

Wie zuverlässig ist operationalisierte Diagnostik? — Die Test-Retest-Reliabilität des Strukturierten Klinischen Interviews für DSM-III-R —¹

Hans-Ulrich Wittchen, Michael Zaudig, Peter Spengler,
Werner Mombour, Wolfgang Hiller, Cecilia A. Essau, Rick Rummeler,
Robert L. Spitzer* und Janet Williams*

Max-Planck-Institut für Psychiatrie, München

* New York State Psychiatric Institute, Biometric Research Department, New York

*How reliable are operationalized diagnoses? — The test-retest reliability
of the Structured Clinical Interview for DSM-III-R*

Abstract. In a test-retest design (2—3 days between examinations) the reliability of the SCID was examined for cross-sectional as well as lifetime diagnoses. 103 psychiatric inpatients were examined twice independently by 2 clinical psychologists and 3 psychiatrists. Kappa values and concordance rates were good to excellent for all diagnostic groups. Test-retest reliability Kappa values for psychotic disorders were 0.86, depressive disorders 0.86, substance abuse/dependence 0.70, bipolar disorders 0.60, anxiety disorders 0.54. Somehow lower kappa values were obtained for the diagnostic subgroups. This might be related to the problems in reliably subtyping psychotic disorders, the differential diagnosis of affective disorders with psychotic disorders and the differential diagnosis of agoraphobia and panic disorder. Possible reasons for these findings were discussed. Nevertheless it can be summarized that the SCID is a time-efficient and reliable instrument for diagnostic classification according to the criteria of DSM-III-R.

Zusammenfassung. In einer Test-Retest-Studie (2—3 Tage Abstand zwischen den Untersuchungen) wurde die Reliabilität des Strukturierten Klinischen Interviews für DSM-III-R (SKID) auf der Diagnosen-(DSM-III-R)- und der Symptomebene überprüft. Zwei Klinische Psychologen und drei Psychiater untersuchten voneinander unabhängig insgesamt 103 stationär-psychiatrische Patienten sowohl bezüglich ihrer Querschnittsdiagnose (4 Wochen) als auch aller im Verlaufe des Lebens aufgetretenen Störungen (Lifetime). Sowohl bezüglich der Lifetime- als auch der Querschnittsdiagnostik ergaben sich mit wenigen Ausnahmen gute bis sehr gute signifikante Übereinstimmungswerte auf der Symptom- und auf der Diagnoseebene. Bei einer zusammenfassenden Analyse nach DSM-III-R-Obergruppen ergaben sich im einzelnen folgende Kappa-Werte: Psychotische Störungen 0,86; Depressive Störungen 0,70; Substanzmißbrauch/-abhängigkeit 0,70; Bipolare Störungen 0,60 und Angststörungen 0,54. Die Prüfung der Reliabilität in den diagnostischen Subgruppen ergab z. T. niedrigere Übereinstimmungskoeffizienten. Probleme scheinen bei der Differentialdiagnostik einiger Psychotischer Störungen, der Abgrenzung Affektiver Störungen mit psychotischen Merkmalen von Psychotischen Störungen sowie bei der Differenzierung von Panikstörungen und Agoraphobie zu bestehen; Gründe hierfür werden auf der Ebene der diagnostischen Kriterien und den Frageformulierungen des SKID diskutiert. Insgesamt ist das SKID als verlässliches und zeitökonomisches Instrument zur Diagnostik psychischer Störungen nach DSM-III-R einzuordnen.

Anschrift der Verfasser: Prof. Dr. H.-U. Wittchen, Max-Planck-Institut für Psychiatrie, Kraepelinstraße 10,
D - 8000 München 40.

¹ Mit Unterstützung der Deutschen Forschungsgemeinschaft (DFG), Gz. DFG Wi 709/2-2. Teile der US-amerikanisch-deutschen Kollaborativstudie wurden vom National Institute of Mental Health (NIMH contract: 278-83-0007 und NIMH Grant 1 RO1 MH 40511-01) gefördert.

Einleitung

Strukturierte und halbstrukturierte Interviewverfahren blicken in der psychiatrischen, der epidemiologischen und der klinisch-psychologischen Forschung bei psychischen Störungen auf eine relativ lange Tradition zurück (z. B. Health Opinion Survey, McMillan, 1957; Health Interview Schedule, Srole, Langer, Michael, Opler & Renie, 1962; zusammenfassend Wittchen, Semler, Schramm & Spengler, 1988). Als diagnostische Instrumente im engeren Sinne, d. h. als Instrumente, aus denen direkt eine definierte Diagnose abzuleiten ist, sind sie jedoch erst in den letzten Jahren „wiederentdeckt“ worden. Dabei spielen zwei Aspekte eine wichtige Rolle: Einerseits das Bemühen um eine Verbesserung des diagnostischen Prozesses in der Klassifikation psychischer Störungen, andererseits sollen diagnostische Instrumente helfen, die durch die sog. expliziten diagnostischen Kriterien und die Operationalisierung der diagnostischen Entscheidungsfindung immer komplizierteren und umfangreicheren Regeln der klassifikatorischen Diagnostik besser zu erlernen und anzuwenden. Beide Aspekte sind vor allem vor dem Hintergrund immer schneller aufeinander folgenden Revisionen der derzeit gebräuchlichen Klassifikationssysteme DSM-III-R (American Psychiatric Association, 1987) und der in Vorbereitung befindlichen 10. Revision der Internationalen Klassifikation von Krankheiten (ICD-10) besonders relevant (Helzer, 1983; Wittchen & Schulte, 1988; Wittchen, Saß, Zaudig & Köhler, 1989).

Im Zuge dieser Entwicklung ist in den letzten 15 Jahren eine nur noch schwer zu überblickende Vielzahl neuer teil-, halb- oder vollstrukturierter bzw. sogar standardisierter Interviewansätze vorgestellt worden, die eine verlässlichere Zuordnung von Patienten im stationären und ambulanten Bereich sowie von Probanden in epidemiologischen Feldstudien ermöglichen sollen (Wittchen & von Zerssen, 1988). Das Spektrum von Interviewhilfen bei der Diagnostik psychischer Störungen läßt sich auf der *formalen Ebene* zunächst in zumindest drei Gruppen unterteilen:

(1) *Checklistenansätze*: Ihr Hauptcharakteristikum ist, daß sie keine expliziten Vorgaben enthalten, wie die Informationssammlung im einzelnen erfolgt, es werden — je nach Instrument in unterschiedlich differenzierter Form — lediglich die für die diagnostische Klassifikation relevanten Beurteilungsgesichtspunkte vorgegeben. Beispiele sind das ältere AMDP-System (Arbeitsgemeinschaft für Methodik und Dokumentation in der Psychiatrie, 1979), die DSM-III-Checkliste (Helzer, 1981) oder die kürzlich eingeführten Münchner Diagnosen Checklisten (MDCL, Hiller, Zaudig & Mombour, 1989).

(2) *Strukturierte Interviews* sind demgegenüber von ihrer Frageform, ihrer Fragenstruktur, ihrer Kodierung und Verrechnung *im Sinne eines Leitfadens vorstrukturiert*, geben aber dem Kliniker noch erheblichen Variationsspielraum. Beispiele für diesen Ansatz sind das Present State Examination (PSE, Wing, Cooper & Sartorius, 1974) bzw. seine erweiterte neueste Revision, die Schedules for Clinical Assessment in Neuropsychiatry (SCAN, Wing, in Druck) sowie das Schedule for Affective Disorders and Schizophrenia (SADS, Endicott & Spitzer, 1978).

(3) *Standardisierte Interviews* versuchen demgegenüber hinsichtlich des Untersuchungs- und Auswertungsprocedere eine *vollständige* Standardisierung und lassen somit fast gar keinen Beurteilungsspielraum des Untersuchers mehr zu. Beispiele hierfür sind das "Diagnostic Interview Schedule" (DIS, Robins, Helzer, Croughan & Ratcliff,

1981, oder das "Composite International Diagnostic Interview" (CIDI, Wittchen et al., in Druck).

Bei dieser Gruppierung ist zu berücksichtigen, daß das damit angesprochene Ausmaß der Standardisierung sich auf Formalisierungsaspekte bezieht, die der Erhöhung der Objektivität und damit auch der Reliabilität dienen und nicht etwa auf eine Normierung im Sinne der klassischen Testtheorie oder auf eine weitergehende Standardisierung sozialer Einflußgrößen in der Befragungssituation abzielen.

