+
	4		3	3	+
	5		2	3	
	6		3	3	+
2. Minute	7		1	2	
	8		0	0	+
	9		3	2	
	10		3	3	+
	11		3	3	+
	12		3	3	+
		Σ	24	26	8
Leistungsmaß:		$\frac{\text{Punktsumme}}{\text{Intervalle} \cdot 3} \cdot 100$	67 %	72 %	
Übereinstimmungsmaße:					
deskriptiv PPU = $\frac{\text{Übereinstimmung}(S)}{\text{Intervalle}} \cdot 100$			66 %		
zufallskritisch $\kappa_O = \frac{S - E(S)}{\sigma(S)}$			1.86		
			(p < .04)		

Berechnung von κ_0 :

Anzahl der möglichen Einstufungen im Kategoriensystem: C = 4

Anzahl der Beobachtungsintervalle: N = 12

Häufigkeit der i-ten Einstufung:

Punkt-wert (i-1)	durch den Beobachter 1 i	durch den Beobachter 2 n_{+i}	durch beide Beobachter simultan n_{ii}	beide $n_{i+} \cdot n_{+i}$
0	1	3	2	6
1	2	1	0	1
2	3	1	0	2
3	4	7	6	49
Summe	12	12	S = 8	58

$$P = \frac{1}{n^2} \cdot \sum_{i=1}^4 n_{i+} \cdot n_{+i} = \frac{58}{144} = 0.4028$$

$$E(S) = nP = 12 \cdot 0.4028 = 4.8333$$

$$\sigma(S) = \sqrt{nP(1-P)} = \sqrt{12 \cdot 0.4028 \cdot 0.5972} = 1.699$$

$$\underline{\kappa_0} = \frac{S - E(S)}{\sigma(S)} = \frac{8 - 4.8333}{1.699} = 1.8629$$

Die Fläche unter der Standardnormalkurve zwischen $z = 1.863$ und ∞ ist kleiner als 0.04. Es ergibt sich Signifikanz auf dem 4%-Niveau.

3. Ergebnisse

Beobachterübereinstimmung. Für jedes Protokollpaar wurde die Übereinstimmung als PPÜ% und κ_0 berechnet. Die durchschnittliche Übereinstimmung der beiden Beobachter in der Trainingsphase (3. und 4. Woche) und der Testphase (5. und 6. Woche) sind in Tab. 3 zusammengestellt. Während κ_0 von 2.19 auf 3.87 monoton ansteigt, erfolgt die Zunahme von PPÜ% nicht proportional. Korrelation der Einzelwerte ergibt keinen Zusammenhang. Daß ein solcher auch nicht erwartet werden kann, zeigt sich bei näherer Betrachtung der die Werte definierenden Formeln.

In der 5. und 6. Woche liegt die durchschnittliche PPÜ zwischen 75 % und 76 %. Die Summenübereinstimmung für den gleichen Zeitraum ist in Tab. 4 angegeben. Sie liegt mit .96 und .98 erwartungsgemäß höher. Aus den Einzelwerten für κ_0 kann für jedes Protokollpaar entschieden werden, ob die Übereinstimmung überzufällig hoch ist. Als Kriterium hatten wir $p < .05$ gesetzt. In der 5. und 6. Woche stimmten von 22 Beobachtungen 21 in überzufälligem Ausmaß überein.

Tabelle 3:

Beobachterübereinstimmung bei zunehmender Beobachtungsübung, berechnet als prozentuale Platz-zu-Platz-Übereinstimmung und zufallskritisch als κ_0 (Durchschnittswerte).
(gemeinsame Arbeit, 2. Schuljahr)

	Woche	Übereinstimmung		Anzahl	
		PPÜ%	$\bar{\kappa}_0$	beobachtete Kinder	davon „signifikante“ Übereinstimmung (κ_0)
Training	3.	67,9	2,19	17	11
		84,1	2,33	11	6
	4.	79,9	3,17	10	9
Test	5.	76,1	3,60	22	21
	6.	75,7	3,87	22	21

In der Entwicklung des hier verwendeten Kategoriensystems hatten wir auch solche mit 2 und 3 Abstufungen erprobt (Ehrhardt et al. 1980). Zum Vergleich geben wir hier die durchschnittlichen PPÜ-Werte an:

2 Stufen: PPÜ% 87,0 (n = 9); 85,5 (n = 10)

3 Stufen: PPÜ% 82,0 (n = 4); 90,5 (n = 22)

Sie sind den PPÜ-Werten in Tabelle 3 aber nicht direkt vergleichbar, da die Beobachtungssituation in einigen Punkten unterschiedlich war. Die Höhe der Werte erklärt sich aus der zufallsbedingten Übereinstimmung, die bei zwei Stufen 50 %, bei drei Stufen 33 % beträgt. Auch aus diesem Grund sind die numerischen Werte nicht direkt vergleichbar.

