Die Klassifikation psychischer Störungen nach DSM-IV mit Hilfe eines strukturierten diagnostischen Interviews (F-DIPS) –

Eine Untersuchung der Retest-Reliabilität und der Validität

Inaugural-Dissertation im Fach Psychologie
zur Erlangung der Doktorwürde
der Fakultät für Sozial- und Verhaltenswissenschaften der
Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Dipl. Psych. Andrea Keller, Dresden

2000

Dekan: Prof. Dr. M. Brumlik

1. Gutachter: Prof. Dr. P. Fiedler

2. Gutachter: Prof. Dr. J. Margraf (Basel)

Danksagung

Meinen Dank an alle, die dazu beigetragen haben, dass diese Arbeit ihre jetzige Form erhalten konnte. Dies gilt insbesondere für meine beiden Betreuer, Herrn Prof. Dr. J. Margraf, der mir das Thema zur Verfügung gestellt hat, und Herrn Prof. Dr. Peter Fiedler, der die Betreuung spontan übernommen hat. Frau Dr. Eni Becker und Herr Dr. Wolfgang Deppe unterstützten mich durch ihre kritische Durchsicht. Alle haben mir wertvolle Anregungen und die notwendige positive Verstärkung gegeben, um an der Arbeit dran zu bleiben. Herrn Prof. Dr. Peter Joraschky möchte ich danken, dass er mich im letzten Stress der Fertigstellung beruflich entlastet und mir die Kooperationsmöglichkeit mit der Klinik in Berggießhübel vermittelt hat. Den Interviewerinnen und Interviewern Melanie Merswolken, Jana Mrose, Katja Lämmerhirt, Harald Gebhardt, Claudia Hille, Dennis Scholz, Yvonne Hofmann, Susan Seyfert und Herrn Dr. Rainer Niethammer aus Heidelberg danke ich für ihr Engagement und Herrn Chefarzt Dr. Höll und Herrn Dr. Roth aus der Median-Klinik Berggießhübel sowie Herrn Prof. Dr. Ch. Mundt der Psychiatrischen Universitätsklinik Heidelberg für die unkomplizierte Kooperation.
Diese Arbeit wurde finanziell unterstützt vom Forschungsverbund Public Health, was eine große Hilfe besonders dafür war, als die Interviewer größere räumliche Entfernungen zurücklegen mussten, um ein Interview führen zu können.

Mein Dank gilt auch allen Patientinnen und Patienten, die an den Interviews teilnahmen und sich geduldig zweimal die gleichen Fragen stellen ließen sowie denjenigen, die an der Schulung der Interviewer teilgenommen haben.

Inhalt

1 Einleitung

„In der Psychologie sowohl wie in der Psychopathologie besteht die Tatsache, daß man nur wenig Behauptungen, ja vielleicht keine Behauptung aufstellen kann, die nicht irgendwie und irgendwo bestritten wird. .... Es ist schon viel, wenn zwei Forscher sich über die Methode einig sind und sich nur über einen mit ihr gewonnenen Befund in einer dann immer fruchtbaren Weise streiten.“ (Jaspers, 1973, 9. Aufl., S. 5)

Die Uneinigkeit zwischen Therapieschulen und unter Psychiatern und Psychologen bzgl. der Einschätzung von Patienten sowie bzgl. der daraus folgenden optimalen Behandlung trug lange Zeit dazu bei, die Diagnostik psychischer Störungen in Misskredit zu bringen. Aufgrund der Befürchtung, mit einer Diagnose mehr Unheil anzurichten als Nutzen, wurde teilweise lieber ganz auf Diagnosen verzichtet. Dies führte dazu, dass sich Psychotherapie schlecht in ihrer Qualität und in ihrem Erfolg überprüfen ließ und dass letztendlich Psychotherapie als Behandlungsmethode, besonders innerhalb der Medizin wenig ernst genommen wurde.

Einen konstruktiveren Versuch, auf die unzureichende Übereinstimmung zwischen mehreren Diagnostikern zu reagieren, als ganz auf Diagnosen zu verzichten, stellt die Operationalisierung der Diagnostik psychischer Störungen durch die Entwicklung bzw. Weiterentwicklung von Klassifikationssystemen dar. Um zudem die systematische Anwendung dieser Klassifikationssysteme zu ermöglichen, wurden Interviewmanuale wie z.B. das „Diagnostische Interview bei psychischen Störungen (DIPS)“ oder das Nachfolgermodell F-DIPS in der Forschungsversion entwickelt.

Aber inwieweit mit Hilfe dieses strukturierten Interviews (F-DIPS) wirklich eine Verbesserung der diagnostischen Übereinstimmungen zwischen zwei Interviewern erreicht werden kann und inwieweit diese Diagnosen zutreffend sind, soll in der vorliegenden Arbeit untersucht werden.

Hierzu wurden unselektiert 191 Patientinnen und Patienten aus der Klinik für Psychotherapie und Psychosomatik am Universitätsklinikum Dresden, der Median Klinik Berggießhübel, der Psychiatrischen

Universitätsklinik Heidelberg und aus dem Projekt „Prädiktoren psychischer Gesundheit junger Frauen in Dresden“ zwei Mal von unabhängigen Ratern interviewt und parallel dazu mit Selbstbeurteilungs-Fragebögen untersucht.

Die Arbeit ist unterteilt in einen theoretischen Teil, der aus Kapitel 1 bis 6 besteht, in einen empirischen (Kapitel 7) und einen Ergebnis- und Diskussionsteil (Kapitel 8 – 10).
Kapitel 2 gibt zunächst einen Überblick über die Probleme bei der Diagnostik psychischer Störungen in den letzten 150 Jahren und Versuche der Verbesserung und stellt die aktuell gültigen Klassifikationssysteme vor. In Kapitel 3 werden die Kriterien dargestellt, anhand derer die Güte eines diagnostischen Instruments beurteilt werden kann. Kapitel 4 referiert wichtige Studien zur Güte von psychiatrischer Diagnostik und von Klassifikationssystemen. Danach gibt Kapitel 5 einen Einblick in verschiedene Formen diagnostischer Interviews und deren Güte und stellt das F-DIPS vor. Nach der detaillierten Darlegung der Fragestellung in Kapitel 6 folgt in Kapitel 7 die Vorstellung des Untersuchungsablaufs und -vorgehens. In Kapitel 8 erfolgt die Ergebnisdarstellung der Untersuchung, Kapitel 9 umfasst die Diskussion und Vorschläge zur Weiterentwicklung des diagnostischen Interviews.