Auf der *inhaltlichen Ebene* unterscheiden sich fast alle angeführten Instrumente beträchtlich, so daß bei der Auswahl eines geeigneten Instruments je nach Zielsetzung und konzeptuellem Hintergrund einer Studie eine Auswahl getroffen werden sollte. Einige Instrumente sind so angelegt, daß auch klinisch wenig erfahrene Interviewer einsetzbar sind (DIS oder CIDI), andere erfordern ein hohes Ausmaß an klinisch-psychiatrischer Sachkompetenz, insbesondere wenn sie einen Schwerpunkt auf die Erfassung spezifischer psychotischer Störungen legen, wie das PSE, die MDCL oder das neue SCAN. Auch spielen konzeptuelle Unterschiede eine entscheidende Rolle. Vor allem Verfahren vor der Einführung von DSM-III im Jahre 1980 halten sich eher an den dimensionalen Ansatz, wobei zumeist Anhaltspunkte für das Vorliegen einer Diagnose oder die Zugehörigkeit zu einer diagnostischen Klasse über die Angabe des Grads der psychopathologischen Normabweichung bestimmt werden (z. B. das AMDP-System oder das PSE). Die neueren, auf den operationalisierten Diagnosekriterien, wie DSM-III aufbauenden Verfahren verfolgen demgegenüber fast immer einen eher kategorialen prototypischen Ansatz (Horowitz, Wright, Lowenstein & Parad, 1981), der einem logischen Entscheidungsbaum ähnlich ist.

Die Beurteilung der Frage, inwieweit mit den neuen diagnostischen Verfahren tatsächlich eine substantielle Erhöhung der Reliabilität der diagnostischen Entscheidungen auf der Diagnose- und Symptomebene erreicht wurde, ist durchaus problematisch. Zu unterschiedlich sind, vor allem in älteren Arbeiten, Methodik und die statistischen Analyseverfahren. Dennoch besteht zwischenzeitlich wohl kaum mehr ein Zweifel, daß die neuen diagnostischen Interviews zu einer substantiellen Verbesserung der Reliabilität klassifikatorischer Diagnostik geführt haben. Dabei mehren sich ferner die Hinweise, daß vollstandardisierte Interviews, die im Gegensatz zu lediglich strukturierten Verfahren alle Elemente des diagnostischen Prozesses festlegen (Frageform, Abfolge der Fragen, Prüffragen, Kodierung, computerisierte Auswertung), bessere Reliabilitätswerte auf der Symptom- und Diagnoseebene aufweisen als strukturierte Verfahren (Semler et al., 1987; Wittchen, Semler & von Zerssen, 1985). Die Test-Retest-Reliabilität der bislang wohl am besten untersuchten neuen diagnostischen Instrumente für DSM-III, des DIS und des CIDI liegt beispielsweise für alle erfaßten 67 Diagnosebereiche mit Kappa-Werten zwischen 0,42 und 0,84 und einer Prozentübereinstimmung von im Mittel 92 % (Range: 72 % bis 98 %) zufriedenstellend hoch (Burke, 1986; Semler et al., 1987; Wittchen et al., in Druck) und damit vor allem angesichts der Breite des von ihnen abgedeckten diagnostischen Spektrums — deutlich über den Werten älterer und neuerer strukturierter Interviewverfahren (PSE, SADS). Obwohl standardisierte diagnostische Instrumente reliabel, inhaltsvalide und ökonomisch sind, stoßen sie in klinischen Institutionen doch noch häufig auf Widerstand. Kliniker bevorzugen nach wie vor „freiere“, ihrem Erfahrungswissen mehr Spielraum eröffnende

Verfahren und somit die weniger strukturierten Interviews oder Checklistenansätze, obwohl diese nicht unbedingt zeitökonomischer oder verlässlicher sind.

Diese Ausgangssituation stimulierte 1985 das Interesse, einen Mittelweg zu versuchen und die Vorteile eines standardisierten Verfahrens mit denen strukturierter Ansätze zu verbinden. Die 1984 vor dem Abschluß stehende Revisionsarbeit an dem 1980 eingeführten DSM-III bot einen weiteren Anlaß, ein neues Interview zu konstruieren, das einerseits die verlässliche Erfassung der neuen und revidierten Kriterien von DSM-III (DSM-III-R) ermöglichen sollte und andererseits die Einwände der Kliniker bezüglich vollstandardisierter Verfahren berücksichtigt. 1984 stellten Spitzer und Williams einen solchen Entwurf in Gestalt einer ersten Vorversion des Structured Clinical Interview (SCID) für DSM-III vor, das im Rahmen einer größeren transnationalen Angststudie erprobt wurde. Stimuliert durch die ersten vielversprechenden Befunde, begann 1986 die Vorbereitung der in der vorliegenden Arbeit beschriebenen amerikanisch-deutschen Studie zur Entwicklung und Reliabilitätsprüfung des SCID. Parallel zur Erstellung der Endversion von DSM-III-R wurden in unserer Münchener und der New Yorker Arbeitsgruppe unabhängig Fragen und Procedere des Interviews erprobt und schließlich 1987 in einer größeren Test-Retest Reliabilitätsstudie getestet. Die Publikation der Gesamtergebnisse dieser Kollaborativstudie an mehr als 500 Patienten ist derzeit in Vorbereitung (Spitzer et al., in Vorbereitung²). Als Teil dieser Studie soll nun in der vorliegenden Arbeit untersucht werden, wie zuverlässig die operationalisierte Diagnostik psychischer Störungen mit diesem Ansatz ist. Dabei werden die Befunde zur Prüfung der Test-Retest-Reliabilität der deutschsprachigen Version (Strukturiertes Interview für DSM-III-R, SKID) vorgestellt. Da die DSM-III-R-Regeln zur Beurteilung vieler Diagnosekriterien Informationen aus dem Querschnitt und aus der gesamten Lebens- und Symptomvorgeschichte erfordern, soll die Reliabilität sowohl auf der Querschnittsebene (4 Wochen Zeitraum) wie auch in bezug auf vergangene Symptome und Syndrome (Lifetime) geprüft werden.

Methodik

Das Strukturierte Klinische Interview für DSM-III-R (SKID)

Das SKID wurde von Spitzer und Williams entwickelt und erlaubt nach den revidierten Kriterien von DSM-III-R (Wittchen et al., 1987) die Ableitung von 62 Diagnosen sowohl für den Querschnitt als auch für die gesamte Lebensspanne. Das Grundprinzip des SKID besteht darin, alle notwendigen Fragen und Prüfungsaspekte explizit vorzugeben, gleichzeitig aber durch den Aufbau dem klinisch erfahrenen Interviewer die Möglichkeit zu geben, bei jedem Item zusätzliche Fragen zu stellen. Erst dann, wenn der Interviewer hinreichend sicher zu sein glaubt, soll er eine Entscheidung kodieren. Mit jeder SKID-Frage ist also das mit der Frage angesprochene *diagnostische Kriterium* zu beantworten und zu kodieren. Es wird nicht die Antwort des Patienten, sondern die jeweilige Gesamtbeurteilung des klinischen Interviewers kodiert. Hieraus ergibt sich die für SKID typische Gliederung des Interviewheftes (Tabelle 1). In der linken Spalte finden sich die wörtlich

² Spitzer, R. L., Williams, J., Gibbon, M. & First, M. B. The Structured Clinical Interview for DSM-III-R (SCID) I. History and Description.
Williams, J., Gibbon, M., First, M. B., Spitzer, R. L., Davies, M., Borns, J., Kane, J., Pope, H. G., Rounsaville, B. & Wittchen, H.-U. The Structured Clinical Interview for DSM-III-R (SCID) II. Multi-site Test-retest Reliability.