Ein wesentlicher Nachteil der Kategoriensysteme mit nur 2 oder 3 Abstufungen liegt darin, daß κ_0 -Werte für zahlreiche Protokollpaare nicht berechnet werden können: Vor allem bei nur 2 Abstufungen ist die Einstufung „aufmerksam“ relativ häufig, und (besonders bei Verwendung nur jeden vierten Beobachtungswertes) kann es vorkommen, daß beide Beobachter stets nur ein und dieselbe Einstufung wählten. Dann ist κ_0 nicht berechenbar, da der Nenner in der κ_0 definierenden Formel gleich Null wird.

Reliabilität. Die Reliabilität des Verfahrens wurde in der 5. und 6. Beobachtungswoche nach der Halbierungsmethode und durch Beobachtungswiederholung bestimmt.

Bei der Halbierungsmethode wurde jedes Protokoll in zwei gleiche Hälften geteilt durch Aufsummieren der Punktewerte der geraden und ungeraden Beobachtungsintervalle (odd-even). Die Produkt-Moment-Korrelationen der Halbwerte betrugen (für jeden Beobachter getrennt) $r = .96$ und $r = .97$.

Für die Bestimmung der Retest-Reliabilität wurden alle 22 Kinder einer Klasse unter weitgehend standardisierten Unterrichtsbedingungen im Abstand von genau 8 Tagen zweimal beobachtet. Der Beobachtungszeitraum betrug immer 20 Minuten

Tabelle 4:

Wiederholte Beobachtung der Kinder einer zweiten Klasse ($N = 22$) unter standardisierten Unterrichtsbedingungen im Abstand von acht Tagen. Aufgelistet sind die Aufmerksamkeitswerte für jedes Kind, getrennt nach Beobachtern und Zeitpunkten. Berechnet sind die Korrelationen (Pearson) der simultanen Beobachtungswerte (Summenübereinstimmung) und zwischen Test- und Retestwert für jeden Beobachter getrennt und für beide gemeinsam (Retest-Reliabilität).

	TEST			RETEST			Σ	
	VPN Nr.	Beobachter 1	Beobachter 2	VPN Nr.	Beobachter 1	Beobachter 2		
Meßwerte für Aufmerksamkeit	1	76	72	148	1	71	72	143
	2	46	49	95	2	51	51	102
	3	77	76	153	3	77	74	151
	
	
	
	22	44	38	82	22	65	67	132
	\bar{x}	72,5	69,8		\bar{x}	71,4	70,9	
	s	16,3	16,2		s	15,1	14,3	
Beobachter-übereinstim. (Summenübereinstimmung)		.96			.98			
Retest - Reliabilität	Beobachter 1		.82					
	Beobachter 2			.84				
	Beobachter 1 + 2				.84			

und enthielt 120 Beobachtungsintervalle. Die Zahl der auswertbaren Intervalle war meist etwas geringer wegen gelegentlicher kurzfristiger Unterbrechungen der Beobachtung. Als anschauliches, von der Länge der Beobachtung unabhängiges Maß für die Aufmerksamkeit eines Kindes wurden die Punktwerte über alle auswertbaren Intervalle summiert und durch die maximal mögliche Punktzahl dividiert („Leistungsmäß“ in Tab. 2). Dieser „prozentuale Aufmerksamkeitswert“ ist in Tabelle 4 für jedes Kind aufgelistet, getrennt nach Beobachtern und Beobachtungszeitpunkten. Die Verteilung der prozentualen Aufmerksamkeit weicht nicht wesentlich in der Normalverteilung ab (Ehrhardt et al. 1981). Deshalb konnte die Retest-Reliabilität als Produkt-Moment-Korrelation berechnet werden und betrug für die Beobachter

$r = .82$ und $.84$, für den Summenwert beider Beobachter $r = .84$.

Stabilität des Merkmals „Aufmerksamkeit im Unterricht“. Die Höhe der Retest-Reliabilität ist abhängig von der Zuverlässigkeit des Meßinstruments (Beobachter) und der Stabilität des gemessenen Merkmals. Bereinigt man die Retest-Reliabilität um den Meßfehler, der durch Abweichungen der Beobachter zustande kommt, kann man die Stabilität des Merkmals nach folgender Formel berechnen:

$$r'_{tt} = \frac{r_{tt}}{r_1(1,2) \cdot r_2(1,2)}$$

Einsetzen der Werte aus Tabelle 3 ergibt $r'_{tt} = .87$. Da die hier verwendete Summenübereinstimmung den Meßfehler eher unterschätzt, ist die Stabilität höher anzusetzen.