2 Die Diagnostik psychischer Störungen

In diesem Kapitel stehen die Besonderheiten und Schwierigkeiten der Diagnostik psychischer Störungen im Vordergrund, sowie Versuche zu deren Bewältigung durch die Fortentwicklung von Klassifikationen bis hin zu ICD-10 und DSM-IV.

Eine psychische Störung wird im allgemeinen definiert durch persönliches Leid, durch die Abweichung von der Norm, die Funktionseinschränkung bzw. die Behinderung sowie durch Selbst- oder Fremdgefährdung (Fydrich, 1997). Dabei ist diese Definition bereits schwierig, da es, besonders bei Persönlichkeitsstörungen, vorkommen kann, dass eine Störung vorliegt, ein Patient aber kein Leid hinsichtlich seiner Verhaltensmuster empfindet, was mit der Ich-Syntonie der Persönlichkeitsstörungen zu tun hat (Fiedler, 1994), die dadurch gekennzeichnet ist, dass der Betroffene die anderen Personen auffallenden Abweichungen eher als normal und zu sich gehörig ansieht. Das Kriterium des Leids ist subjektiv, die Betroffenen entscheiden selbst, wie sehr sie leiden und inwieweit sie darüber berichten. Auch die Abweichung von der Norm mit dem statistischen Kriterium der Normalverteilung, die für jede einzelne Eigenschaft die Populationsmehrheit in der Mitte ansiedelt, ist problematisch, da die Abnormität damit als Seltenheit definiert wird. Wenn ein Merkmal selten in der Gesellschaft ist, heißt dies jedoch noch nicht, dass es auch als abnorm angesehen wird (z.B. besondere sportliche Fähigkeiten, vgl. Davison & Neale, 1996). Zudem ist Abnormität von den Normen und Werten der jeweiligen Gesellschaft abhängig.

Der Punkt Funktionseinschränkungen oder Behinderung durch die Symptomatik ist ebenfalls ein Faktor, der von der Umgebung abhängig ist. Beispielsweise kann es bei einer spezifischen Phobie (z.B. Hundephobie) vorkommen, dass zwar alle Angst-Symptome und Vermeidungsverhalten vorhanden sind, aber, da nur ein kleiner Teilbereich Probleme bereitet, keine besonderen Funktionseinschränkungen bestehen, solange die betroffene Person in einer Umgebung ohne Angst auslösenden Reiz (Hunde) lebt.

Um diesen Punkt eindeutiger zu machen, gilt eine Symptomatik laut DSM-IV (Diagnostisches und Statistisches Manual psychischer Störungen) erst dann als psychische Störung, wenn eine bedeutsame Beeinträchtigung oder Leiden in sozialen, beruflichen oder anderen wichtigen Funktionsbereichen vorhanden sind (Saß et al., 1996). Hiermit soll die Schwelle für nicht-pathologische Formen von Symptomen höher gesetzt werden. Auch in der ICD-10 (Internationale Klassifikation psychischer Störungen) wird der Begriff „Störung“ verwendet, um einen klinisch erkennbaren Komplex von Symptomen oder Verhaltensauffälligkeiten zu umschreiben. Eine Störung sollte danach immer auf der individuellen und oft auch auf der Gruppen- oder sozialen Ebene mit Belastung und mit Beeinträchtigung von Funktionen verbunden sein. Eine alleinige Beeinträchtigung auf der sozialen Ebene reicht nach dieser Definition nicht aus (Dilling et al., 1991).
Von Relevanz ist die allgemeine Definition der psychischen Störung insofern, dass die Beurteilung, ob Probleme das Ausmaß einer „Störung“ innehaben, auch entscheidet, ob Behandlungsbedarf besteht und ob die Behandlungskosten von Krankenkassen oder Rentenversicherungsträgern übernommen werden.

2.1 Einteilung und Beschreibung von Merkmalen

In der psychiatrischen Klassifikation wird der Begriff der Störung benutzt im Gegensatz zum Begriff der Krankheit, der eher dann gebraucht wird, wenn es für Störungskomplexe eine spezifische, gemeinsame Ätiologie gibt und ein einheitliches Ansprechen auf eine bestimmte Therapie angenommen wird (Stieglitz, 2000). Als Klassifikation bezeichnet man die Einteilung eines Merkmals nach verschiedenen Kriterien wie beispielsweise groß oder klein, hoch oder niedrig, dumm oder schlau. Solche Einteilungen geschehen aus bestimmten Wertsystemen heraus, in denen festgelegt wird, ab welcher Höhe z.B. ein Gebäude hoch ist, und kann sich über ein Jahrhundert hinweg gewaltig verändern.

Auch die Klassifikation psychischer Störungen veränderte sich erheblich in den letzten 50 Jahren durch wachsendes Störungswissen und den Wunsch, eindeutiger klassifizieren zu können. Dies führte zu immer differenzierteren Klassifikationssystemen mit immer mehr Diagnosen.

Die Klassifikation psychischer Störungen kann auf der Grundlage einzelner Symptome (symptomatologisch), anhand von Gruppen gemeinsam auftretender Symptome (syndromatologisch) oder die Krankheitslehre betreffend (nosologisch) erfolgen.

Nach Helmchen (1975) können Störungen nach folgenden Dimensionen und Kriterien klassifiziert werden:

Dimension	Kriterium
Symptomatologie	◆ Art der Symptome ◆ Konfiguration von Symptomgruppen bzw. Syndromen
Zeit (Verlauf)	◆ Erkrankungsalter ◆ Tempo des Erkrankungsbeginns (Akuität) ◆ Verlauf (intermittierend, chronisch) ◆ Dauer ◆ Ausgang
Ätiologie	◆ Disposition, genetisch ◆ Disposition, Persönlichkeitsstruktur ◆ Auslösung, psychoreaktiv ◆ Auslösung, somatisch ◆ Auslösung, therapeutisch ◆ Verlaufsbeeinflussung, morbogen ◆ Verlaufsbeeinflussung, psychoreaktiv ◆ Verlaufsbeeinflussung, sozial ◆ Verlaufsbeeinflussung, therapeutisch
Intensität	der meisten Kriterien auf den ersten drei genannten Dimensionen
Sicherheit	◆ der Merkmalseinschätzungen auf den ersten drei genannten Dimensionen ◆ der verbalen Diagnose ◆ der kodierten Diagnose

Die präzise Festlegung von Merkmalen, die einzelne *Klassen* definieren bei Beibehaltung des Einteilungsgrundes (kategoriale oder auch klassifikatorische Systeme), ist dann sinnvoll, wenn alle Mitglieder einer diagnostischen Klasse homogen sind, wenn klare Grenzen zwischen den