Tabelle 1. Beispielseite (S. 44) des SKID aus der Sektion F: Angststörungen (F1 – F64)

F1 Hatten Sie jemals einen Angstanfall, bei dem Sie sich ganz plötzlich und unerwartet in panischen Schrecken versetzt fühlten oder panische Angst hatten?	A Eine oder mehrere Panikattacken (abgegrenzte Perioden intensiven Unbehagens oder Angst) irgendwann während der Störung, die 1) unerwartet auftraten, (d.h. traten nicht unmittelbar vor oder in einer Situation auf, die fast immer Angst hervorruft) und 2) nicht von Situationen ausgelöst wurden, in denen die Aufmerksamkeit anderer auf den Betroffenen gerichtet war	? 1 2 3 1 = Gehe zu F15 (Agoraphobie)
Wenn ja: Berichten Sie mir davon. Wann passiert das meist? (Hatten Sie auch mal völlig unerwartet einen Angstanfall/Panikattacke?)		
F2 Hatten Sie jemals <u>vier</u> solcher Anfälle in einem Zeitraum von vier Wochen?	B Entweder traten die vier Panikattacken des A-Kriteriums innerhalb eines Zeitraums von vier Wochen auf oder einer bzw. mehreren Attacken folgte ein Zeitraum von mindestens einem Monat, in dem anhaltende Angst vor einer erneuten Attacke bestand.	? 1 2 3 1 = Gehe zu F15 Agoraphobie
Wenn nein: Hatten Sie große Angst vor einem erneuten (Anfall)? (Wie lange hatten Sie Angst?)		
F3 Wann trat der letzte schwere (Anfall) auf? (erwartet oder unerwartet?)	Zeitpunkt: _____	
F4 Nun werde ich Ihnen einige Fragen über diese Angstanfall stellen. Während dieser Attacke Hatten Sie Atemnot? (Hatten Sie Schwierigkeiten durchzuatmen?)	C Wenigstens vier der folgenden Symptome traten bei mindestens einer der Attacken auf:	
Fühlten Sie sich benommen, unsicher oder einer Ohnmacht nahe?	1) Atemnot (Dyspnoe) oder Beklemmungsgefühle 2) Benommenheit, Gefühl der Unsicherheit oder Ohnmachtsgefühl	? 1 2 3 ? 1 2 3
Hatten Sie Herzrasen, Herzklopfen oder Herzstolpern?	3) Palpationen oder beschleunigte Herzfrequenz (Tachykardie)	? 1 2 3
Hatten Sie gezittert oder gebebt?	4) Zittern oder Beben	? 1 2 3
Haben Sie geschwitzt?	5) Schwitzen	? 1 2 3
Hatten Sie das Gefühl zu ersticken?	6) Erstickungsgefühle	? 1 2 3
War es Ihnen übel oder hatten Sie das Gefühl, Durchfall zu bekommen?	7) Übelkeit oder abdominelle Beschwerden	? 1 2 3
Erschien Ihnen alles (oder Teile des Körpers) unwirklich/wie entfremdet?	8) Depersonalisation	? 1 2 3
Hatten Sie ein Kribbeln oder Taubheit in Teilen Ihres Körpers?	9) Taubheit oder Kribbelgefühl (Parästhesie)	? 1 2 3
Hatten Sie Hitzewallungen oder Kälteschauer?	10) Hitzewallungen oder Kälteschauer	? 1 2 3
Hatten Sie Schmerzen oder ein Engegefühl in der Brust?	11) Schmerzen oder Unwohlsein in der Brust	? 1 2 3
Hatten Sie Angst zu sterben?	12) Furcht zu sterben	? 1 2 3
Hatten Sie Angst, verrückt zu werden oder die Kontrolle zu verlieren?	13) Furcht verrückt zu werden oder etwas Unkontrolliertes zu tun	? 1 2 3
Mindestens 4 der Kriterien C1 - 13) sind erfüllt und mit 3 kodiert (Beachte Attacken mit weniger als 4 Symptomen werden im Rahmen der Agoraphobie ohne Panikstörung (F 4) kodiert)		1 3 1 = Gehe zu F15 Agoraphobie

zu stellenden Fragen, in der Mitte das zu beurteilende Kriterium und in der rechten Spalte sind die Kodierungen und sich daraus ergebende Sprungentscheidungen zu ersehen.

Der Beurteilungsspielraum des Kliniklers wird allerdings durch eine Reihe von weiteren Besonderheiten begrenzt. Hierzu gehört, daß das SKID mit einem grob vorstrukturierten traditionell klinisch aufgebauten Explorationsteil beginnt, dessen Informationen im Verlauf des Interviews im Zusammenhang mit Merkmalskodierungen und zeitlichen Zuordnungen verwendet werden sollen.

Der klinischen Tradition entsprechend, sind auch die Beurteilungsdimensionen der Symptombkodierung angelegt: ? = unsicher / zu wenig Informationen, 1 = nicht vorhanden / nein, 2 = vorhanden, nicht kriteriumsgemäß ausgeprägt, 3 = sicher vorhanden und kriteriumsgemäß. Das SKID ist, den Regeln von DSM-III-R entsprechend, hierarchisch, einem Entscheidungsbaum ähnlich, aufgebaut. Vielfache Sprungregeln erhöhen die Effizienz, können allerdings bei Fehlentscheidungen zum unrichtigen Überspringen ganzer Diagnosebereiche führen und reduzieren somit möglicherweise die Itemübereinstimmung. Die Diagnosen werden vom Interviewer im Verlaufe des Interviews sukzessiv gestellt und am Ende des Interviews auf einen Diagnosenbogen übertragen. Das SKID ermöglicht die Diagnostik von insgesamt 62 DSM-III-R-Diagnosen, die ohne Angabe der zusätzlichen Subtypen in Tabelle 2 wiedergegeben sind.

Tabelle 2. *DSM-III-R Achse I Diagnosen des SKID nach der Sektionsgliederung A bis I
(Die diagnostischen Untergruppen für Affektive,
Psychotische und Angststörungen sind weiter aufgeschlüsselt)*

A./D. Affektive Störungen

Bipolare Störung*
Bipolare Störung NNB
Major Depression*
Dysthymie (nur derzeitige)*
Depressive Störung NNB

B./C. Psychotische Störungen

Schizophrenie*
Schizophrenieforme Störung*
Schizoaffective Störung*
Wahnhafte Störung
Kurze reaktive Psychose*
Psychotische Störung NNB

E. Störungen durch psychotrope Substanzen

Alkohol
Sedativa, Schlafmittel, Anxiolytika
Cannabis
Stimulantien
Opiate
Kokain
Halluzinogene-PCP
Polytoxikomanie
Andere

F. Angststörungen

Panikstörung*
Agoraphobie*
Soziale Phobie
Einfache Phobie
Zwangsstörung
Generalisierte Angststörung*

G. Somatoforme Störungen

Somatisierungsstörung (nur derzeitige)
Hypochondrie (nur derzeitige)
Undifferenzierte Somatoforme Störung

H. Eßstörungen

Anorexia Nervosa
Bulimia Nervosa

I. Anpassungsstörungen

(nur derzeitige)

Zusätzlich zu den diagnostischen Entscheidungen hinsichtlich des Vorliegens derzeitiger und vergangener psychischer Störungen gemäß der Achse I von DSM-III-R ermöglicht das SKID ferner: (a) die Kodierung weiterer klinischer Variablen wie Alter bei Beginn der Störung, Dauer der Störung, Persistenz, Remissionsgrad sowie Schwere der psychosozialen Einschränkungen, (b) die Beurteilung der derzeitigen psychosozialen Einschränkungen auf einer Ratingskala gemäß der Achse V von DSM-III-R, (c) Kodierungsmöglichkeiten für körperliche Erkrankungen (Achse III) sowie (d) die Diagnostik von Persönlichkeitsstörungen in einem gesonderten Interview- und Fragebogenteil (Achse II). Diese optimalen Ergänzungen werden in einer gesonderten Arbeit referiert.

