Arch. Psychiat. Nervenkr. 225, 73—85 (1978)

Archiv für Psychiatrie und Nervenkrankheiten
Archives of Psychiatry and Neurological Sciences

Interrater-Reliabilität von AMP-Symptomen

B. Woggon[1], U. Baumann[2] und J. Angst[1]

[1] Psychiatrische Universitätsklinik Zürich, Forschungsdirektion (Direktor: Prof. Dr. med. J. Angst), Zürich
[2] Institut für Psychologie der Christian-Albrecht-Universität, Kiel

Interrater Reliability of AMP Symptoms

Summary. Two psychiatrists examined 48 patients (25 depressed and 23 schizophrenic). Each documented the symptoms on AMP sheets 3 (psychopathologic symptoms) and 4 (somatic signs, first column only). The study deals with 139 AMP symptoms. Seventy could be judged concerning symptom exists/does not exist. Of these 70 symptoms, 45 showed a good or moderate interrater reliability. Specific symptoms had a better reliability than nonspecific. Symptoms described by the patients had a better reliability than those judged by the doctor alone. The results indicate that expanded use of the AMP system in its present form is problematic. Work on a new version of the AMP system has already begun.

Key words: AMP system – Interrater reliability.

Zusammenfassung. 48 Patienten (25 depressive und 23 schizophrene Patienten) wurden jeweils von 2 Untersuchern interviewt und die Befunde auf dem AMP-Bogen 3 (psychischer Befund) und 4 (somatischer Befund, nur erste Spalte) festgehalten. Die Studie bezieht sich auf 139 AMP-Symptome, von denen 70 bezüglich der Einstufung vorhanden/nicht vorhanden beurteilt werden konnten. Von diesen weisen 45 eine gute bis mittlere Übereinstimmung zwischen den Beurteilern auf. Spezifische Symptome zeigen eine bessere Reliabilität als unspezifische, auf Selbstbeurteilung beruhende eine bessere als auf Fremdbeurteilung zurückzuführende. Die Ergebnisse lassen eine verbreitete Anwendung des AMP-Systems in der vorliegenden Form als kritisch erscheinen. Veränderungen des AMP-Systems sind in Vorbereitung.

Schlüsselwörter: AMP-System – Interrater-Reliabilität.

Sonderdruckanforderungen an: Dr. med. Brigitte Woggon, Psychiatrische Universitätsklinik, Forschungsdirektion, Postfach 68, CH-8029 Zürich, Schweiz

0003-9373/78/0225/0073/$ 02.60

1. Einleitung

Im psychiatrischen Bereich nehmen Fremdbeurteilungsverfahren — sogenannte Ratings — eine besondere Stellung ein (Mombour, 1972; Pichot, 1974). Bei klinischen Prüfungen von Psychopharmaka haben sie sich zur unentbehrlichen Untersuchungsmethode entwickelt; das Sammelreferat von Angst und Dinkelkamp (1974) belegt dies z. B. deutlich im Bereich der somatischen Therapie Schizophrener. Fremdbeurteilungsverfahren werden wie psychologische Selbstbeurteilungsverfahren (Fragebögen, Intelligenztests usw.) nach formalen und inhaltlichen Kriterien auf ihre Güte geprüft (Lienert, 1969). Bei den formalen Kriterien ist das Gütemaß meistens die Beurteilerübereinstimmung, die Interrater-Reliabilität. Diese gibt Auskunft darüber, wie weit die Beurteilung eines Patienten durch den Beurteiler (Rater) generalisiert werden kann. Bei einem Rating haben wir als Varianzquelle den Patienten, den Rater und ihre Interaktion. Nur wenn die beiden letzteren Quellen wenig zur Varianz beitragen, ist eine hohe Beurteilerübereinstimmung zu erwarten. Zur Varianz tragen bei der Beurteiler (z. B. Halo-Effekte, die interindividuell verschieden sind) und die Beurteilungs-Skala, die z. B. durch unklare Definitionen eine Übereinstimmung erschwert (Hasemann, 1964; Langer und Schulz von Thun, 1974). Durch entsprechendes Ratertraining und sorgfältige Konstruktion der Beurteilungs-Skala läßt sich daher die Übereinstimmung verbessern. Bei der sogenannten Verhaltensbeobachtung wird versucht, den Einfluß des Beurteilers (Ratergröße) gering zu halten; im Extremfall liegt die automatische Auswertung von Ausdrucksverhalten vor.

Auf Grund unserer Überlegungen darf für Symptombeurteilungen keine allzu hohe Übereinstimmung erwartet werden, da Symptome meistens auf Grund komplexer Interaktionen zwischen Beurteiler und Patient im Rahmen eines Interviews beurteilt werden. Dennoch stellt sich die Frage, welche Symptome einer Beurteilungs-Skala (eines Ratings) relativ gut und welche schlechter einzustufen sind.