Kategorien existieren und die vorhandenen Klassen einander vollständig ausschließen. Dem kategorialen Ansatz liegt die Annahme zugrunde, dass es sinnvolle Gruppierungen der beobachteten Merkmale gibt, die überzufällig häufig gemeinsam auftreten und somit eine Störung kennzeichnen, die sich von einer anderen abgrenzen lässt (Margraf, 1996).
Den fließenden Übergängen zwischen den verschiedenen Klassen, wie z.B. zwischen Panikstörung und Hypochondrie oder zwischen Panikstörung und Somatisierungsstörung, könnte die Formulierung von *Typen* aufgrund ihrer eigenen Randunschärfe am ehesten gerecht werden, da diese eine Abstraktion von realen Gegebenheiten darstellt (Möller, 1998). Dabei sollen prototypische Merkmale benannt werden, die für das jeweilige Störungsbild als besondere Markierungspunkte gelten (Fiedler, 1994). So könnte man (hypothetisch) formulieren: Ein Mensch mit einer Panikstörung wechselt die Symptome selten und ist in seiner Körpersymptomatik eindeutiger auf Herz-bezogene Symptome fixiert als ein Mensch mit einer Somatisierungsstörung, der ein vielfältigeres Bild mit jahrelang wechselnder Symptomatik zeigt.

Neben der Einteilung in Klassen oder Typen gibt es auch die Möglichkeit, psychische Störungen auf *dimensionaler Basis* einzuschätzen.
Im Gegensatz zum kategorialen Ansatz wird beim dimensionalen Ansatz kein Kriterium festgelegt, das die Störung als gegeben belegt, sondern es wird davon ausgegangen, dass die feststellbaren Unterschiede in bezug auf die Merkmale vor allem quantitativer Natur und kontinuierlich verteilt sind. Am Beispiel der Angst und Depression könnte ein zweidimensionales System mit Angst und Depression formuliert werden (vgl. Margraf, 1996): Sind die Symptome folgendermaßen verteilt: Niedergeschlagene Stimmung, Interesseverlust, gestörter Schlaf, Herzrasen, Angst, Konzentrationsstörung, kann die Depressionsausprägung als hoch, die Angstausprägung als niedrig eingestuft werden, ohne eine Trennung zwischen den Diagnosegruppen vollziehen zu müssen. Ein Beispiel für den dimensionalen Ansatz ist das dreidimensionale Modell von Eysenck, bestehend aus den Dimensionen Extraversion, Neurotizismus und Psychotizismus. Eysenck (1986) wirft der DSM-Diagnostik ei-

nen „kategorialen Irrtum“ vor und zweifelt kategoriale Unterschiede zwischen Gruppen an.
Besonders in der Diagnostik von Persönlichkeitsstörungen wird die Forderung nach Abkehr von der kategorialen Diagnostik zugunsten einer dimensionalen Einschätzung entsprechender Persönlichkeitsmerkmale diskutiert (Fydrich et al., 1997).

Unter *nosologischer Zuordnung* wird die eindeutige, logische Zuordnung in ein einheitliches und logisches System der Krankheiten verstanden. Dabei werden außer der Symptomatik auch Verlauf, Ätiologie, Pathogenese und das Ansprechen auf therapeutische Maßnahmen mit einbezogen. Am Beispiel des Ulcus ventriculi, dem Magengeschwür, kann das Prinzip der nosologischen Zuordnung dargestellt werden: Das Ulcus ventriculi ist durch die Art der Gewebsschädigung (Histologie) einerseits und durch die Lokalisation andererseits definiert. Symptomatisch liegen Magenschmerzen, Erbrechen oder Blut im Stuhl vor, die sich bei Gabe von säurehemmenden Medikamenten verbessern.
Das momentane Wissen über psychische Störungen ist jedoch noch lange nicht umfassend genug, um ein vollständiges System der Krankheiten zu schaffen (Wittchen, 1994a). Somit existieren im Bereich der Nosologie besonders viele unterschiedliche Klassifikationsversuche. Basis ist häufig noch die von Kraepelin klinisch intuitiv entwickelte Klassifikation nach ursächlichen Faktoren von meist hypothetischer Natur (Möller, 1998).

In der Diagnostik psychischer Störungen haben sich besonders in jüngerer Zeit kategoriale Systeme mit syndromatologischer (mehrere häufig gemeinsam vorkommende Symptome) Klassifikation durchgesetzt. Diese Entwicklung entstand vor allem aufgrund von Reliabilitätsproblemen bei der klinisch-psychiatrischen Klassifikation (s.u.), aber auch durch die Pharmakotherapie, die eher syndromorientiert als nosologisch orientiert vorgeht (Möller, 1998).

2.2 Historische Entwicklung der Klassifikation psychischer Störungen

Die formale Klassifikation psychischer Störungen begann mit Philippe Pinel (1745 – 1826) nach dem Vorbild der biologischen Klassifikationen. Er unterschied zwischen Melancholie und Manie mit und ohne Delirium, zwischen Demenz und Idiotie.

Von Möbius (1892) wurde Ende des 19. Jahrhunderts in der Tradition der Degenerationslehre der Endogenitätsbegriff eingeführt (Mundt, 1991). Dieser Begriff sollte psychische Störungen kennzeichnen, die weder als Reaktion auf ungünstige Entwicklungsbedingungen und aktuelle Belastungen noch als Folge körperlicher Störungen erklärt werden konnten. Emil Kraepelin (1856 – 1926) ging von der Annahme aus, dass psychische Störungen ebenso wie körperliche Krankheiten somatische Ursachen haben und vor allem anhand ihrer Symptome klassifiziert werden sollten. Er entwickelte ein triadisches nosologisches System der organischen Störungen, der abnormen Variationen des Seelenlebens und der endogenen Psychosen. Darin sind endogene Psychosen durch das Zusammentreffen von Psychotizität und Indirektheit der Somatopathogenese definiert.

Ernst Kretschmer (1888 – 1964) bezeichnete das „manisch-depressive Irresein“ und die Schizophrenie als erbkonstitutionell bedingte endogene Psychosen. Karl Jaspers (1883 – 1969) begründete vor dem 1. Weltkrieg die Psychopathologie als Wissenschaft mit eigenem Forschungsgegenstand, eigener Methodik und kritischem Methodenbewusstsein. Er diskutierte die Frage „Entwicklung einer Persönlichkeit“ oder „Prozess“ für die Zuordnung „abnormer seelischer Phänomene“ (Janzarik, 1974). Die Tradition der deskriptiven Schule mit ätiologischer und pathogenetischer Konnotation wurde auch bei Kurt Schneider (1887 – 1967) beibehalten, wobei mehr Wert auf das Spezifische der psychopathologischen Phänomene der Psychosen gelegt wurde. Er postulierte die Einteilung in „Zyklothymie und Schizophrenie“ und außerdem in verschiedene „psychopathische Persönlichkeiten“, in „abnorme Erlebnisreaktionen“, „Schwachsinn“ und „körperlich begründbare Psychosen“ (Schneider,

1987). Eine besondere vererbbare Persönlichkeitsstruktur bildete für ihn oft die Voraussetzung für psychoreaktive Störungen.