Patienten

Untersucht wurden im Rahmen des hier referierten Münchener Untersuchungsteils der Kollaborativstudie insgesamt 107 Patienten. 82 von ihnen waren neu stationär aufgenommene Patienten des Max-Planck-Institut für Psychiatrie der geschlossenen (überwiegend Suizidgefährdete und schwere psychotische Störungen) und der offenen Station (überwiegend Neurosen und Persönlichkeitsstörungen), 25 waren nach vergleichbaren Kriterien stationär aufgenommene psychiatrische Patienten des McLean Hospital in Boston. Die in Boston untersuchten Patienten wurden von zwei der Autoren, HUW und MZ, in englischer Sprache untersucht. Alle erfüllten folgende Einschlusskriterien: (a) Alter 20 bis 50 Jahre, (b) keine ausgeprägten hirnrnorganischen Auffälligkeiten oder Anfallsleiden, (c) schriftliches Einverständnis, an beiden Interviewterminen teilzunehmen, (d) Untersuchung in den ersten 2 Wochen nach Aufnahme. Die Auswahl der Patienten wurde von den zuständigen Stationsärzten vorgenommen. Zwei der 82 ausgewählten Patienten in der Münchener Gruppe verweigerten das Retest-Interview, zwei weitere waren zum Zeitpunkt des Test- oder Retest-Interviews akut psychotisch und konnten nicht regelrecht untersucht werden. Damit verbleiben N = 103 Patienten für die Auswertung.

Das Durchschnittsalter der untersuchten 103 Patienten betrug 34,0 Jahre (Standardabweichung 13,0); 61 waren weiblichen, 42 männlichen Geschlechts. 28 % waren verheiratet, 10 % geschieden, 2 % getrennt oder verwitwet, 59 % waren noch nie verheiratet. Die soziale Schichtverteilung wurde nach Hollingshead und Readlich wie folgt beurteilt: Oberschicht 4 %, Mittelschicht 63 %, Unterschicht 25 %. Einige Patienten in der Bostoner Gruppe (8 %) konnten wegen Fehlens von Information nicht beurteilt werden. Nur 41 der 103 Patienten waren erstmalig hospitalisiert, wiesen zum Teil mehrere stationäre Aufnahmen in ihrer Vorgeschichte auf (Durchschnitt: 2,3 stationäre Behandlungen).

Design

Bei der einmal pro Woche stattfindenden Fallkonferenz der Stationen wurden durch den Stationsarzt alle potentiellen Interviewpatienten benannt. Ein Forschungsmitarbeiter übernahm dann zusammen mit dem behandelnden Arzt die Unterrichtung des Patienten und holte die Einverständniserklärung ein. Insgesamt 7 Patienten verweigerten diese, weitere 11 konnten aus stationstechnischen Gründen (Therapie, andere Diagnosestudien) nicht teilnehmen. Der Einweisungsgrund und die im Überweisungsschein aufgeführten Hauptbeschwerden wurden vom Forschungsmitarbeiter auf ein Vorinformationsblatt übertragen, wobei alle diagnostischen Termini unberücksichtigt blieben. Der Forschungsmitarbeiter teilte dann jeden Patienten nach einem vorbestimmten Schema einem Interviewer-Paar zu, um sicherzustellen, daß alle 5 Untersucher in etwa die gleiche Anzahl von SKID-Interviews durchführten.

Alle 103 Patienten wurden zweimal von jeweils verschiedenen Interviewern mit dem SKID befragt. Die Patienten wurden über die Gründe informiert, warum sie mit dem gleichen Untersuchungsinstrument zwei Mal untersucht wurden und wurden angehalten, die beiden Interviews als unabhängig voneinander zu betrachten.

Um Varianzquellen aufgrund möglicher Veränderungen des psychopathologischen Zustandes des Patienten auszuschalten, wurde der Abstand zwischen dem Test- und dem Retest-Interview mit 2—3 Tagen kurz gehalten. Die Interviews dauerten im Durchschnitt 75 Minuten. Jedes Interview, sowohl das Test- als auch das Retest-Interview, wurde mit Video aufgenommen, um so in späteren Auswertungsschritten Abweichungsquellen gezielt analysieren zu können.

Interview

Das SKID ist ein Interview für Kliniker und erfordert sowohl Kenntnis des DSM-III-R Manuals als auch klinische Erfahrung in der Beurteilung psychopathologischer Symptome. Die SKID Interviews wurden deshalb von einem erfahrenen Psychiater (WM) und 2 am Ende ihrer Facharztzeit stehenden Psychiater sowie von einem psychopathologisch erfahrenen Klinischen Psychologen (HUW) und einem Klinischen Psychologen mit einjähriger Psychiatrieerfahrung (WH) durchgeführt. HUW, MZ und RR wurden im Rahmen eines einwöchigen Trainingskurses in den USA durch die Originalautoren in Gebrauch des SCID trainiert, die übrigen wurden nachträglich in München in das SCID eingeübt. Alle Interviewer führten vor Beginn der Studie fünf SKID Interviews unter Videokontrolle durch, die in der Gruppe diskutiert wurden.

Auswertung

Zur Bestimmung der Übereinstimmung benutzten wir drei verschiedene Maße: die prozentuale Übereinstimmung, den Kappa-Koeffizienten (Cohen, 1960) und den Y-Koeffizienten (Yule, 1912). Prozentübereinstimmung und Kappa-Koeffizient gehören zu den am häufigsten verwendeten Reliabilitätsmaßen (Spitznagel & Helzer, 1985; Zubin, 1967). Zusätzlich zur statistischen Signifikanzprüfung für Kappa-Werte, benutzen wir die Konvention von Bartko und Carpenter (1976), Kappa-Werte zwischen 0,50–0,70 als akzeptabel zu interpretieren, und Werte größer als 0,70 als sehr gute Übereinstimmung zu bezeichnen (Burke, 1986). Die gleiche Konvention gilt für den Yule-Koeffizient. Der Y-Koeffizient ist dem Kappa-Koeffizienten ähnlich, jedoch vollständig unabhängig von der Grundrate (Aufretenshäufigkeit des untersuchten Merkmals). Wenn eine einzelne Zelle der Häufigkeitstabelle Null ist, wird es problematisch, den Y-Koeffizient zu interpretieren, weil er den Endpunkt seines Wertebereiches erreicht (-1 oder +1). Dies wurde mit der Pseudo-Bayes-Methode (Bishop, Feinberg & Holland, 1975) ausgeglichen. Wir empfehlen bei einer Grundrate von kleiner als 10 % (d. h. in der vorliegenden Arbeit, weniger als 10 Fälle mit dem jeweiligen Merkmal) den Konventionen Spitznagels folgend eher den Y-Wert zu interpretieren und den Kappa-Wert unberücksichtigt zu lassen. Eine besondere Problematik bei der Auswertung ist die Handhabung der Sprungregeln. Da die Items z. T. nicht unabhängig sind, sondern ihre Beantwortung in einigen Sektionen von der Beantwortung von Screening-Fragen abhängt, ergeben sich bei der Untersuchung der Itemreliabilität einerseits Verletzungen einzelner Voraussetzungen der Auswertungsverfahren, andererseits auswertungspraktische Probleme. Wir haben uns trotzdem entschieden, die Itemübereinstimmungen zu berechnen und haben im Falle von Sprunginstruktionen die übersprungenen Items als nicht vorhanden mitberücksichtigt.

Ergebnisse

Test-Retest-Reliabilität für DSM-III-R Lifetime- und Querschnittsdiagnose für Diagnose-Gruppen

Sowohl hinsichtlich Lifetime wie auch im 4-Wochen-Querschnitt ergaben sich in der Anzahl gestellter Diagnosen keine signifikanten Unterschiede zwischen Test- und Retestinterview. Im Testinterview wurden im Mittel 2,1 Lifetime- (Querschnitt 1,3) im Retest-Interview 2,2 (Querschnitt: 1,4) Lifetime-Diagnosen je Patient gestellt. Ein Abfall der Diagnosenhäufigkeit vom Test zum Retest-Interview wurde nicht festgestellt. Drei Patienten erhielten übereinstimmend keine psychiatrische Diagnose. Mehr als 60 % aller Patienten erfüllten im Verlaufe ihres Lebens die Kriterien für mehr als eine DSM-III-R-Diagnose, im Querschnitt wiesen mehr als 40 % mehr als eine Diagnose auf. Tabelle 3 zeigt die Übereinstimmungsbefunde zwischen Test- und Retestinterview (Prozentübereinstimmung, Kappa-Koeffizient, Yule-Koeffizient) der Lifetime- und

Tabelle 3. *Test-Retest Reliabilität der SKID-Diagnosengruppen (Lifetime und 4 Wochen Querschnitt)*