In unserer Studie wird diese Frage bezüglich des AMP-Systems untersucht. Das AMP-System (Angst et al., 1969; Scharfetter, 1972), das der Dokumentation des psychiatrischen Befundes dient, ist im deutschen Sprachgebiet weit verbreitet, doch liegt bisher nur eine Interrater-Studie vor, die eine formale Beurteilung des Systems zuläßt. In der Arbeit von Gebhardt und Helmchen (1973) wird die Übereinstimmung zwischen mehreren Beurteilern (6 der 20 Beurteiler waren bei allen Patienten anwesend) berechnet. Bei dieser Arbeit sind aber einige Punkte problematisch: Beurteilerselektion (nur 6 von 20 Beurteilern gelangten in die Auswertung), Trainingscharakter der Studie (jeder Fall wurde nach der Registrierung des psychopathologischen Befundes ausführlich diskutiert), künstliche Versuchssituation (Patient wurde vor 13—20 Beurteilern exploriert), Trennung von Beurteilern und Interviewer (entspricht meist nicht der Realität; Interviewer gehen als Konstante in den Versuch, müßten aber auch variiert werden). Es scheint daher gerechtfertigt, eine mehr der klinischen Situation entsprechende Studie durchzuführen, insbesondere da bekannt ist, daß die Güte der Übereinstimmung durch eine künstliche Versuchssituation eher erhöht wird. Auf die in einem AMP-Seminar durchgeführte Studie (Busch et al., 1975) kann hier nicht

weiter eingegangen werden, da zum Teil die erwähnten Punkte auch hier zutreffen; vor allem aber war die Variabilität der Rater so groß, daß exakte Ergebnisse nicht zu erwarten waren (zum Teil keine Kenntnisse, zum Teil mehrjährige Erfahrung mit dem AMP-System).

Ebenso wird an dieser Stelle auf die Arbeit von Maurer-Groeli (1976) nicht eingegangen, da sie sich mit der Interrater-Reliabilität auf Skalenebene beschäftigt hat; dieser Punkt soll in einer weiteren Arbeit behandelt werden. Zur Zeit ist eine größere Beurteiler-Übereinstimmungsstudie in Auswertung, die auf der Einstufung von Videobändern in mehreren Kliniken beruht.

2. Fragestellung

In unserer Studie sollten im einzelnen folgende Fragen geklärt werden:

— Wie gut ist die Übereinstimmung von Beurteilern bei den einzelnen Symptomen?
— Ist die durchschnittliche Übereinstimmung in Symptomuntergruppen gleich oder verschieden? Dazu wird einerseits die von Baumann vorgeschlagene Unterteilung in unspezifisch häufige und spezifische Symptome verwendet, die für die psychischen (Baumann et al., 1976) und die somatischen Symptome durchgeführt worden ist (Baumann und Angst, 1977). Andererseits sollte eine Unterteilung der Symptome nach der Beobachtungsebene gemacht werden (Woggon, in Vorbereitung): Selbstbeurteilung, Fremdbeurteilung, gemischte Beurteilung.

Neben diesen im Vordergrund der Studie stehenden Fragen, über die in dieser Arbeit berichtet wird, wurden weitere Probleme untersucht, die hier aber nur kurz erwähnt werden können (ausführlich: Baumann und Woggon, in Vorbereitung): Übereinstimmung auf Skalenebene; Übereinstimmung in Abhängigkeit von Diagnosegruppen (schizophrene und depressive Patienten); Abhängigkeit der Ergebnisse vom Ausbildungsgrad der Beurteiler, — vom Schweregrad der Erkrankung des Patienten, — von der diagnostischen Sicherheit.

3. Versuchsplan[1]

Da das AMP-System für den Bereich der endogenen und der körperlich begründbaren Psychosen konzipiert worden ist und in den Psychopharmakastudien vorwiegend schizophrene und depressive Patienten untersucht werden, interessierte vor allem die Interrater-Reliabilität des AMP-Systems bei schizophrenen und depressiven Zustandsbildern. Es war daher geplant, beide Gruppen gleich stark in die Untersuchung miteinzubeziehen. Die Zahl der Patienten war durch die Zahl der Untersucher (vergleiche Kapitel 3.1.) bedingt und wurde auf 48 festgelegt. Von einem nicht an den Untersuchungen beteiligten Arzt wurden die Patienten ausgewählt und nach einer Randomisierungs-Liste den Untersucher-Paaren zugeteilt. Dieses Vorgehen wurde gewählt, damit die Untersucher die jeweiligen Patienten nicht vor der Exploration kennenlernten.

[1] Den an der Untersuchung beteiligten Assistenzärzten W. Felder, R. Frey, D. Lamparter, B. Radanov und H. Reiser sei für ihre Mitarbeit bestens gedankt

3.1. Raterauswahl

Die Zahl der Beurteiler ($N=4$) war durch die in der Forschungsdirektion tätigen Assistenzärzte, die im pharmakopsychiatrischen Bereich tätig waren, bedingt. Die Ärzte sollten paarweise die Patienten untersuchen (einer führte das Interview durch, und der andere konnte Zusatzfragen stellen). Mit dieser Konstellation sollte die Realität möglichst gut nachgeahmt werden, in der 1—2 Rater einen Patienten beurteilen. Jeder Rater wurde randomisiert mit jedem andern gekoppelt, damit keine systematische Verzerrung durch einzelne Teams erfolgte. Bei 4 Ratern ergeben sich 6 Kombinationen und bei gleichmäßiger Berücksichtigung der einzelnen Kombinationen ein mehrfaches von 6 für die Patientenzahl ($8 \times 6 = 48$). Jeder Untersucher war 12mal als Interviewer und 12mal als Nebeninterviewer tätig. Um Ermüdungserscheinungen zu vermeiden, wurden pro Tag nicht mehr als 3 Interviews durchgeführt, wobei die Dauer des Interviews nicht begrenzt war. Die Untersuchung der Patienten erfolgte in der Regel nicht auf der Krankenabteilung, sondern in den Arztzimmern der Forschungsabteilung. Nur sehr unruhige Patienten bildeten diesbezüglich eine Ausnahme. Die Patienten wurden durch einen nicht an den Untersuchungen beteiligten Arzt ins Untersuchungszimmer gebracht und dort erstmals vom Interviewer und Rater begrüßt.