Seit den 70er Jahren ist ein Versuch der nosologischen Klassifikation, der sich im Endogenitätsbegriff ausdrückt (z.B. ob die Depression eher reaktiv auf ein äußeres Ereignis eintritt oder endogen begründet ist, also einem „inneren Ablauf“ folgt bzw. kein äußerer Auslöser zu finden ist), umstritten (Mundt, 1991) und wird zugunsten von Begriffen ersetzt, von denen eine geringere theoretische Vorannahme erwartet wird.
Die mangelnde Übereinstimmung von psychiatrischen Diagnosen tat das ihre dazu, dass die Operationalisierbarkeit von Diagnosen in den Mittelpunkt der diagnostischen Bemühungen trat und damit die Entwicklung von detaillierteren Klassifikationssystemen mit der Unterteilung in immer neue Störungsgruppen, die damit besser beforscht werden können. Die aktuell vorliegenden gebräuchlichen Klassifikationssysteme sind die International Classifikation of Diseases der Weltgesundheitsorganisation ICD-10 (1994) und das Diagnostic and Statistical Manual of Mental Disorders der American Psychological Association DSM-IV (1994).
In der Forschung ist es inzwischen kaum noch möglich, Arbeiten in Fachzeitschriften zu publizieren, ohne Diagnosen mit Hilfe eines aktuellen Klassifikationssystems abgesichert zu haben (Stieglitz & Freyberger, 1999b).

Die folgende Tabelle gibt einen Überblick über die Entwicklung von Klassifikationssystemen:

Zeit	Diagnostik	Möglichkeiten
1840	Volkszählung in den USA	Kategorie für Schwachsinn /Wahnsinn
1880	Volkszählung in den USA	7 Kategorien psychischer Erkrankungen: Manie, Melancholie, Monomanie, Parese, Demenz, Dipsomanie und Epilepsie
1853	International Classification of Diseases (ICD)	Ohne psychische Störungen

Zeit	Diagnostik	Möglichkeiten
1917	Klassifikationssystem der „American Medico-Psychological Association	22 Kategorien, auf Prinzipien Kraepelins basierend
1933	Standard Classified Nomenclature of Diseases	24 Kategorien mit 82 Untergruppen
1948	ICD-6	Erste offizielle Klassifikation der WHO mit einem Kapitel über psychische Störungen: 10 Kategorien für Psychosen, 9 für Psychoneurosen und 7 für Charakterstörungen, Verhaltensstörungen und Störungen der Intelligenz
1952	DSM-I (Diagnostic and Statistical Manual of Mental Disorders der APA)	Schwerpunkt auf klinische Anwendung. Einfluss der psychobiologischen Sicht Adolf Meyers, der annahm, dass psychische Störungen die Reaktion eines Individuums auf psychische, soziale und biologische Faktoren sind
1965	ICD-8	Erweiterung um neue Krankheitsgruppen; internationale Kooperation bei der Entwicklung
1968	DSM-II	Traditionell intrapsychische Sichtweise
1975	ICD-9	Ohne diagnostische Kriterien, ohne multiaxiales System. Klassifikation in „Organische Psychosen, Andere Psychosen, Neurosen, Persönlichkeitsstörungen und andere nichtpsychotische psychische Störungen (z.B. Abhängigkeit, funktionelle Störungen, depressive Reaktion etc.)“
1980	DSM-III	Definitionen für einzelne Störungen, multiaxiale Klassifikation mit Feldstudien vor der Einführung
1987	DSM-III-R	Einführung des Komorbiditätsprinzips
1992	ICD-10	Klinisch-diagnostische Leitlinien
1994	ICD-10	Forschungskriterien
1994	DSM-IV	Erfassung von 395 Störungen ist möglich. Auf der Basis empirischer Resultate entwickelt

Nach diesen allgemeinen Entwicklungen in der Klassifikation psychischer Störungen, die aktuell weg von ätiologischen Modellen hin zu ausdifferenzierteren Störungsgruppen geht, ist nun von Relevanz, wie die Klassifikationssysteme den Anforderungen im Forschungs- und klinischen Gebrauch gerecht werden.

2.3 Ziele und Probleme einer Klassifikation psychischer Störungen

Die Klassifikation psychischer Störungen erfolgt, um subjektive und laienhafte Beschwerdeschilderungen einer Person in ein Bewertungsschema zu überführen und um damit Diagnosen zu gewinnen. Margraf (1996) geht aufgrund empirischer Untersuchungen in der Klinischen Psychologie und aus der Sozialpsychologie davon aus, dass auch ohne eine explizite Klassifikation ständig Annahmen zur Erklärung von geschilderten Problemen gebildet und Hypothesen-konforme Informationen aktiv gesucht werden. Widersprechende Informationen werden ignoriert, Beurteilungsfehler können leichter auftreten (Halo-Effekt, logische Fehler, s.u.). Durch die explizite Klassifikation besteht im Gegensatz zur impliziten die Möglichkeit der Überprüfbarkeit der Regeln, nach denen klassifiziert wird, was Entscheidungen weniger willkürlich machen soll.

Die Güte der Klassifikation beruht auf der Erfassung geeigneter Merkmale für die verschiedenen Störungen. Hierüber kann sie wiederum zur Erklärung psychischer Störungen beitragen. Durch Klassifikationen können umfangreiche Informationen reduziert und der wissenschaftliche bzw. fachliche Austausch vereinheitlicht werden. Darüber hinaus ermöglicht eine einheitliche Verwendung von Diagnosen eine diagnosenspezifische Behandlung und wissenschaftliche Überprüfbarkeit.
Nach Stieglitz & Freyberger (1999b, S. 32f) werden mit Klassifikationssystemen folgende Ziele verfolgt:

- forschungsrelevante Ziele
  - Charakterisierung von Patientengruppen in empirischen Studien
  - Fallidentifikation in epidemiologischen Studien
  - Grundlage empirischer Untersuchungen zu Ätiologie und Verlauf von Störungen
  - Grundlage empirischer Studien zur Entwicklung und Überprüfung therapeutischer Interventionen
  - Dokumentation von therapeutischen Interventionen psychiatrischer Versorgungseinrichtungen
  - Verbesserung der Kommunikation von Forschungsergebnissen

- klinisch relevante Ziele
  - Vereinfachung und Homogenisierung des psychiatrischen Denkens, Reduktion der Komplexität klinischer Phänomene durch Trennung einzelner Betrachtungsebenen (z.B. deskiptive Diagnostik, psychosoziale Funktionseinschränkungen)
  - Verbesserung der Kommunikation zwischen Klinikern
  - Grundlage der klinisch-psychiatrischen Ausbildung
  - Grundlage für die Indikationsstellung und Einleitung von Behandlungsmaßnahmen sowie für ihre Überprüfung am Therapieerfolg
  - Grundlage für kurz- wie langfristige Prognosestellungen
  - Bedarfsplanung für psychiatrische Versorgungseinrichtungen.