DSM-III-R-Diagnosen	Lifetime				Derzeit			
	NB	%	K	Y	NB	%	K	Y
Depressive Störungen	57	84,5	0,70***	—	38	86,4	0,68***	—
Bipolare Störungen	34	86,4	0,66***	—	26	86,4	0,55***	0,63
Psychotische Störungen	36	94,2	0,86***	0,90	34	94,2	0,86***	0,88
Substanzmißb./abhängigkeit	34	87,4	0,70***	0,82	11	98,1	0,89***	0,82
Angststörungen	43	78,6	0,54***	0,90	28	89,3	0,69***	0,74
Somatoforme Störungen	4	94,2	0,22***	0,67	— nur Lifetime Diagnosen —			
Eßstörungen	13	97,1	0,87***	0,92	10	97,1	0,81***	0,89
Anpassungsstörungen	— nur Querschnitt —				5	96,0	0,32*	0,70

Anmerkungen: * $p < .05$; *** $p < .001$; NB = Grundrate; % = Prozentübereinstimmung; K = Kappa-Koeffizient; Y = Yule-Koeffizient; — nicht berechenbar (Anzahl positiver Diagnosen)

Querschnittsdiagnosen für DSM-III-R-Störungsgruppen. Die Prozentübereinstimmung für Lifetimedialdiagnosen liegt zwischen 78,6 % (Angststörungen) und 94,2 % (Psychotische Störungen und Somatoforme Störungen), die Kappa-Werte sind alle signifikant mit den niedrigsten Übereinstimmungswerten für akute Anpassungsstörungen (0,32), Angststörungen (0,54 Lifetime), Bipolare Störungen (0,55 derzeit) und Somatoforme Störungen (0,22). Hierbei ist allerdings zu berücksichtigen, daß für Somatoforme und Anpassungsstörungen nur 4 Fälle in der Stichprobe enthalten waren, so daß der Kappa-Wert nur mit Einschränkungen zu interpretieren ist. Der in diesen Fällen verlässlichere Y-Wert ist in beiden Fällen mit 0,67 für Somatoforme Störungen und 0,70 für Anpassungsstörungen befriedigend.

Test-Retest-Reliabilität diagnostischer Subgruppen

Um zu kleine Fallgruppen in den diagnostischen Untergruppen zu vermeiden, wurde die Auswertung der diagnostischen Subgruppen lediglich für die Lifetime-Diagnosen durchgeführt. Somatoforme und Eßstörungen sowie die Untertypen für Abhängigkeit und Mißbrauch werden wegen der geringen Fallzahlen für diese Störungen nicht berücksichtigt.

Angststörungen

Wie schon in der Tabelle 3 angedeutet, ergeben sich in der Klassifikation von Angststörungen eine Reihe von Nicht-Übereinstimmungen. Insbesondere die Differential-

diagnose von Panikstörung und Agoraphobie fällt mit einer relativ hohen Anzahl von Abweichungen zwischen Test- und Retestinterview aus dem Rahmen (vgl. Tabelle 4). Eine genauere Analyse der Kodierungen von Test- und Retestinterview auf der Symptomebene ergibt, daß die Interviewer bei einer zentral wichtigen Frage, und zwar der, ob die Person jemals einen plötzlichen Angstanfall erlebte, häufig nicht übereinstimmen. Mit einer Ausnahme sind alle Abweichungen auf diese eine Frage zurückzuführen. Da die Antwort auf diese eine Frage im SKID mit einem Sprungbefehl zur Agoraphobie verknüpft ist, wird bei Verneinen der Panikattacke nur noch die Diagnose Agoraphobie möglich. Eine versuchsweise zusammenfassende Übereinstimmungsanalyse Panikstörung und/oder Agoraphobie in Gegenüberstellung zu dem Nicht-Vorliegen einer oder beider Störungen ergibt deshalb auch ein höchst zufriedenstellendes Kappa von 0,91. Hier scheint also sowohl ein Problem im diagnostischen Klassifikationssystem wie auch in der Frageformulierung und Strukturierung des SKID zu liegen.

Tabelle 4. *Test-Retest Reliabilität der Subgruppen von Angststörungen (Lifetime)*

DSM-III-R-Diagnosen	N	%	K	Y	1. Int.		2. Int.-		
					-	+	-	+	
						A	B	C	D
Panikstörung	103	85,4	0,27	0,62	84	6	9	4	
Agoraphobie	103	89,3	0,36*	0,56	88	4	7	4	
Soziale Phobie	103	89,3	0,50***	0,67	85	3	8	7	
Einfache Phobie	103	90,3	0,56***	0,68	85	4	6	8	
Zwangsstörung	103	95,1	0,71***	0,85	91	4	1	7	
Generalisierte Angststörung	103	99,0	0,80***	0,86 #	100	1	0	2	

Anmerkungen. * $p < .05$; *** $p < .001$; # Pseudo-Bayes Schätzung von Y: % Prozentübereinstimmung; K = Kappa-Koeffizient; Y = Yule-Koeffizient.

Nur ein mittelhoher Kappa-Wert findet sich auch für Soziale Phobien, hier ergeben sich die meisten Abweichungen bei der Beurteilung, ob die Symptomatik auf eine der vorher diagnostizierten psychischen Störungen (Affektive, Psychotische, Substanz-

abhängigkeiten) zurückzuführen ist. In derartigen Fällen sollte die Diagnose ausgeschlossen werden. Ohne Berücksichtigung dieser Kriteriumsfrage steigt der Kappawert auf 0,86. Die Reliabilität der übrigen Angstdiagnosen ist trotz z. T. kleiner Fallzahlen als zufriedenstellend bis gut zu bezeichnen.

Psychotische Störungen

Die Reliabilität differentialdiagnostischer Entscheidungen bei psychotischen Störungen ist (Tabelle 5) zumindest bezüglich schizoaffektiver Störungen und der Kategorie „Nicht Näher Bezeichnete Psychotische Störungen (NNB)“ problematisch. Eine genauere Analyse der nicht übereinstimmend klassifizierten Patienten ergibt, daß alle Abweichungen auf Fälle zurückzuführen sind, die neben eindeutigen psychotischen Symptomen auch die Kriterien für ein depressives oder manisches Syndrom erfüllen. In diesen Fällen lassen sich im SKID zwei Hauptursachen für abweichende Beurteilungen identifizieren. Zum einen bereitet die Beurteilung des DSM-III-R Kriteriums „C“ für Schizophrenie Schwierigkeiten. Bei diesem Kriterium muß der Interviewer das Zeitverhältnis „Dauer affektive Syndrome“ zur Gesamtdauer florider bzw. ggf. residualer psychotischer Symptomatik beurteilen. Mit den drei Kodierungsoptionen „unsicher“, „affektiv länger als psychotisch“ und „psychotisch länger als affektiv“ entscheidet sich

Tabelle 5. *Test-Retest Reliabilität der Subgruppen Psychotischer Störungen (Lifetime)*

DSM-III-R-Diagnosen	N	%	K	Y	2. Int.-	1. Int.	
						-	+
						A	B
						+ C	D
Schizophrenie	102	89,2	0,56***	0,69		82	8
						3	9
Schizophrenieforme Störung	103	97,1	0,56*	0,73 #		98	0
						3	2
Schizoaffektive Störung	102	95,1	-0,02	—		97	4
						1	0
Wahnhafte Störung	101	98,0	0,79***	0,90		95	1
						1	4
Psychotische Störung NNB	103	92,2	0,29	0,65		93	3
						5	2

Anmerkungen. * $p < .05$; *** $p < .001$; # Pseudo-Bayes Schätzung von Y; % Prozentübereinstimmung; K = Kappa-Koeffizient; Y = Yule-Koeffizient; - nicht berechenbar; NNB = Nicht Näher Bezeichnet

hier die Differentialdiagnose für Schizophrenie, Psychotische Störungen NNB und Affektive Störungen mit psychotischen Merkmalen. Offensichtlich besteht hier ein zu großer Beurteilungsspielraum für den Interviewer, denn die Übereinstimmung der Interviewer bei dieser Frage ist gering. In vier Fällen beurteilte ein Interviewer eine kurze Dauer affektiver Syndrome, oder beurteilte die affektiven Symptome als Ausdruck einer psychotischen Residualsymptomatik und kam in der Folge zur Diagnose Schizophrenie, während der andere Interviewer eine längere Dauer affektiver Symptomatik kodierte und im weiteren zur Diagnose einer Schizoaffektiven Störung kam. Weitere 5 Abweichungen ergaben sich dadurch, daß einer der beiden Interviewer die Kodierung „unsicher“ vornahm. In diesen Fällen ist nur noch die Diagnose Psychotische Störung NNB möglich.