3.2. Untersuchungsverfahren, Auswertung

Im Anschluß an das Interview füllten die beiden Untersucher getrennt die AMP-Bogen 3 (psychischer Befund) und 4 (somatischer Befund, dabei nur 1. Spalte), einen Interviewbogen und einen Diagnosebogen aus.

Die Datenauswertung[2] erfolgte im Rechenzentrum der Universität Zürich (IBM 370-155) mit Programmen der Programmbibliothek PSYCHLIB (U. Baumann) und ad-hoc-Programmen.

4. Statistische Auswertung

4.1. Allgemeine Probleme

Die Wahl der statistischen Auswertungsmethode bei Interrater-Studien richtet sich nach der Zahl der Beurteiler, Zahl der beurteilten Personen sowie Anzahl und Skalenniveau der beurteilten Merkmale (Langer und Schulz von Thun, 1974; Tinsley et al., 1975).

Für Nominaldaten, wie sie in dieser Arbeit vorliegen, hat der von Cohen entwickelte Kappa-Koeffizient Verbreitung gefunden (Cohen, 1960, 1972; Signifikanzprüfung: Everitt, 1968; Fleiss et al., 1969).

Kappa ist wie folgt zu interpretieren: +1 vollständige Übereinstimmung, 0 Übereinstimmung ist von Zufallsübereinstimmung nicht zu unterscheiden, −1 minimale Übereinstimmung, Zufallsübereinstimmung kann durch „blindes Raten" (Randverteilungen 50:50) oder durch Raten auf Grund von Konzepten erfolgen (Wissen, daß Symptom X im allgemeinen selten vorkommt; dies bedingt eine einseitige Randverteilung von 10:90). Dabei darf man aber Kappa = 0 nicht mit „keine Übereinstimmung" interpretieren, wie das folgende Beispiel zeigt:

Wir haben eine Vierfeldertafel mit Symptom vorhanden (+), nicht vorhanden (−) und berechnen Kappa und das Übereinstimmungsprozent (Felder ++ und −− auf N bezogen).

---

[2] Fräulein G. Schneidewind, die mit viel Sorgfalt die Programme erstellte und bei der Aufbereitung der Tabellen behilflich war, sei besonders gedankt

a) Feld		b) Feld	
++	48	++	100
−−	8	−−	0
+−	32	+−	0
−+	12	−+	0
Kappa = 0.0		Kappa = 0.0	
Übereinstimmungsprozent = 56%		Übereinstimmungsprozent = 100%	
Diskrepanzprozent = 44%		Diskrepanzprozent = 0%	

Im Beispiel a ist das Übereinstimmungsprozent 56%, der Prozentsatz an Diskrepanz aber 44%, was sich in einem niedrigen Kappa niederschlägt. Daher ist die Interpretation gerechtfertigt, daß die Übereinstimmung zwischen den Ratern von der Zufallsübereinstimmung nicht zu unterscheiden ist: die Interrater-Reliabilität ist gering. Im Beispiel b dagegen ist das Übereinstimmungsprozent 100% und dennoch Kappa = 0. Dies ist dadurch bedingt, daß das Merkmal nicht variiert; dabei läßt sich nicht auseinanderhalten, ob die Beurteiler keine Variation wahrgenommen haben oder ob bei den Patienten keine Variation vorkam. Kappa ist daher nicht zu interpretieren, ebenso nicht das Übereinstimmungsprozent. Dies gilt auch für den von Maxwell (1977) vorgeschlagenen Koeffizienten RE (Übereinstimmungsprozent — Diskrepanzprozent).

Auf Grund unserer Überlegungen soll bei der Auswertung Kappa jeweils mit der prozentualen Übereinstimmung in Bezug gesetzt werden. Bei hoher Übereinstimmung und niedrigem Kappa können wir keine Aussage machen; dies gilt insbesondere, wenn eine oder zwei Randsummen Null sind. Diesen Fall werden wir mit „Kappa nicht berechenbar" bezeichnen. Bei der Interpretation von Kappa verwenden wir aus rechentechnischen Gründen die Näherungsformel zur Signifikanzberechnung[3] (Cohen, 1960; Fleiss et al., 1969). Über die zu fordernde Höhe ist aus der Literatur nichts bekannt.

Folgende Überlegungen können aber behilflich sein. Die Gesamtzahl einer Vierfeldertafel können wir aufgliedern in: $N = U_z + D_z + U + D$. Dabei sind U_z und D_z die auf Grund der Randverteilung zu erwartenden Übereinstimmungen, resp. Diskrepanzen (Zufallsübereinstimmungen, -diskrepanzen). U und D sind die Differenzen zwischen beobachteter und zufälliger erwarteter Übereinstimmung, resp. Diskrepanz ($U = -D$). U ist dann am größten, wenn U sämtliche Diskrepanzfälle, die man zufällig erwarten kann, umfaßt ($= D_z$). Kappa ist daher definiert als Zahl der überzufällig erreichten Übereinstimmungen gemessen an der Zahl der maximal möglichen überzufälligen Übereinstimmungen: Kappa $= U/D_z$. Verlangt man, daß dieser Prozentsatz mehr als die Hälfte beträgt, so gelangt man zu Kappa ≥ 0.50 für die numerische Mindestanforderung an gute Übereinstimmung. In unserer Arbeit interpretieren wir verschärft Kappa $\geq 0{,}6$ als gute und $0{,}4 \geq$ Kappa $\leq 0{,}6$ als mittlere Übereinstimmung.