4. Diskussion

Objektivität und Zuverlässigkeit der mittel SVB erhobenen Daten hängen u.a. von der Eindeutigkeit und Vollständigkeit der Kategorien, der Genauigkeit der Kodierregeln und dem Beobachtertraining ab (Ehrhardt et al. 1981). Für die empirische Prüfung der Güte ist die Beobachterübereinstimmung das wichtigste Kriterium. In den Arbeiten, die bisher versucht haben, Aufmerksamkeit im Unterricht mittels SVB zu erfassen, ist die Beobachterübereinstimmung – soweit überhaupt angegeben – als prozentuale PPÜ berechnet worden. Die Werte liegen zwischen 82 % und 100 % (Lahaderne 1968; Cobb 1972; Kazdin 1973; Samuels & Turnure 1974; Kupietz & Richardson 1978). Wie in der Einleitung dargestellt, kann die prozentuale PPÜ ohne Berücksichtigung der Zahl der verwendeten Kategorien und ihrer tatsächlichen Auftretenshäufigkeit nicht eindeutig interpretiert werden. Außerdem handelt es sich nicht um ein zufallskritisches Maß.

Deshalb haben wir ein statistisches Prüfverfahren, das einem „bedingten Test“ entspricht, aus der Literatur entnommen (Light 1971), standardisiert (κ_0) und so dargestellt, daß es bei SVB für beliebiges Verhalten und unabhängig von der Zahl der Kategorien und der benutzten Auftretenshäufigkeiten angewendet werden kann zur Beantwortung der Frage: *In wieviel Beobachtungsfällen stimmen die Beobachter signifikant überein?* In der vorliegenden Untersuchung haben die beiden Beobachter nach 4 Wochen Training die Anwendung des Kategoriensystems so gut beherrscht, daß sie in 95 % der Fälle übereinstimmten. Hierin sehen wir einen Hinweis, daß mit der SVB unter geeigneten Bedingungen eine Objektivität erzielt werden kann, die derjenigen von Testverfahren vergleichbar ist.

Für die Reliabilitätsuntersuchung war es notwendig, die Unterrichtssituation so weit wie möglich zu standardisieren, und zwar bezüglich Schulstunde, Unterrichtsinhalt, Lehrerverhalten und Unterrichtsdauer. Die zweite und vierte Schulstunde wurden ausgewählt, weil in der ersten Stunde das Verhalten der Kinder recht schwankend ist, desgleichen in der dritten Stunde nach der langen Pause. Durch die zufällig festgelegte Reihenfolge, in der die Kinder über die Wochentage beobachtet wurden, kann erwartet werden, daß sich Einflüsse wie „gute oder schlechte Tage“ beim Lehrer oder Störeinflüsse von seiten der Kindergruppe im Sinne einer Randomisation ausgleichen.

Angaben zur Retest-Reliabilität haben wir in der Literatur bei SVB von Aufmerksamkeit nicht gefunden. Bei Testverfahren liegen die Werte zur Halbierungsreliabilität etwa gleich, zur Retest-Reliabilität in der Regel etwas höher als unsere Werte. Dafür hat die SVB den Vorteil, daß reale Situation und Meßsituation identisch sind. Die Retest-Reliabilitäten von .82 und .84 können als zufriedenstellend beurteilt werden.

Das Vorgehen bei der Bestimmung der Retest-Reliabilität war auf eine getrennte Abschätzung der Zuverlässigkeit des Meßinstrumentes (Beobachterübereinstimmung) und der (u.U. nur geringen) Stabilität des gemessenen Merkmals angelegt. Die Beobachterübereinstimmung konnte dabei jedoch nur als Summenübereinstimmung berechnet werden (Tab. 4), die die tatsächliche Übereinstimmung, wie eingangs dargestellt, überschätzt. Die durchschnittliche Stabilität der Aufmerksamkeit im Unterricht ist mit .87 erstaunlich hoch und ist bei Berücksichtigung der rechnerisch zu hohen Beobachterübereinstimmung eher noch höher anzusetzen.

Nach diesen Vorarbeiten meinen wir, daß das Verfahren jetzt eingesetzt werden kann, um den Einfluß von Unterrichtsbedingungen (z.B. bestimmtes didaktisches oder methodisches Vorgehen, Unterrichtsdauer, Gruppenstärke) oder den Erfolg von Förderprogrammen (z.B. Aufmerksamkeitstraining nach Wagner 1976) auf die Aufmerksamkeit der Kinder im Unterricht prüfen zu können.