Trotz dieser problematischen Situation, ist die Reliabilität für Schizophrenie, Schizophrenieforme und Wahnhafte Störungen mit Kappawerten zwischen 0.56 und 0.79 zumindest als akzeptabel zu beurteilen.

Affektive Störungen

Mit Ausnahme der Residualkategorien (NNB) ist die Test-Reliabilität für die meisten Subgruppen Affektiver Störungen als gut zu bezeichnen (Tabelle 6). Als Hauptursache für Abweichungen ergab sich wiederum die Zuordnung psychotischer Auffälligkeiten

Tabelle 6. *Test-Retest Reliabilität der Subgruppen Affektiver Störungen (Lifetime)*

DSM-III-R- Diagnosen	N	%	K	Y	2. Int.-	1. Int.	
						-	+
						A	B
						C	D
Major depression	102	85,3	0,69***	0,70		55 9	6 32
Bipolare Störung	103	92,2	0,70***	0,78		83 4	4 12
Dysthyme Störung	102	99,0	0,80***	0,86		99 1	0 2
Depressive Störung NNB	103	89,3	0,29	0,50		89 6	5 3
Bipolare Störung NNB	103	87,4	0,36*	0,52		85 6	7 5

Anmerkungen. * $p < .05$; *** $p < .001$; % = Prozentübereinstimmung; K = Kappa-Koeffizient; Y = Yule-Koeffizient; - nicht berechenbar; NNB = Nicht Näher Bezeichnet

im oben diskutierten „C“ Kriterium der Schizophrenie. Acht der Abweichungen bei Bipolaren Störungen NNB bzw. Depressiver Störung NNB ergaben sich dadurch, daß jeweils nur ein Interviewpartner zur Entscheidung einer Bipolaren Störung bzw. Major Depression mit stimmungsinkongruenten psychotischen Merkmalen kam. Vier Abweichungen ergaben sich dadurch, daß einer der Interviewer eine psychotische Störung (2 Fälle Schizoaffektive Störung, 2 Fälle Schizophrenie, Residualer Typus) ohne eine gleichzeitige Affektive Störung NNB diagnostizierte, während der andere Interviewpartner der depressiven Störung eine eigenständige Bedeutung zuwies und zusätzlich zur psychotischen Störung noch eine Affektive Störung NNB kodierte.

Substanzmißbrauch/-abhängigkeit

Wie in Tabelle 7 dargestellt, ist die Übereinstimmung bei Substanzmißbrauch/-abhängigkeit als gut zu bezeichnen. Wegen der geringen Grundrate (NB) in den einzelnen Substanzgruppen sollte eher der Y-Wert zur Beurteilung herangezogen werden.

Tabelle 7. *Test-Retest Reliabilität der Subgruppen von Substanzmißbrauch/-abhängigkeit Lifetime)*

DSM-III-R-Diagnosen	NB	%	K	Y
Alkoholmißb./abhängigkeit	23	94,1	0,75***	0,82
Alkoholmißbrauch	17	88,2	0,68***	0,60
Alkoholabhängigkeit	5	80,0	—	—
Drogenmißb./abhängigkeit	13	92,2	0,75***	0,82
Sedativa	8	96,0	0,82***	0,90
Cannabis	4	94,1	0,38	0,90
Stimulantien	5	96,1	0,41	0,90
Opiate	1	98,0	0,33	0,90
Halluzinogene — PCP	2	99,0	0,80***	0,86
Polytoxikomanie	3	98,7	0,60*	0,75
Andere	3	97,1	-0,01	0,75

Anmerkungen. * $p < .05$; *** $p < .001$; — nicht berechenbar; NB = Grundrate (Anzahl positiver Diagnosen); % = Prozentübereinstimmung; K = Kappa-Koeffizient; Y = Yule-Koeffizient.

Test-Retest-Reliabilität der SKID-Symptomfragen

Die Verteilung der Kappa-Werte für die Symptomübereinstimmung zeigt z. T. erhebliche Unterschiede zwischen den einzelnen SKID-Diagnosebereichen. Die Befunde sind

mittels eines "Boxplot" (Abbildung 1) dargestellt (Tukey, 1977). Innerhalb der Box finden sich 50 % der Werte. Der Median ist durch einen Querstrich innerhalb der Box dargestellt. Die Längsstriche an den beiden Enden der Boxen geben die Streuung der verbleibenden Werte an. In unserer Analyse des "Box-plots" werden nur Items mit einer Grundrate über 10 % in betracht gezogen. Die Anzahl der jeweils berücksichtigten SKID-Fragen in jeder Sektion ist in Klammern angegeben. Die höchste Übereinstimmung findet sich bezüglich der Symptome für Abhängigkeit und Mißbrauch von Alkohol und anderen psychotropen Substanzen, depressive Symptome sowie die hier aus Gründen der Übersichtlichkeit nicht dargestellten Bereiche für Somatoforme Störungen (Median Kappa: 0,71), Eßstörungen (Median Kappa: 0,74) und Anpassungsstörungen (Median Kappa: 0,68). Aber auch die Mehrzahl der Items aus der SKID-Sektion für psychotische Störungen sowie die Soziale Phobie weist eine gute Test-Retest-Reliabilität auf. Demgegenüber liegen die Übereinstimmungswerte für Items aus dem

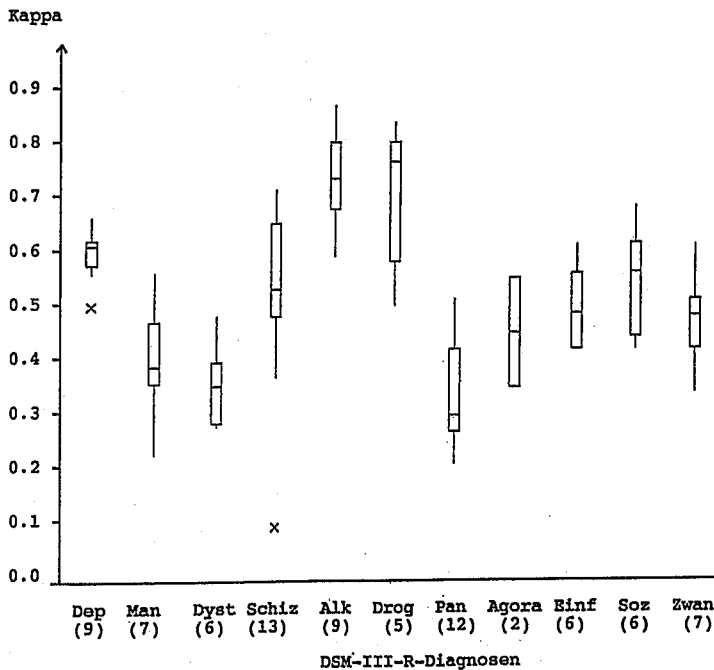


Abbildung 1. Boxplot-Darstellung der Kappa-Werte zur Reliabilität der SKID-Symptomfragen.

Die Boxplots geben die Werteverteilung der entsprechenden Variable an. Innerhalb der Box finden sich 50 % der Werte, der Median ist durch den Markierungsstrich in der Box wiedergegeben. Die Linien unterhalb und oberhalb der Box geben die Streuung für das erste bzw. dritte Quartal an. Die Kreuze markieren Ausreißer.

Dep = depressives Syndrom, Man = manisches Syndrom, Dyst = Dysthymie, Schiz = Schizophrenie, Alk = Alkoholmißbrauch /-abhängigkeit, Pan = Panikstörung, Agora = Agoraphobia, Einf = Einfache Phobie, Soz = Soziale Phobie, Zwan = Zwangsstörung, () = Anzahl der jeweils berücksichtigten SKID-Symptomfragen in jeder Sektion.

Bereich Manisches Syndrom, Dysthyme Störung, aber auch Panikstörungen und Agoraphobie deutlich niedriger.