4.2. Untersuchungsspezifische Probleme

Für die statistische Bearbeitung der AMP-Symptome werden die Kategorie „nicht beurteilbar" und „fraglich vorhanden" zusammengefaßt (vergleiche Baumann et

[3] Bei extremer Randverteilung kann es bei Anwendung der Schätzformel, die eher zu Unrecht die Nullhypothese beibehalten läßt, zum paradoxen Fall kommen, daß Kappa = 1, der z-Wert aber gering ist, Beispiel Feld ++47, +−0, −+0, −−1. Kappa = 1, $z = 1{,}4$

al., 1975) und — da diese Sammelkategorie sehr selten vorkam — diese der Kategorie „nicht vorhanden" zugeordnet. Eine Auswertung der Symptome bezüglich der Graduierung (leicht, mittel, schwer vorhanden) ist bei einer Patientenzahl von 48 problematisch; wir haben daher das Hauptgewicht auf die Auswertung nach „Symptom vorhanden/nicht vorhanden" gelegt.

Wie in Kapitel 3.1. erwähnt, ist die Zusammensetzung der Untersucher variiert worden, dabei hat jeder Rater an 24 Interviews teilgenommen. Primär kann man Vierfeldertafeln vom Typus „Beurteiler X versus alle übrigen" rechnen. Aus Platzgründen werden wir bei den Ergebnissen diese Koeffizienten nicht diskutieren. Für die Beurteilung der Symptome wichtiger ist eine globale Charakterisierung, die sich auf alle 48 Patienten bezieht. Dazu könnte man die Einzelkoeffizienten (X versus Rest) mitteln, doch hat man keine unabhängigen Werte.

In Analogie zur Intraklasskorrelation haben wir die Einzeltafeln (X versus Rest) durch Aufeinanderlegen zu einer Summentafel zusammengefaßt und davon Übereinstimmungsprozente und Kappa berechnet. Die Ergebnisse sind mit den Mittelwerten der Einzelwerte fast identisch, weswegen wir nur die Kennwerte der Summentafeln berichten werden.

5. Beschreibung der Beurteiler

Bei den Untersuchern (3 männlich, 1 weiblich) handelt es sich um Assistenzärzte unserer Forschungsabteilung. Zum Zeitpunkt der Untersuchung waren sie zwischen 6 und 20 Monaten in unserer Abteilung tätig. Während dieser Zeit hatten sie an dem 2mal wöchentlich stattfindenden 2stündigen AMP-Interrater-Training teilgenommen und waren daher mit dem AMP-System vertraut. 3 Untersucher hatten während ihrer gesamten Tätigkeit in unserer Forschungsabteilung bei Psychopharmakaprüfungen mitgearbeitet und waren daher in der täglichen Anwendung des AMP-Systems gut geübt. 1 Untersucher hatte nur zeitweilig bei Psychopharmakaprüfungen mitgearbeitet. 2 Untersucher hatten ihre psychiatrische Ausbildung in unserer Abteilung begonnen, 1 Rater verfügte über eine vorgängige $2\frac{1}{2}$jährige Ausbildung, und ein anderer war vorher 5 Jahre psychiatrisch tätig gewesen.

6. Patientenbeschreibung

Die diagnostische Zuordnung (ICD) der Paienten ist aus Tabelle 1 ersichtlich. Zahlenmäßig am häufigsten waren die paranoide Schizophrenie und die endogene Depression mit je 11 Patienten vertreten.

Abweichend von der ursprünglich geplanten diagnostischen Zusammensetzung der Patientenstichprobe aus 24 depressiven und 24 schizophrenen Patienten wurden wegen diagnostischer Unsicherheit und dadurch bedingtem Diagnosewechsel bei einem Patienten schließlich 25 depressive und 23 schizophrene Zustandsbilder in der Studie untersucht.

25 Patienten waren männlich und 23 weiblich. Das durchschnittliche Lebensalter betrug $M=38{,}4$ ($s=12{,}7$) Jahre mit einer Spannbreite von 17—68 Jahren.

Bei 6 Patienten handelte es sich um eine Ersterkrankung. Die durchschnittliche Krankheitsdauer betrug $M=9{,}4$ ($s=8{,}2$) Jahre mit einer Spannbreite von

Tabelle 1. Diagnosen

ICD-Nr.	Diagnose	Patientenzahl
295.1	Hebephrene Schizophrenie	7
295.2	Katatone Schizophrenie	1
295.3	Paranoide Schizophrenie	11
295.7	Schizoaffektive Psychose	4
296.0	Involutionsdepression	3
296.1	Manie	1
296.2	Endogene Depression	11
296.3	Zirkuläre Verlaufsform manisch-depressiver Psychosen	2
298.0	Reaktive depressive Psychose	3
300.4	Depressive Neurose	4
300.7	Hypochondrische Neurose	1
		48

0—26 Jahren. Die Dauer bis zur Aufnahme bei der jetzigen Erkrankung war bei 3 Patienten unbekannt, betrug bei 8 Patienten weniger als 1 Woche, bei 14 Patienten weniger als 1 Monat, bei 13 bis zu 6 Monaten, bei 3 weniger als 1 Jahr und bei 7 Patienten mehr als 1 Jahr. Weitere Angaben zu den Patienten (Syndromwerte) sind in Baumann und Woggon (in Vorbereitung) aufgeführt.