Neben diesen Vorteilen, die mit der Verwendung von Klassifikationsschemata einher gehen, gibt es auch Skepsis, besonders bei Phänomenologen, ob die syndomatologische Klassifikation dem Individuum in seiner Komplexität gerecht werden kann.
So ist es im organ-medizinischen Bereich zwar üblich, Diagnosen zu stellen, im psychischen Bereich wurde die Klassifikation jedoch lange kontrovers diskutiert. Ein weiterer Grund hierfür ist die Angst vor Etikettierung und Stigmatisierung.
Da die Gefahr einer Etikettierung, etwa mit Bezeichnungen wie „die Hysterikerin“ oder „der Schizophrene“, tatsächlich bei Klassifikationen besteht, wurde bereits im DSM-III-R davon abgekommen, dies sprachlich zu unterstützen, indem z.B. nicht mehr von „dem Schizophrenen“ gesprochen wurde, sondern auf Formulierungen wie „bei der Schizophrenie treten ... Symptome auf“ zurückgegriffen wurde. Im DSM-IV liegt die Betonung wieder auf den Menschen mit den Störungen. Jedoch wurde versucht, die Stigmatisierung weiter zu reduzieren, indem z.B. die Rede ist von „Menschen mit Schizophrenie können ... Symptome zeigen“ (Saß et al., 1996).

Neben der Gefahr der Stigmatisierung werden weitere Nachteile der Klassifikation beschrieben (vgl. Margraf, 1996):

- die Festschreibung künstlicher Einheiten, denen dann ein unangemessener Realitätsgehalt zugebilligt wird,
- die Verwechslung von Deskription mit der Erklärung einer Störung und
- die Verschleierung basaler Dimensionen (evtl. existieren keine qualitativ verschiedenen Kategorien, sondern jede Störung ist durch bestimmte Ausprägungen auf verschiedenen Dimensionen gekennzeichnet).

Ein weiteres Problem ist die Komplexität psychischer Störungen im Erscheinungsbild. Es gibt fließende Übergänge zwischen verschiedenen Störungen und ein unzureichendes Wissen über Entstehungsbedingungen (Möller, 1998). Daraus wird bereits ein Großteil der Probleme mit Klassifikationen psychischer Störungen deutlich.
Um der Vielfalt der Erscheinungsbilder gerecht zu werden, werden teilweise bei der Konstruktion von Klassifikationssystemen verschiedene Einteilungskriterien miteinander vermengt, z.B. Ätiologie, Erscheinungsbild, Verlauf oder therapeutische Ansprechbarkeit. Dies verleiht auch den gegenwärtigen Klassifikationssystemen noch einen „eher vorläufigen Charakter“ (Stieglitz & Freyberger, 1999b).
Da Personen meist heterogen in Bezug auf die zu definierenden Merkmale einer Diagnose sind, werden bei Grenzfällen zusätzliche klinische Informationen zur Diagnosestellung berücksichtigt. Im DSM-IV wird versucht, die Heterogenität klinischer Bilder durch die Vorgabe von Kriterienlisten zu vereinfachen, bei denen die Person nur eine Teilmenge von Items einer längeren Liste aufweisen muss. Das Ziel, mit Klassifikationen möglichst homogene Patientengruppen zu schaffen, konnte somit bisher nur eingeschränkt erreicht werden (Stieglitz & Freyberger, 1999b).

Die Zeiten, in denen einzelne Therapieschulen, z.B. die Gesprächspsychotherapie, Diagnosen und Klassifikationen völlig ablehnten, sind jedoch vorbei. Die Entwicklung ist inzwischen sogar so weit, dass in der

Verhaltenstherapie diskutiert wird, ob aufwendige individuelle Verhaltensanalysen noch sinnvoll sind, oder ob Behandlungen nicht auch störungsspezifisch ohne eingehende Verhaltensanalyse, aber nach einer Klassifikation, durchgeführt werden können.
Auch die Psychoanalytiker entwickelten inzwischen Instrumente zur operationalisierten Diagnostik wie die OPD (*Operationalisierte Psychodynamische Diagnostik*, Arbeitskreis OPD, 1998). Mit der OPD soll zwischen rein deskriptiven Systemen und psychodynamischer Diagnostik vermittelt werden. Dabei wird unter anderem eine ICD-Diagnose ermittelt.

Zur Festlegung auf syndromatologische Klassifikationssysteme kommt es schließlich auch dadurch, dass die Forschungsenergie besonders in den letzten 25 Jahren verstärkt in die Verbesserung der Klassifikationssysteme geflossen ist, die nun sehr ausdifferenziert sind und dadurch, dass Alternativen zu den Klassifikationssystemen quasi nicht existieren.

2.3.1 Das DSM-IV und seine Besonderheiten

Das DSM-IV (Diagnostisches und Statistisches Manual Psychischer Störungen, dt.: Saß et al., 1996) enthält etwa 1000 Kriterien für die Erfassung von 395 Störungen. Es handelt sich um das auf englisch 1994 erschienene Klassifikationssystem der American Psychiatric Association (APA). Mit dem DSM wird seit der Version III (1980) ein von theoretischen und ätiologischen Annahmen fast unabhängiger, deskriptiver Ansatz verfolgt.
Mehr als die Vorgänger-Klassifikationssysteme beruht das DSM-IV auf empirischen Grundlagen. Jedoch haben einige diagnostische Kriterien mehr als andere von empirischen Resultaten profitiert. Im Vorfeld der Entwicklung wurden 150 Literaturreviews und 40 Datenreanalysen durchgeführt und die Kriterienlisten von DSM-III, DSM-III-R und die ICD-10 Forschungskriterien in 12 Feldstudien an insgesamt 6000 Patienten verglichen. Es wurde eine Kompatibilität mit der ICD-10, dem

Klassifikationssystem der Weltgesundheitsorganisation (s.u.), angestrebt (Saß et al., 1996; Stieglitz & Freyberger, 1999b).