Die Kappa-Werte der Items aus dem Syndrombereich Manie liegen fast ausnahmslos niedrig zwischen 0,23 („Übermäßige Aktivitäten mit möglicherweise negativen Konsequenzen“) und 0,56 („Vermindertes Schlafbedürfnis“). Der Median liegt bei 0,39. Die Hauptursache für die Abweichungen liegt in der getrennten Erfassung dieser Symptome für derzeitige und vergangene Episoden (ein Interviewer beurteilt die Symptomatik als noch gegenwärtig vorhanden, d. h. in den letzten 4 Wochen, der andere Interviewer beurteilt die Symptomatik als mehr als 4 Wochen zurückliegend). Weniger häufig (3 Fälle) finden sich Abweichungen bereits in der Screeningfrage, ob überhaupt schon einmal manische oder hypomanische Symptome aufgetreten sind.

Im Bereich Dysthymie finden sich am häufigsten Abweichungen in der Beantwortung der Screening-Frage, die eine zeitliche Beurteilung erfordert, ob jemals eine Phase von 2 Jahren oder mehr bestand, in der der Patient durchgängig depressiv war. Infolge des darauf folgenden Sprungbefehls bleiben bei Verneinen die 6 Folgeitems unkodiert. Deshalb liegt die Übereinstimmung mit Kappa-Werten zwischen 0,28 und 0,48 und einem Median von 0,35 niedrig. Die Items mit der niedrigsten Übereinstimmung sind „Hoffnungslosigkeit“ (0,28) und „schlechte Konzentrationsfähigkeit“ (0,29). Das Item mit der höchsten Übereinstimmung ist „Schlaflosigkeit oder vermehrter Schlaf“ (0,48).

Gründe für die herabgesetzte Reliabilität im Bereich Panikstörung und Agoraphobie wurden bereits oben diskutiert. Wiederum ist der Sprungbefehl nach der Screeningfrage ausschlaggebend. Verneint der Interviewer diese Eingangsfrage, springt er zur Agoraphobie und überspringt damit die Beurteilung aller panikbezogenen Symptomfragen; beurteilt er jedoch ein Paniksyndrom, überspringt er andererseits die Beurteilung der Sektion für Agoraphobie ohne Panikstörung in der Vorgeschichte. Die Kappa-Werte liegen zwischen 0,21 („Hitzewallungen oder Kälteschauer“) und 0,51 („Benommenheit“) mit einem Median von 0,30. Wie in der Box dargestellt, liegen 50 % von allen Kappa-Werten zwischen 0,25 und 0,41. Relativ niedrige Kappa-Werte erhalten wir bei den Items „Schwitzen“ (0,27), „Übelkeit“ (0,29), „Depersonalisation“ (0,29), „Hitzewallungen“ (0,21), und „Furcht zu sterben“ (0,23).

Diskussion

Unsere Ergebnisse zeigen zusammenfassend, daß die Mehrzahl der untersuchten SKID/DSM-III-R-Diagnosen eine befriedigende Test-Retest-Reliabilität aufweisen. Legen wir einen Kappa-Wert von 0,50 als akzeptable Übereinstimmung fest, so ergeben sich lediglich für fünf diagnostische Kategorien unbefriedigende Übereinstimmungswerte: Affektive Störungen NNB, Schizoaffective Störung, Psychotische Störung NNB, Panikstörung und Agoraphobie.

Als Erklärung für die eingeschränkte Reliabilität dieser fünf Diagnosen lassen sich verschiedene Aspekte anführen:

(1) Design und Stichprobencharakteristika: Wir haben in unserer Untersuchung schwerer gestörte, hospitalisierungsbedürftige psychiatrische Patienten untersucht.

Mehr als 80 % der untersuchten Patienten weisen die Hauptdiagnose Affektive und Psychotische Störungen auf, mit einem Überwiegen von oft langjährigen komplizierten Erkrankungen, bei denen sich psychotische und affektive Symptome mischen. Der hohe Anteil schwerer Affektiver und Psychotischer Erkrankungen ist einerseits typisch für die Gesamtsituation stationär psychiatrischer Krankenhäuser, andererseits ist zu berücksichtigen, daß zum Zeitpunkt der Untersuchung Patienten mit affektiven und psychotischen Symptomen ein spezieller Forschungsschwerpunkt des MPI-P waren. Somatoforme, Anpassungs-, Angst- und Eßstörungen sowie Abhängigkeiten von psychotropen Substanzen wurden in unserer Studie mit 6 Ausnahmen nur als Zweit- oder Dritt-diagnosen gestellt. Es ist zu erwarten, daß bei der Untersuchung primärer Angststörungen wesentlich bessere Übereinstimmungswerte erzielt werden. Zu berücksichtigen ist ferner, daß die Variabilität durch Einbeziehung der englischsprachigen Interviews möglicherweise erhöht wurde.

(2) Das SKID versucht nicht nur — ungleich der meisten psychiatrisch-diagnostischen Interviews — Lifetime- und Querschnittsdiagnosen abzuleiten, sondern auch möglichst hierarchiefrei alle möglichen Zusatzdiagnosen zu erheben. Die von uns gewählte Strategie, alle gestellten Diagnosen in die Reliabilitätsauswertung einzubeziehen und nicht nur die Hauptdiagnose, könnte zu einem insgesamt herabgesetzten Test-Retest-Reliabilitätswert führen. In einer für die Hauptdiagnose und alle gestellten Diagnosen getrennt durchgeführten Analyse der diagnostischen Reliabilität fanden Spitzer et al. (siehe Fußnote Seite 139) z. B. wesentlich höhere Übereinstimmungswerte. So betrug die Test-Retest-Reliabilität (Kappawerte) für Hauptdiagnosen in unserem Datensatz 0,79 (anstatt 0,56 in der vorliegenden Analyse) für Schizophrenie, 0,81 (statt 0,69) für Major Depression, 0,62 (statt 0,27) für Panikstörung und 0,78 (statt 0,36) für Agoraphobie. Interessant ist aber der Hinweis, daß die Reliabilität der Lifetime-Diagnosen insgesamt nicht schlechter als die der Querschnittsdiagnosen ist.

(3) Als ein weiteres methodisches Problem sollte noch einmal auf die Problematik der Grundratenabhängigkeit des Kappa-Wertes und speziell auf insgesamt die zur Prüfung aller diagnostischen Untergruppen zu geringe Fallzahl unserer Studie hingewiesen werden. Der von uns zur teilweisen Lösung des Grundratendilemmas angebotene Y-Wert ergibt zwar in den Fällen mit der Grundratenproblematik substantiell höhere Übereinstimmungskoeffizienten, ohne aber die beschriebenen unzureichenden Reliabilitätsbefunde für die 5 Diagnosen erklären zu können.

(4) Als vierte und wichtigste Varianzquelle sind das Instrument selbst bzw. die diagnostischen Kriterien anzuführen. Die meisten Abweichungen, sowohl auf der Item- wie auch auf der Diagnoseebene, lassen sich auf einige wenige, offensichtlich noch unzureichend explizit formulierte DSM-III-R-Kriterien bzw. die entsprechenden Fragen im SKID zurückführen. Hierzu gehören z. B. die Beurteilung des „C“-Kriteriums der Schizophrenie (Zeitverhältnis, Dauer affektiver Symptomatik gegenüber Zeitdauer psychotischer Symptomatik) sowie die neue Hierarchieregel für Panikstörungen, die die Diagnose Agoraphobie. Darüber hinaus bedarf das SKID offensichtlich ergänzender Anweisungen, wie zu verfahren ist, wenn eine affektive Störung vorhanden ist, die nur noch zum Teil in das 4-Wochen-Querschnittsbild hineinfällt. Häufig sind Patienten im Verlauf der letzten 4 Wochen schon teilweise remittiert, so daß die SKID-Querschnitts-

beurteilung bei der Symptomeinschätzung der Schwere der Störung ein verzerrtes Bild ergibt.