7. Angaben zum Interview

Bei jedem Interview wurde die Dauer registriert. Die mittlere Interviewdauer betrug $M=45$ min ($s=13.7$), am häufigsten war eine Dauer von 30 min (25 min: 1mal; 30 min: 11mal; 35 min: 3mal; 40 min: 7mal). Die benötigte Interviewdauer ist damit der von Mombour (1972) angegebenen Interviewdauer von 35—45 min vergleichbar, die zum Ausfüllen der IMPS nach Lorr notwendig ist. Gerade für ein so detailliertes Beurteilungs-System wie das AMP-System scheint es uns wichtig, daß mit der ermittelten Interviewdauer die praktische Verwendbarkeit des AMP-Systems unterstrichen wird.

Für jeden Patienten mußten der Interviewer und der Rater beurteilen, wie gut der Patient zu explorieren war. Dabei wurde entsprechend den in der Schweiz gebräuchlichen Schulnoten graduiert (6 = ausgezeichnet, 5 = gut, 4 = befriedigend, 3 = mäßig, 2 = schwach, 1 = ungenügend).

37 Patienten (77,2%) wurden als ausgezeichnet bis befriedigend explorierbar eingestuft. 11 Patienten (22,8%) wurden als mäßig bis ungenügend explorierbar eingestuft. Dabei fand sich zwischen schizophrenen und depressiven Patienten kein Unterschied in der Explorierbarkeit. Eine genaue Übereinstimmung bezüglich der Einstufung der Explorierbarkeit zwischen Interviewer und Rater lag

Tabelle 2. Verteilung von Kappakoeffizienten unter Berücksichtigung der Übereinstimmungsprozente. In Klammern: Zahl der Koeffizienten mit $P \geq 0{,}1$ (n.s.)

Kappa	Übereinstimmungsprozent														Anzahl	
	50—59		60—69		70—79		80—89		90—99		100					
≥0,9											7	(5)			7	(5)
0,8—0,9									5						5	
0,7—0,8							3		6						9	
0,6—0,7							5		8	(4)					13	(4)
0,5—0,6					6		3		6						15	
0,4—0,5					3		3	(1)	9	(9)					15	(10)
0,3—0,4			1		4		5	(3)							10	(3)
0,2—0,3			9	(7)	3	(3)	1	(1)	3	(3)					16	(14)
0,1—0,2	2	(2)			1	(1)	1	(1)							4	(4)
0,0—0,1	2	(2)			1	(1)									3	(3)
<0,1—0,0	1	(1)	1	(1)					11	(11)					13	(13)
Nicht berechenbar											29				29	
Anzahl	5	(5)	11	(8)	18	(5)	21	(6)	48	(27)	36	(5)			139	(56)

20mal vor. Bei der Einteilung in die 2 oben erwähnten Kategorien (ausgezeichnet bis befriedigend versus mäßig bis ungenügend) wurde eine Übereinstimmung in der Beurteilung in der Explorierbarkeit zwischen Interviewer und Rater 37mal erzielt.

8. Ergebnisse

8:1. Auswertung Symptom vorhanden/nicht vorhanden

Bei den folgenden Abschnitten beziehen wir uns auf 139 der 181 AMP-Symptome; die Kategorien „weitere körperliche Symptome" und „neurologische Sym-

Tabelle 3. Zahl der Symptome mit verschiedener Güte an Interrater-Reliabilität

Interrater-Reliabilität	Kappa	Signifikanz ($P \leq 0{,}1$)	Übereinstimmungsprozent Ü	Anzahl
Gut	≥0,6	s.	≥80%	25
Mittel	0,4–0,6	s.	≥80%	11 } 20
			70≤Ü<80%	9 }
Gering	<0,4	s. oder n.s.	<80%	25
Nicht beurteilbar	≥0,4	n.s.	≥80%	19 } 69
	<0,4	s. oder n.s.	≥80%	21 }
	nicht berechenbar, da Randsumme(n) Null			29 }

Tabelle 4. Symptomliste mit Angabe der Interrater-Reliabilität (B: Interrater-Reliabilität ist in der Berliner Studie niedrig)

1. Interrater-Reliabilität gut

Kappa ≥ 0,6, signifikant, Übereinstimmungsprozent ≥ 80%

Kappa ≥ 0,8

44 Beeinträchtigungs-, Verfolgungswahn, 48 Stimmenhören, 84 affektinkontinent (B), 99 abends besser, 142 Obstipation, 143 Diarrhoe. 152 Schwindel

0,7 ≤ Kappa < 0,8

11 Konzentrationsstörungen (B), 18 verlangsamt (B), 34 Wahnstimmung, 54 Störung der Ichidentität, 57 Gedankenausbreitung, 79 Schuldgefühl, 129 verstärkte Traumtätigkeit, 131 Appetit vermindert, 150 Kopfdruck o. ä.