Zusammenfassung

Eine neue Anwendung der systematischen Verhaltensbeobachtung zur fortlaufenden Registrierung von Aufmerksamkeit einzelner Kinder während des Unterrichts wurde geprüft, ob sie ausreichend objektive und zuverlässige Daten liefert. Dazu wurden zwei Beobachter in der Anwendung des Kategoriensystems trainiert und 60 Kinder der 1. und 2. Grundschulklasse wiederholt während des Unterrichts beobachtet. Verschiedene Verfahren zur Berechnung der Beobachterübereinstimmung werden erläutert und deren Probleme an den empirischen Daten vergleichend dargestellt. Nach 4 Wochen Training der Beobachter wurde bei Anwendung eines zufallskritischen Prüfmodells eine Übereinstimmung in 95 % der Beobachtungen erreicht. Die Wiederholungs-Reliabilität im Abstand von 8 Tagen war mit .84 zufriedenstellend.

Summary

A new application of systematic behavior observation, the continuous recording of classroom attention in individual children was examined for its objectivity and reliability. To this end, two observers were trained in the use of the categorial system, and 60 children in the first and second grade were repeatedly observed during classes. Several methods for the calculation of inter-observer agreement are discussed in light of our empirical data and a statistically meaningful model is suggested. After training the observers for four weeks, we found agreement in 95 % of the observations. Retest-reliability after 8 days (.84) was satisfactorily.

Literatur

- Bartenwerfer, H.: Allgemeine Leistungstests. In R. Heiss et al. (Hg.), Handbuch der Psychologie, Bd. VI. Göttingen: Hogrefe, 1964, S. 386–410.
- Bijou, S. W., Peterson, R. F. & Ault, M. H.: A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. Journal Applied Behavior Analysis, 1968, 1, 175–191.
- Brickenkamp, R.: Handbuch psychologischer und pädagogischer Tests. Göttingen: Hogrefe, 1975.
- Cattell, R. B.: Die empirische Erforschung der Persönlichkeit. Weinheim: Beltz, 1973.
- Cobb, J. A.: Relationship of discrete classroom behaviors to fourthgrade academic achievement. Journal Educational Psychology, 1972, 63, 74–80.
- Ehrhardt, K. J., Haack, R., Klich, C., Marinello, G., Plassmann, M. L., Wenzel, H. & Winzer, A.: Entwicklung eines Kategorien-systems zur systematischen Verhaltensbeobachtung verschiedener Aspekte der Aufmerksamkeit von Kindern im Unterricht. (Unveröffentlicht. Manuscript, Psychologisches Institut der Univ. Düsseldorf, 1980).
- Ehrhardt, K. J., Findeisen, P., Marinello, G. & Wenzel, H.: Systematische Verhaltensbeobachtung von Aufmerksamkeit bei Grundschülern während des Unterrichts. Psychologie in Erziehung und Unterricht, 1981, 28, 204–213.
- Fassnacht, G.: Systematische Verhaltensbeobachtung. München: Reinhardt, 1979.
- Fricke, R.: Testgütekriterien bei lehrzielorientierten Tests. Zeitschrift Erziehungswissenschaftliche Forschung, 1972, 150–175.
- Heller, K. & Nickel, H. (Hg.): Psychologie in der Erziehungswissenschaft, Band IV: Beurteilen und Beraten. Stuttgart: Klett, 1978.
- Kazdin, A. E.: The effect of vicarious reinforcement on attentive behavior in the classroom. Journal Applied Behavior Analysis, 1973, 6, 71–78.
- Kupietz, S. S. & Richardson, E.: Children's vigilance performance and inattentiveness in the classroom. Journal Child Psychology and Psychiatry, 1978, 19, 145–154.
- Lahaderne, H. M.: Attitudinal and intellectual correlations of attention: A study of four sixth-grade classrooms. Journal Educational Psychology, 1968, 59, 320–324.
- Langhorst, E.: Beobachtung und Beurteilung des Schülerverhaltens im Unterricht. In K. Heller (Hg.), Leistungsbeurteilung in der Schule. Heidelberg: Quelle & Meyer, 1974, S. 230–252.
- Lienert, G. A.: Testaufbau und Testanalyse, Weinheim: Beltz, 1969³.
- Light, R. J.: Measures of response agreement for qualitative data. Some generalizations and alternatives. Psychological Bulletin, 1971, 76, 365–377.