Ein wesentliches Unterscheidungsmerkmal zwischen der DSM-IV- und der ICD-10- Klassifikation ist das Eingangskriterium des DSM-IV bei fast jeder Störung „Das Störungsbild verursacht in klinisch bedeutsamer Weise Leiden oder Beeinträchtigungen in sozialen, beruflichen oder anderen wichtigen Funktionsbereichen“.

Dabei definiert das DSM-IV eine psychische Störung als klinisch bedeutsames Verhaltens- oder psychisches Syndrom oder Muster, das bei einer Person auftritt und das mit momentanem Leiden oder einer Beeinträchtigung oder mit einem stark erhöhten Risiko einhergeht, zu sterben, Schmerz, Beeinträchtigung oder einen tiefgreifenden Verlust an Freiheit zu erleiden. Zusätzlich darf dieses Syndrom oder Muster nicht nur eine verständliche und kulturell sanktionierte Reaktion auf ein bestimmtes Ereignis sein, wie z.B. beim Tod eines geliebten Menschen (Saß et al., 1996).

Während die erste DSM-Version zunächst auf der psychobiologischen Sicht Adolf Meyers basierte, die zweite aber einen traditionell intrapsychischen Standpunkt einnahm, versucht das DSM-IV wieder biologische, psychologische und soziale Faktoren einzubeziehen, um Störungen besser verstehen, vorzubeugen und behandeln zu können. Auch das Ansprechen auf medikamentöse Behandlung, genetische und neurobiologische kausale Bedingungen von Störungen werden stärker im DSM-IV berücksichtigt als in früheren Versionen (Nathan, 1994).

Das multiaxiale System des DSM, bestehend aus fünf Achsen, wurde derart verändert, dass auf Achse II nur noch die Geistige Behinderung und Persönlichkeitsstörungen kodiert werden. Die Entwicklungsstörungen zählen nun zu den Störungsbildern der Achse I. Auf Achse III werden medizinische Krankheiten, auf Achse V das psychosoziale Funktionsniveau auf einer GAF-Skala (Global Assessment of Functioning: Globale Erfassung des Funktionsniveaus), jetzt bis 100 (im DSM-III-R bis 90) bewertet.

Tabelle: Das multiaxiale System

Achse	Inhalt	Beispiele
I	Klinische Störungen und andere klinisch relevante Probleme	Schizophrenie, Affektive Störungen, Angststörungen, Schlafstörungen, Störungen der Impulskontrolle
II	Persönlichkeitsstörungen und Geistige Behinderung	Paranoide, Borderline, histrionische, dependente Persönlichkeitsstörung
III	Medizinische Krankheitsfaktoren	Infektiöse Erkrankungen, Erkrankungen des Kreislaufsystems, angeborene Störungen, Vergiftungen
IV	Psychosoziale und umgebungsbedingte Probleme	Probleme mit der Hauptbezugsgruppe, im sozialen Umfeld, berufliche Probleme, wirtschaftliche Probleme, Probleme beim Zugang zu Einrichtungen der Krankenversorgung
V	Globale Erfassung des Funktionsniveaus	Zwischen: hervorragender Leistungsfähigkeit bis zu: ständiger Gefahr, sich oder andere schwer zu verletzen

Das DSM-IV ist nicht für die Anwendung durch ungeübte Personen konzipiert, die es mechanistisch verwenden wollen, sondern als Unterstützung eines klinischen Urteils gedacht (Saß et al., 1996). So kann die klinische Beurteilungserfahrung es z.B. rechtfertigen, dass einem Menschen eine bestimmte Diagnose gegeben wird, obwohl das klinische Bild dem Kriterienkatalog der Diagnose nicht vollständig entspricht, die Symptome jedoch anhaltend und ausgeprägt sind. Es darf jedoch nicht so flexibel und idiosynkratisch angewendet werden, dass der Nutzen als allgemeines Kommunikationsmittel dadurch grundlegend beeinträchtigt wird.

2.3.2 Die ICD-10

Die 10. Revision der International Statistical Classification of Diseases (ICD-10) ist ein Klassifikationssystem der Weltgesundheitsorganisation (WHO, dt. Dilling et al., 1991) und dient nicht nur der Klassifikation psychischer Störungen (Kodierung mit dem Buchstaben F), sondern auch von somatischen Störungen (Kodierung mit unterschiedlichen

Buchstaben je nach Störungsart z.B. infektiöse Erkrankungen mit A). Unter Z werden Faktoren, die den Gesundheitszustand beeinflussen und zur Inanspruchnahme von Gesundheitsdiensten führen, zusammengestellt (z.B. Probleme in Verbindung mit Berufstätigkeit und Arbeitslosigkeit Z56.x). Die internationale Klassifikation psychischer Störungen versteht sich als Zusammenstellung von Symptomen und Kommentaren zu Störungen, auf die sich über hundert Experten und psychiatrische Fachgesellschaften aus verschiedenen Ländern geeinigt haben. Die 1987 erstellte Version wurde in über 30 Ländern mit Feldstudien überprüft. In der ICD-10 wurden wie im DSM im Vergleich zur ICD-9 die diagnostischen Kategorien operationalisiert und versucht, einem atheoretischen Ansatz zu folgen, weswegen auf Begriffe wie Neurose, Psychose und Endogenität verzichtet wurde. Der Begriff psychische Krankheit wurde durch den Begriff der Störung ersetzt. In Bezug auf die psychischen Störungen ist die ICD-10 mit zehn diagnostischen Hauptgruppen und 1000 Unterscheidungsmöglichkeiten, die allerdings noch nicht ausgeschöpft sind, sondern noch Entwicklungsmöglichkeiten offen lassen, drei Mal umfangreicher als die ICD-9. Eine weitere Annäherung an das DSM ist die Einführung von Achsen, wobei sich die ICD auf drei Achsen beschränkt (Achse I: klinische Diagnosen (psychische und somatische Störungen), Achse II: soziale Funktionseinschränkungen, Achse III: umgebungs- und situationsabhängige Ereignisse / Probleme der Lebensführung und Lebensbewältigung).

Die Jasperssche Schichtenregel, die in der Vorgängerversion noch gültig war, wurde von dem Komorbiditätsprinzip (s.u.) abgelöst. Trotzdem verfährt die ICD-10-Klassifikation in einer etwas anderen Weise mit Komorbiditäten als das DSM-IV, indem z.B. eine Panikstörung nicht als Hauptstörung klassifiziert wird, wenn gleichzeitig die Kriterien für eine depressive Störung vorliegt. Dem geht die Annahme voraus, dass Panikattacken häufig auch im Zusammenhang mit depressiven Störungen auftreten.