0,6 ≤ Kappa < 0,7

16 gehemmt, 23 Vorbeireden, 25 Gedankenabreißen (B), 27 Hypochondrie (B), 36 Wahneinfälle/Wahngedanken (B), 51 Körperhalluzinationen/Körperfühlstörungen (B), 97 logorrhoisch, 108 Ablehnung der Behandlung, 112 Aggressionstendenzen (B)

2. Interrater-Reliabilität mittel

0,4 ≤ Kappa < 0,6, signifikant, Übereinstimmungsprozent ≥ 70%

0,5 ≤ Kappa < 0,6

17 gesperrt, 22 beschleunigt/ideenflüchtig, 55 andere Entfremdungserlebnisse, 59 Gedankeneingebung, 64 Gefühl der Gefühllosigkeit, 66 deprimiert/traurig, 67 hoffnungslos/verzweifelt, 69 gehoben/euphorisch, 74 klagsam/jammerig, 77 Insuffizienzgefühle, 80 Verarmungsgefühl, 85 affektstarr, 91 antriebsgesteigert, 104 Kontakt vermehrt, 105 Krankheitsgefühl

0,4 ≤ Kappa < 0,5

13 Gedächtnisstörungen, 35 Wahnwahrnehmungen, 56 Autismus, 109 Suizidtendenzen (B), 137 Mundtrockenheit

3. Interrater-Reliabilität gering

Kappa < 0,4, signifikant oder n.s., Übereinstimmungsprozent < 80%

0,3 ≤ Kappa < 0,4

62 ratlos (B), 68 ängstlich (B), 71 mißtrauisch/feindselig (B), 82 affektiv inadäquat (B), 125 Durchschlafstörungen

0,2 ≤ Kappa < 0,3

19 eingeengt (B), 20 umständlich, 24 inkohärent/zerfahren (B), 63 gefühlsverarmt/affektarm (B), 70 mürrisch gereizt/dysphorisch (B), 81 ambivalent, 92 motorisch unruhig, 107 Mangel an Krankheitseinsicht, 121 Beschäftigung erschwert, 103 Kontakt vermindert (B), 124 Einschlafstörungen, 128 Müdigkeit

Kappa < 0,2

10 Auffassungsstörungen (B), 65 Störung der Vitalgefühle (B), 72 gespannt (B), 73 innerlich unruhig (B), 87 antriebsarm (B), 88 antriebsgehemmt (B), 94 negativistisch, 126 Verkürzung der Schlafdauer

ptome" sowie die Sammelsymptome „andere" wurden nicht berücksichtigt. Tabelle 2 gibt einen Überblick über die Kappawerte und die Übereinstimmungsprozente. Bei 29 der 139 Symptome wurde kein Kappa berechnet, da eine oder zwei Randsummen Null waren. Der Median der 110 Werte liegt bei Kappa = 0,45.

In Tabelle 3 werden die Symptome bezüglich der Interrater-Reliabilität beurteilt, indem der numerische Wert und die Signifikanz von Kappa und das Übereinstimmungsprozent berechnet werden. In knapp der Hälfte aller Symptome kann keine Aussage gemacht werden, ob die Beurteilungsübereinstimmung genügend hoch ist (vgl. Kap. 4). In Tabelle 4 sind die 70 Symptome mit der berechenbaren Interrater-Reliabilität aufgeführt.

Obwohl die Untersuchung von Gebhardt und Helmchen (1973) mit unserer direkt nicht vergleichbar ist (vgl. Kap. 1), können dennoch allgemeine Vergleiche angestellt werden. Auch bei der Berliner Studie waren nur ein Teil der Symptome analysierbar (65 der 132). In der Arbeit von Gebhardt und Helmchen haben Symptome mit hohem Q-Wert eine niedrige Interrater-Reliabilität. Von den 25 Symptomen, die bei uns eine niedrige Interrater-Reliabilität haben, sind bei der anderen Studie 18 ausgewertet worden; davon haben 15 auch in Berlin eine niedrige Interrater-Reliabilität (Q-Wert mit $P \leq 0{,}1$):

In Tabelle 4 sind Symptome, die in Berlin eine niedrige Beurteilerübereinstimmung aufwiesen, mit einem B versehen. Von den 20 Symptomen mit mittlerer Übereinstimmung wurden in Berlin 13 ausgewertet, davon 2 mit niedriger Übereinstimmung. Bei den 25 Symptomen mit guter Übereinstimmung finden wir bei den 14 in Berlin analysierten Symptomen 8 mit geringer Übereinstimmung.

Von Interesse sind also die 15 Symptome, die an beiden Orten wegen geringer Interrater-Reliabilität auffallen. Es handelt sich um: Auffassungsstörungen, Störung der Vitalgefühle, gespannt, innerlich unruhig, antriebsarm, antriebsgehemmt, eingeengt, inkohärent/zerfahren, gefühlsverarmt/affektarm, mürrisch gereizt/dysphorisch, Kontakt vermindert, ängstlich, ratlos, mißtrauisch/feindselig, affektiv inadäquat.

Aus klinischer Sicht ist von diesen 15 Symptomen mit geringer Interrater-Reliabilität die Mehrzahl sehr wichtig, und es ist schwer verständlich, daß diese Symptome eine so geringe Untersucher-Übereinstimmung aufweisen. Wichtig könnten die zum Teil nicht sehr exakten Definitionen dieser Symptome sein.

8.2. Auswertung nach Symptomklassen

Baumann et al. (1976) und Baumann und Angst (1977) haben die AMP-Symptome in folgende drei Klassen eingeteilt: unspezifisch selten, unspezifisch häufig, spezifisch (selten oder häufig). Wie zu erwarten, finden wir bei den Symptomen, über deren Güte wir keine Aussage machen können, vermehrt unspezifisch seltene Symptome ($\chi^2 = 45{,}8$; $df = 2$; $P \leq 0{,}01$) (Tabelle 5).

Vergleicht man die drei Reliabilitätsklassen von Kap. 6.1 (gut, mittel, gering) mit den zwei Symptomklassen „unspezifisch häufig" und „spezifisch" (Klasse „unspezifisch selten" wurde zu „spezifisch" gezählt, da nur ein Element enthaltend), so finden wir auch hier signifikante Zusammenhänge ($\chi^2 = 13{,}5$; $df = 2$; $P \leq 0{,}01$): Symptome mit niedriger Übereinstimmung sind eher unspezifisch häufig, während Symptome mit guter Übereinstimmung eher spezifisch sind.