Der Vergleich unserer Befunde mit denen anderer Test-Retest-Untersuchungen ist, wie eingangs diskutiert wurde, nicht unproblematisch; nur wenige Studien haben mit ähnlicher Methodik Lifetime- und Querschnittsdiagnosen nach DSM-III-R erstellt. Nur zwei Untersuchungen haben Vorfassungen von DSM-III-R untersucht. Verglichen mit den Befunden von Semler et al. (1987) an einem ähnlichen Patientengut, liegen unsere Kappa-Werte etwas höher. Dies trifft besonders für den Bereich affektive Störungen zu. Mit einem standardisierten psychiatrischen Interview, dem CIDI, erzielten Semler et al. (1987) beispielsweise Kappa-Werte von 0,66 (SKID: 0,69) für Major Depression, 0,47 (SKID: 0,70) für Bipolare Störung, und 0,47 (SKID: 0,80) für Dysthymie. Bessere Test-Retest-Reliabilitäten wurden auch für die meisten Angststörungen, allerdings mit Ausnahme der Panikstörung und Agoraphobie gefunden, die in der Arbeit von Semler sehr gute Kappawerte aufwiesen (Panikstörung 0,84, Agoraphobie 0,65). Eine zweite Untersuchung von Mannuzza et al. (1989) mit dem SADS, die sich allerdings auf die Untersuchung von Angststörungen beschränkte, ergab folgende Kappawerte: Panikstörung 0,67, Agoraphobie 0,81, Einfache Phobie 0,31, Soziale Phobie 0,68. Diese Befunde zeigen, daß zumindest im Angstbereich grundsätzliche Verbesserungen in Hinblick auf eine verlässlichere Erfassung von Panikstörung und Agoraphobie möglich sind. Aus dem Vergleich mit dieser Studie kann jedoch mit methodischen Vorbehalten geschlossen werden, daß der Versuch des SKID, einen für den Kliniker eher akzeptablen Mittelweg zwischen Standardisierung und Checkliste zu gehen, durchaus als vielversprechend bezeichnet werden darf.

Literatur

- American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders* (3rd. edn., revised). Washington, DC: APA. (Deutsche Bearbeitung: Wittchen, H.-U., Saß, H., Zaudig, M. & Kochler, K. (1989). Diagnostisches und Statistisches Manual psychischer Störungen (DSM-III-R Revision. Weinheim und Basel: Beltz).
- Arbeitsgemeinschaft für Methodik und Dokumentation in der Psychiatrie (Hrsg.) (1979). *Das AMDP-System. Manual zur Dokumentation psychiatrischer Befunde* (3. Aufl.). Heidelberg: Springer-Verlag.
- Bartko, J. J. & Carpenter, W. T. (1976). On the methods and theory of reliability. *Journal of Nervous and Mental Disease*, 163, 307—317.
- Bishop, Y. M. M., Feinberg, S. E. & Holland, P. W. (1975). *Discrete multi-variate analysis: Theory and practice*. Cambridge: MIR Press.
- Burke, J. D. (1986). Diagnostic categorization by the Diagnostic Interview Schedule (DIS): A comparison with other methods of assessment. In J. Barrett & R. M. Rose (Eds.), *Mental disorders in the community* (pp. 255—285). New York: The Guilford Press.
- Cohen, J. (1960). A coefficient of agreement of nominal scales. *Educational Psychological Medicine*, 20, 37—46.
- Endicott, J. & Spitzer, R. L. (1978). A diagnostic interview: The schedule for Affective Disorders and Schizophrenia. *Archives of General Psychiatry*, 35, 837—844.
- Helzer, J. E. (1981). The use of a structured diagnostic interview for routine psychiatric evaluations. *Journal of Nervous and Mental Disease*, 169, 45—49.
- Helzer, J. E. (1983). Standardized interviews in psychiatry. *Psychiatric Developments*, 2, 161—178.
- Hiller, W., Zaudig, M. & Mombour, W. (1989). *MDCL — Münchner Diagnosen Checklisten für DSM-III-R*. München: Logomed.

- Horowitz, L. M., Wright, J. C., Lowenstein, E. & Parad, H. W. (1981). The prototype as a construct in abnormal psychology: 1. a method for deriving prototypes. *Journal of Abnormal Psychology*, 90, 568—574.
- MacMillan, A. M. (1957). The Health Opinion Survey: Techniques for estimating prevalence of psychoneurotic and related types of disorders in communities. *Psychological Reports*, 3, 325—339.
- Mannuzza, S., Fyer, A. J., Martin, L. Y., Gallops, M. S., Endicott, J., Gorman, J., Liebowitz, M. R. & Klein, D. F. (1989). Reliability of anxiety assessment: I. Diagnostic agreement. *Archives of General Psychiatry*, 46, 1093—1101.
- Robins, L. N., Helzer, J. E., Croughan, J. & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: its history, characteristic and validity. *Archives of General Psychiatry*, 38, 381—389.
- Semler, G., Wittchen, H.-U., Joschke, K., Zaudig, M., Geiso, T. v., Kaiser, S., Cranach, M. von & Pfister, H. (1987). Test-Retest reliability of a standardized psychiatric interview (DIS/CIDI). *European Archives of Psychiatry and Neurological Sciences*, 236, 214—222.
- Spitznagel, E. L. & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry*, 42, 725—728.
- Stole, L., Langer, T. S., Michael, S. T., Opler, M. K. & Renie, T. A. C. (1962). *Mental health in the metropolis: the Midtown Manhattan Study* (Vol. 1). New York: McGraw Hill.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley.
- Wing, J. K. (in Druck). Schedules for Clinical Assessment in Neuropsychiatry — SCAN, PSE-10. Part I. Genf: World Health Organisation.
- Wing, J. K., Cooper, J. E. & Sartorius, N. (1974). *Measurement and classification of psychiatry symptoms*. Cambridge: Cambridge University Press.
- Wittchen, H.-U., Semler, G. & Zerksen, D. von (1985). A comparison of two diagnostic methods — clinical ICD diagnoses vs DSM-III and Research Diagnostic Criteria using the Diagnostic Interview Schedule (Version 2). *Archives of General Psychiatry*, 42, 677—684.
- Wittchen, H.-U., Zaudig, M., Schramm, E., Spengler, P., Mombour, W., Klug, J. & Horn, R. (1987). *Strukturiertes Klinisches Interview für DSM-III-R (SKID) (Testversion)*. Weinheim: Beltz.
- Wittchen, H.-U. & Schulte, D. (1988). Diagnostische Kriterien und operationalisierte Diagnosen. Grundlagen der Klassifikation psychischer Störungen. *Diagnostica*, 34, (1), 3—27.
- Wittchen, H.-U., Semler, G., Schramm, E. & Spengler, P. (1988). Diagnostik psychischer Störungen mit strukturierten und standardisierten Interviews: Konzepte und Vorgehensweisen. *Diagnostica*, 34, (1), 58—84.
- Wittchen, H.-U. & Zerksen, D. von (Hrsg.) (1988). *Verläufe behandelter und unbehandelter Depressionen und Angststörungen*. Heidelberg: Springer-Verlag.
- Wittchen, H. U., Burke, J. D., Semler, G., Pfister, H., Cranach, M. von & Zaudig, M. (1989). Recall and dating reliability of psychiatric symptoms. *Archives of General Psychiatry*, 46, 437—443.
- Wittchen, H.-U., Saß, H., Zaudig, M. & Koehler, K. (1989). Von DSM-III zu DSM-III-R. Erfahrungen und Perspektiven. In American Psychiatric Association (Ed.), *Diagnostisches und Statistisches Manual Psychischer Störungen DSM-III-R Revision* (S. 10—12). Weinheim und Basel: Beltz.
- Wittchen, H.-U., Robins, L. N., Cottler, L., Sartorius, N., Altamura, A. C., Andrews, G., Dingemans, R., Droux, A., Essau, C. A., Farmer, A., Halikas, J., Ingebringsten, G., Isaac, M., Jenkins, P., Kueche, G. E., Krause, J., Lepine, J. P., Lyketsos, G., Maier, W., Miranda, C. T., Smeets, R., Pfister, H., Pull, C., Rubio-Stipec, M., Sandanger, I., Tacchini, G., Teherani, M. (in Druck). *Interrater Reliability of the Composite International Diagnostic Interview (CIDI) — Results of the Multicenter WHO/ADAMHA Field Trials (Wave I)* (Proceedings of the VIIIth World Congress of Psychiatry, Athens, GR) Amsterdam: Elsevier.
- Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of The Royal Statistical Society*, 75, 581—642.
- Zubin, J. (1967). Classification of the behavior disorders. *Annual Review of Psychology*, 18, 373—401.