2.3.3 ICD-10 und DSM-IV im Vergleich

Während es noch deutliche Unterschiede zwischen den beiden Vorgängersystemen ICD-9 und DSM-III-R gegeben hat, dadurch, dass das DSM bereits früher eine detailliertere Entwicklung von Störungsgruppen vorgenommen hatte, gibt es nur noch geringe Unterschiede nach der Entwicklung der aktuellen Systeme, da die ICD-10 in jahrelanger Vorarbeit an das DSM angeglichen und beide Systeme aufeinander abgestimmt wurden.

Folgende Unterschiede blieben:

- Die Operationalisierung im DSM-IV ist etwas genauer in Bezug auf die Zeitkriterien und die Anzahl benötigter Symptome für eine Diagnose, die in der ICD-10 nicht bei jeder Störung formuliert sind.
- Die Beschreibung der Störungen sind nur zum Teil miteinander kompatibel (z.B. wird eine Bulimia nervosa in der ICD-10 als Störung mit Untergewicht beschrieben, im DSM-IV würde dies einer Anorexia nervosa entsprechen).
- Die klinische Relevanz (Beeinträchtigung und Belastung in verschiedenen Funktionsbereichen) einer Störung wird im DSM-IV stärker hervorgehoben. In der ICD-10 werden die Funktionsbeeinträchtigungen lediglich auf Achse II kodiert.
- Die Diagnosekategorien des DSM-IV sind empirisch noch besser abgesichert als die ICD-10-Kriterien.
- Für das DSM-IV existiert nur eine Version der Klassifikation, für die ICD-10 mehrere, wobei die ICD-10-Forschungskriterien die größte Ähnlichkeit mit dem DSM-IV zeigen.
- Die multiaxialen Systeme sind verschieden eingeteilt, so dass die ICD-10 nur 3 Achsen, das DSM-IV 5 Achsen aufweist. So werden die Achsen I, II und III des DSM als Achse I (alle Störungen, auch organische) in der ICD kodiert. Hierbei wird deutlich, dass die ICD-10 als internationales Klassifikationssystem für alle Krankheiten konzipiert ist und nur als ein Kapitel psychische Störungen beinhaltet, wogegen das DSM-IV speziell der Erfassung psychischer Störungen dient.

- Die Kodierungen der beiden Systeme sind nicht nur verschieden (die DSM-IV-Kodierung entspricht noch eher der ICD-9-Kodierung), sondern in der ICD-10 differenzierter, so dass hier durch die 5. Stelle eine Ebene mehr differenziert werden kann als im DSM. Auch kann die ICD-10-Kodierung leichter logisch hergeleitet werden und ist daher leichter einzuprägen.

EBENE	ICD-10		DSM-IV	
2	F3	Affektive Störungen		Affektive Störung
3	F33	Rezidivierende depressive Störung	296	Depressive Störungen, Bipolare Störungen
4	F33.1	gegenwärtig mittelgradige Episode	296.3	Major Depression, Rezidivierend
5	F33.11	mit somatischen Symptomen	296.32	mittelschwer

- Die ICD-10 wurde im internationalen Konsens entwickelt, das DSM-IV ist ein US-amerikanisches System, das die bisherige diagnostische Tradition fortsetzen soll, wobei in der Forschung beide Systeme international eingesetzt werden.

Die Bewertung der beiden Klassifikationssysteme kann in besser oder schlechter nicht pauschal erfolgen. Teilweise werden in der Forschung Diagnosen nach ICD-10 und DSM-IV zum Vergleich gestellt. Der Vergleich der Diagnosen kann dann neue Aufschlüsse über Störungsbilder geben. Meist bezieht sich die Forschung, auch bei vielen epidemiologischen Studien, jedoch allein auf die DSM-IV-Klassifikation.
In der klinischen Praxis werden im Gegensatz dazu im allgemeinen ICD-10-Diagnosen verwandt, da somit alle Krankheiten nach einem Klassifikationssystem klassifiziert werden können, was verwaltungstechnisch und für die Kostenträger von Relevanz ist.

2.3.4 Probleme der aktuellen Klassifikationssysteme

Insgesamt existieren mehr Forschungsarbeiten zu ersten Erfahrungen mit dem DSM-IV als zur ICD-10. Deshalb beziehen sich die folgenden Kritikpunkte primär auf das DSM-IV, wobei durch die enge Anlehnung des ICD an das DSM die Kritik auch auf die ICD-10 zutrifft.
Wakefield (1997a) sieht in seiner Kritik nicht nur bei den Vorgängerklassifikationssystemen, sondern auch beim DSM-IV noch das Problem, dass die Kriterien zu breit formuliert sind und zu viele falsch Positive eingeschlossen werden, da Unterschiede zwischen psychischen Störungen und normalen Lebensproblemen nicht getroffen werden können. Beispielsweise existiert zwar für die Depressive Störung ein Ausschluss von Trauerreaktionen, nicht aber für eine andere normale unkomplizierte Traurigkeit nach dem Erfahren einer terminalen Erkrankung oder nach dem Verlust des Arbeitsplatzes oder einer intensiven Liebesbeziehung. Zwar haben einige dieser depressiven Reaktionen Störungscharakter, jedoch können die DSM-IV-Kriterien nicht unterscheiden, welche davon Reaktionen im Sinne einer Störung und welche normal und unkompliziert sind.
Das gleiche gilt für die Bipolar-II-Störung (mindestens eine depressive und eine hypomane Episode). Hier reicht eine vier Tage andauernde leidenschaftliche Liebesbeziehung mit schlaflosen Nächten etc. und einem anschließenden Verlassenwerden mit Traurigkeit, Appetitverlust etc., um die notwendigen Kriterien für die Störung zu erfüllen. Die Bedingung einer klinisch signifikanten Belastung und Beeinträchtigung soll die Schwelle für eine Diagnose erhöhen und damit falsch Positive ausschließen. Dies kann das oben geschilderte Problem jedoch nicht reduzieren, da die normalen Reaktionen auf schwierige Lebensereignisse natürlich auch mit Belastung und Beeinträchtigung einher gehen (Wakefield, 1997a).

Während Eysenck (1986) jedoch noch erwartete, dass der Versuch einer psychiatrischen Klassifikation mit dem DSM-IV ohne das Aufgeben von Prämissen und Annahmen des DSM-III noch weniger wissenschaftlich

als das von ihm als unzulänglich bewertete DSM-III werden würde, hat das DSM ab der Version III-(R) doch entgegen diesen Erwartungen gerade in der Forschung noch mehr an Bedeutung erlangt.