Tabelle 5. Symptomklassen und Interrater-Reliabilität

Symptomklassen	Kappa	
	berechenbar	nicht beurteilbar
Unspezifisch häufig	34	5
Unspezifisch selten	1	27
Spezifisch	35	37
	70	69

Symptomklassen	Interrater-Reliabilität		
	gut	mittel	gering
Unspezifisch häufig	6	9	19
Spezifisch	19[a]	11	6

[a] inkl. 1 unspezifisch seltenes Symptom

Vergleicht man die Kappawerte der drei Symptomklassen „unspezifisch selten", „unspezifisch häufig" und „spezifisch" miteinander, so fallen 29 Symptome aus der Berechnung, da deren Kappawerte nicht berechenbar sind. Da dadurch nur noch wenige Symptome (8) der Kategorie „unspezifisch selten" übrigbleiben, sollen im folgenden nur die beiden Kategorien „spezifisch" und „unspezifisch häufig" miteinander verglichen werden. Der U-Test ergab einen auf dem 5%-Niveau signifikanten Wert ($z = 2{,}08$). Der Mittelwert der spezifischen Symptome beträgt Kappa = 0,48 ($s = 0{,}29$, $N = 63$), bei den unspezifisch häufigen Symptomen ist Kappa 0,38 ($s = 0{,}24$, $N = 39$). Damit werden die mittels χ^2-Methode erhobenen Befunde bestätigt: spezifische Symptome sind reliabler als unspezifisch häufige. Dies kann darauf beruhen, daß die Rater in ihrer Ausbildung stärker bezüglich der spezifischen Symptome geschult werden, wodurch die unspezifisch häufigen vernachlässigt werden, da sie als bekannt vorausgesetzt werden. Eventuell sind aber spezifische Symptome klarer definiert als die unspezifischen, so daß sie eindeutiger beurteilt werden können.

8.3. Auswertung nach Fremd-/Selbstbeurteilung

Woggon (in Vorbereitung) hat die AMP-Symptome nach folgenden drei Kategorien durch verschiedene Beurteiler einstufen lassen:

— Symptom ist nur über Selbstbeurteilung des Patienten einstufbar.
— Symptom ist allein über Verhaltensbeobachtung des Raters einstufbar.
— Symptom wird unter Berücksichtigung beider Informationsquellen eingestuft.

Von den 110 Symptomen (Symptom vorhanden/nicht vorhanden), deren Kappa berechenbar ist, verteilen sich 38 auf die Kategorie Selbsteinstufung, 38 auf Fremdeinstufung, und 34 sind gemischt. Die Kappawerte lassen sich mit dem H-Test von Kruskal-Wallis miteinander vergleichen, der einen H-Wert von $H = 7{,}55$ ergibt (nach χ^2 verteilt mit $df = 2$; $P \leq 0{,}05$). Die Mittelwerte lauten wie folgt:

— Selbstbeurteilung, Kappa = 0,55, $s = 0{,}28$,
— gemischte Beurteilung, Kappa = 0,42, $s = 0{,}28$,
— Fremdbeurteilung, Kappa = 0,37, $s = 0{,}27$.

Beziehen sich die Beurteiler auf Patientenaussagen (Selbstbeurteilung), so stimmen sie eher überein, als wenn sie den Patienten aufgrund ihrer eigenen Beobachtung (Fremdbeurteilung) einstufen (beim U-Test signifikant). Symptome mit gemischter Beurteilung stehen dabei in der Mitte. Dieses Ergebnis stimmt mit der Literatur überein, wonach Fremdbeurteilung bei nicht exakt definierten Kategorien ungenau ist, was auch für das AMP-System zutrifft. Übernimmt der Rater die Patientenaussagen, so ist die Übereinstimmung größer. Es bleibt dabei offen, ob zwei getrennt vorgehende Rater vom Patienten ähnliche Information bekämen; wäre dies nicht der Fall, so würde eventuell auch die Übereinstimmung bei Selbstbeurteilungssymptomen niedriger.

8.4. Auswertung nach Graduierung

Wie schon in Kap. 4.2. hingewiesen, sind die Kappa bei Berücksichtigung der Graduierung (leicht, mittel, schwer) kaum zu berechnen, da bei 48 Patienten auf die 4×4 Felder durchschnittlich nur 3 Patienten fallen. Von den 139 Symptomen waren daher nur 17 auswertbar. Den größten Kappakoeffizienten finden wir bei dem Symptom Wahnstimmung (Kappa = 0,80); es folgt dann das Symptom gehemmt (Kappa = 0,60). Die restlichen 15 Symptome sind unter 0,5.

9. Schlußbemerkungen

Die vorliegende Studie bezieht sich auf 139 AMP-Symptome; die restlichen 42 Symptome sind ohne Aussage (Kategorie andere), sehr selten (weitere körperliche Symptome) oder für Interrater-Studien wenig geeignet (neurologische Symptome). Von diesen 139 Symptomen konnten in der Studie 69 nicht beurteilt werden, davon 27 unspezifisch seltene Symptome. Diese unspezifisch seltenen Symptome sind zum größten Teil wegen ihrer Seltenheit für ein allgemeines Dokumentationssystem von geringem Interesse; zum Teil betrifft es aber Verlaufssymptome, die zur Charakterisierung von Nebenwirkungen von Psychopharmaka (Baumann und Angst, 1977) benötigt werden. Aufgrund der Analyse der 69 in dieser Studie bezüglich Interrater-Reliabilität nicht beurteilbaren Symptome wären weitere Interrater-Studien mit Patienten durchzuführen, die die betreffenden Symptome im Ausgangsbefund oder im Verlauf besser repräsentieren als die hier untersuchten Patienten.