- L i n d n e r , K.: Die Überprüfbarkeit des Konkordanzmaßes "0". Zeitschrift empirische Pädagogik, 1980.
- M a r i n e l l o , G.: Unterrichtsbezogenes Aufmerksamkeitsverhalten von Grundschülern. Entwicklung und Überprüfung eines Beobachtungsverfahrens. Diplomarbeit der Pädagogischen Fakultät der Universität Bonn, 1980.
- M e e s , U.: Verhaltensbeobachtung in der natürlichen Umgebung. In U. Mees & H. Selg (hg.), Verhaltensbeobachtung und Verhaltensmodifikation. Anwendungsmöglichkeiten im pädagogischen Bereich. Stuttgart: Klett, 1977.
- R e i n a r z t , H.: Die Wirkung audio-visueller Medien auf das Aufmerksamkeitsverhalten von Grundschülern. Diplomarbeit der Pädagogischen Hochschule Rheinland, Abt. Neuß, 1979.
- S a m u e l s , S. J. & T u r n u r e , J. E.: Attention and reading achievement in first-grade boys and girls. Journal Educational Psychology, 1974, 66, 29–32.
- S c h o u t e n , J. J. A.: Measuring pairwise agreement among many observers. Biometrical Journal, 1980, 22, 497–504.
- W a g n e r , I.: Aufmerksamkeitstraining mit impulsiven Kindern. Stuttgart: Klett, 1976.

Anschrift der Verfasser:

Dr. med. Klaus Jürgen Ehrhardt
 Dr. rer. nat. Peter Findeisen, Dipl. math.
 Psychologisches Institut der Universität
 Universitätsstraße 1
 D 4000 Düsseldorf

Untersuchungen zur Validität von Kontroll-Skalen für Soziale Erwünschtheit und Akquieszenz¹⁾

Manfred Amelang & Peter Borkenau

1. Einleitung und Fragestellung

Bekanntlich sind Fragebogen zur Erfassung von Persönlichkeitsmerkmalen leicht verfälschbar. Wie mehrere Untersuchungen belegen, führen experimentell variierte Instruktionen im Sinne von „Faking Good“ bzw. „Faking Bad“ zu Skalen-Mittelwerten, die sich nicht nur voneinander, sondern auch jeweils von den unter Normalinstruktionen erhaltenen Resultaten unterscheiden (z.B. Irvine & Gendreau 1974). Darüberhinaus bereitet offenkundig auch die anweisungsgemäße Übernahme oder vorstellungsmäßige Vergegenwärtigung verschiedener Rollen bei der Bearbeitung von Fragebogen für die Probanden keinerlei Probleme – mit der Konsequenz verschiedener gruppenspezifischer Testwertprofile (Hoeth, Büttel & Feyereabend 1967; Kröger & Turnbull 1975). Basierend auf derartigen Studien hat sich die Einsicht allgemein durchgesetzt, daß der Einsatz von Persönlichkeitstests in der Ernstsituation von Selektionsprozeduren, etwa bei der Auslese von Stellenanwärtern, nicht sinnvoll ist (Hampel & Klinkhamer 1978; Thornton & Gerasch 1980).

Wenn verschiedene Anweisungen zu deutlich unterscheidbaren Testwerten führen, liegt die Vermutung nahe, daß gewisse Verfälschungsbedingungen auch im Rahmen der „Normal“-Instruktion von Forschungssituationen auftreten bzw. fortbestehen. Zur Erfassung der interindividuellen Differenzen im Ausmaß solcher Reaktionstendenzen behilft man sich mit Kontrollskalen. Besonders häufig ist die Verwendung von „Lügen“-Fragen, d.h. von Items, die sich auf die Tendenz der Versuchspersonen beziehen, Antworten überwiegend im Sinne von Sozialer Erwünschtheit (SE) zu liefern. Verschiedentlich wird auch versucht, die individuelle Akquieszenz-Neigung zu erfassen.

Implizit wird hier wie dort davon ausgegangen, daß die Punktwerte in Kontrollskalen Hinweise darauf vermitteln, inwieweit auch die Testwerte der „inhaltlichen“ Skalen durch den Einfluß von SE- und Ja-Sage-Tendenz überlagert und damit in ihrer Aussagekraft beeinträchtigt sind.

Konsequenterweise empfehlen deshalb die Autoren von Tests gewöhnlich, Pbn mit hohen Punktwerten in Lügenskalen zu eliminieren (in diesem Sinne z.B. Eysenck 1971), ihre „E- und besonders die N-Werte mit Skepsis (zu betrachten“ (Egger 1974) oder die entsprechenden Protokolle „nur sehr zurückhaltend zu interpretieren“ (Fahrenberg & Selg 1970).

1) Die Untersuchung wurde von der Deutschen Forschungsgemeinschaft unterstützt (Az Am 37/5).