Von den 70 beurteilten Symptomen (Einstufung vorhanden/nicht vorhanden) weisen 45 eine gute bis mittlere Übereinstimmung zwischen den Beurteilern auf. Dabei haben spezifische Symptome eine bessere Reliabilität als unspezifische, auf der Selbstbeurteilung basierende eine bessere als auf Fremdbeurteilung beruhende. Eine globale Bewertung der Ergebnisse ist in Ermangelung von Vergleichswerten schwierig. Die Ergebnisse sind aber insofern zu relativieren, als die an dieser Untersuchung beteiligten Untersucher über intensive Erfahrung mit dem AMP-System verfügten. Ungeübtere Untersucher würden vermutlich niedrigere Übereinstimmungen aufweisen. Da unsere Ergebnisse — trotz Erfahrung der Untersucher mit dem AMP-System — bei einem Drittel der Symptome unbe-

friedigend sind, muß vor einer unkritischen Anwendung des AMP-Systems gewarnt werden.

Verbesserungen sind in Vorbereitung und ließen sich wie folgt erreichen: Präzisierung der Definition von Symptomen und intensives Ratertraining, in dem besonders auf unspezifische Symptome, resp. Fremdbeurteilungssymptome zu achten wäre.

Literatur

Angst, J., Battegay, R., Bente, D., Berner, P., Broeren, W., Cornu, F., Dick, P., Engelmeier, M.-P., Heimann, H., Heinrich, K., Helmchen, H., Hippius, H., Pöldinger, W., Schmidlin, P., Schmitt, W., Weis, P.: Das Dokumentationssystem der Arbeitsgemeinschaft für Methodik und Dokumentation in der Psychiatrie (AMP). Arzneimittel-Forsch. (Drug Res.) **19**, 339—405 (1969)

Angst, J., Dinkelkamp, Th.: Die somatische Therapie der Schizophrenie. Stuttgart: Thieme 1974

Baumann, U., Schneidewind, G., Angst, J., Helmchen, H., Hippius, H.: Zur Entscheidungslogik bei Symptombeurteilungen. Arch. Psychiat. Nervenkr. **220**, 225—235 (1975)

Baumann, U., Schneidewind, G., Angst, J., Helmchen, H., Hippius, H.: Untersuchung zum AMP-System. Symptomvergleich zwischen Berlin und Zürich. Arzneimittel-Forsch. (Drug Res.) **26**, 1111—1114 (1976)

Baumann, U., Angst, J.: AMP-System: Verlaufsanalyse psychopathologischer und somatischer Symptome. Arch. Psychiat. Nervenkr. **223**, 227—238 (1977)

Baumann, U., Woggon, B.: Interraterreliabilität von AMP-Skalen (in Vorbereitung)

Busch, H., Fähndrich, E., Freudenthal, K., Renfordt, E.: Zur Anwendung und Weiterentwicklung des AMP-Systems. Bericht über ein Symposium und ein Trainingsseminar. Pharmacopsychiat. **7**, 170—175 (1975)

Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measmt. **20**, 37—46 (1960)

Cohen, J.: Weighted chi square: an extension of the kappa method. Educ. Psychol. Measmt. **32**, 61—74 (1972)

Everitt, B. C.: Moments of the statistics kappa and weighted kappa. Brit. J. Math. Statist. Psychol. **21**, 97—103 (1968)

Fleiss, J. L., Cohen, J., Everitt, B. S.: Large sample standard errors of kappa and weighted kappa. Psychol. Bull. **72**, 323—327 (1969)

Gebhardt, R., Helmchen, H.: Zur Zuverlässigkeit psychopathologischer Symptomerfassung. Schweiz. Arch. Neurol. Neurochir. und Psychiat. **112**, 459—469 (1973)

Hasemann, K.: Verhaltensbeobachtung. In: Psychologische Diagnostik — Handbuch der Psychologie (R. Heiss, Hrsg.), Bd. 6. Göttingen: Hogrefe 1964

Langer, F., Schulz von Thun, F.: Messung komplexer Merkmale in Psychologie und Pädagogik. München: Reinhardt 1974

Lienert, G. A.: Testaufbau und Testanalyse. 3. Auflage. Weinheim: Beltz 1969

Maurer-Groeli, Y.: Untersuchung zur Interraterreliabilität des AMP-Systems. Arch. Psychiat. Nervenkr. **221**, 321—330 (1976)

Maxwell, A. E.: Coefficients of agreement between observers and their interpretation. Brit. J. Psychiat. **130**, 79—83 (1977)

Mombour, W.: Verfahren zur Standardisierung des psychopathologischen Befundes, Teil 1, 2. Psychiat. Clin. **5**, 73—120 (1972)

Pichot, P. (ed.): Psychological measurement in psychopharmacology. Basel-München: Karger 1974

Scharfetter, Chr.: Das AMP-System. Manual. 2. Aufl. Berlin-Heidelberg-New York: Springer 1972

Woggon, B.: Analyse des AMP-Systems bezüglich Fremdbeurteilungs- und Selbstbeurteilungsebene der verschiedenen Symptome. (in Vorbereitung)

Eingegangen am 5. Juli 1977