Zwischen der weiten Verbreitung und Akzeptanz des Systems und der klinischen Praxis bzw. dem sorgfältigen Gebrauch des DSM besteht jedoch eine Kluft. Wahrscheinlich orientieren sich viele Kliniker implizit noch stärker an traditionellen Modellen oder im Sinne individueller klinischer Gepflogenheiten. Auch zusätzliche Diagnosen als Komorbidität werden nicht immer berücksichtigt (Saß et al., 1996).

2.4 Umgang mit Komorbidität

Der Begriff Komorbidität bedeutet das gemeinsame Auftreten verschiedener Störungen bei einer Person in einem definierten Zeitraum. Schon in der Zeit Kraepelins wurden kombinierte Psychosen diskutiert, wobei jedoch triftige Gründe bestehen mussten, um beim selben Kranken zwei verschiedene Krankheiten anzunehmen. Danach setzte sich die bis zum DSM-III-R übliche diagnostische Hierarchieregel (Jasperssche Schichtenregel) durch, die bedeutet, dass psychische Erkrankungen in Schichten angeordnet sind, wobei es tiefer liegende Störungen gibt, die sich in einer darüber liegenden Störung ausdrücken, weshalb nur eine, die tiefer liegende Störung, diagnostiziert wird.

Die oberste Schicht besteht aus den „neurotischen Symptomen (das Psychasthenische, Hysterische), dann die manisch-depressiven, dann die Prozeßsymptome (das Schizophrene), schließlich die organischen (psychischen und körperlichen) Symptome. Die tiefste Schicht, die bei der Untersuchung des Einzelfalles erreicht wird, gibt den Ausschlag für die Diagnose. Was zuerst als Hysterie erschien, erweist sich so als multiple Sklerose, eine Neurasthenie als Paralyse, eine melancholische Depression als Prozeß usw.“ (Jaspers, 1973, S. 512)

Für die Anwendung dieser Regel sprechen (vgl. Stieglitz, 1999b) die Identifizierung der wichtigsten Diagnose für die Behandlung, Therapie und Prognose, die Identifizierung derjenigen Diagnose mit der sparsamsten Erklärung der Phänomenologie, die Hilfe im differenzialdiagnostischen Prozess und die Identifizierung sogenannter reiner Fälle.

Die Abkehr von dieser hierarchischen Verfahrensweise erfolgte, da für die Hierarchisierung keine empirische Begründung vorliegt und außerdem ein zu hoher Verlust an Information über den Patienten, an therapeutischen Möglichkeiten und an Validierungsmöglichkeiten damit einher geht (Stieglitz, 2000). Das DSM erlaubt seit der Version III-R multiple Diagnosen. Voraussetzung für die Vergabe mehrerer Diagnosen ist allerdings, dass die Symptome der einen Diagnose nicht unter eine andere Störung subsummiert werden können (Saß et al., 1996).

Beim Vorliegen zweier Störungen A und B (Komorbidität) bestehen folgende Interpretationsmöglichkeiten (aus Stieglitz, 2000, nach Frances et al., 1990):

- A prädisponiert oder verursacht B (beispielsweise könnte eine Essstörung eine Sozialphobie nach sich ziehen, wenn die Patientin ihren Körper unförmig findet und deswegen Angst vor den Blicken und Beurteilungen anderer hat und sich entziehen möchte)
- B prädisponiert oder verursacht A (die Essstörung könnte auch die Folge einer vorher bestandenen Sozialphobie sein, so dass die Patientin sich isolierte aufgrund der Sozialphobie und nachfolgend als Möglichkeit zur Anerkennung und Schaffung von Erfolgen zu fasten beginnt und ihre Figur verändert)
- A und B werden beide beeinflusst von einem weiteren, bisher nicht bekannten Faktor (die Patientin könnte evtl. unter einer Borderline-Persönlichkeitsstörung leiden und aufgrund schlechter Beziehungserfahrungen Angst vor sozialen Situationen haben und eine Essstörung haben aufgrund der Schwierigkeit mit heftigen Gefühlen und dem andauernden Gefühl innerer Leere umzugehen)

- die Assoziation von A und B ist ein Zufall (Essstörung und Sozialphobie könnten unabhängig voneinander sein)
- A und B treten gemeinsam auf aufgrund der den Diagnosen zugrunde liegenden gemeinsamen diagnostischen Kriterien (z.B. generalisierte Angststörung und Major Depression)
- A und B treten gemeinsam auf, weil sie artifiziell getrennt wurden (evtl. z.B. Sozialphobie und abhängige Persönlichkeitsstörung)

Eine Studie von Zimmerman & Mattia (1999) zeigte, dass beim Vergleich von 500 klinischen Routine-Interviews (unstrukturiert) mit 500 SCID-Interviews für DSM-IV (s.u.) von vergleichbaren Patienten-Stichproben im SCID-Interview signifikant mehr Achse I-Diagnosen gestellt wurden als mit unstrukturierten Interviews. Mehr als ein Drittel erhielt nach einem SCID-Interview drei oder mehr Diagnosen, während dies bei den klinischen Interviews nur weniger als 10 % waren. Die Untersucher gehen demzufolge davon aus, dass in klinischer Routine-Diagnostik Komorbidität zu selten erkannt wird, besonders bei Angst-, somatoformen und nicht anders spezifizierten Störungen.

In epidemiologischen Studien, wie der Münchener Sieben-Jahres-Follow-up-Studie (Wittchen & von Zerssen, 1987) und einer anderen Komorbiditätsstudie (National Comorbidity Survey, Kessler et al., 1994) sind alle Angststörungen statistisch signifikant mit depressiven und Substanzstörungen assoziiert (Wittchen & Vossen, 1996). Das Lebenszeitrisiko von Patienten mit einer Generalisierten Angststörung, auch eine depressive Störung zu entwickeln, liegt dabei im Vergleich zu anderen ohne Generalisierte Angststörung etwa sieben Mal höher. Im klinischen Bereich kann im Querschnittsbefund mit fast 50 %-iger Komorbidität gerechnet werden (Wittchen & Vossen, 1996). Alpert et al. (1997) untersuchten die Komorbidität von Major Depression, Sozialphobie und ängstlich-vermeidender Persönlichkeitsstörung und stellten einen hohen Zusammenhang zwischen den Störungen fest (27 % der depressiven Patienten erfüllten die Kriterien einer Sozialphobie, 28 % die einer ängstlich-vermeidenden Persönlichkeitsstörung und davon wiederum zwei Drittel beides). Bei den depressiven Patienten mit einer Sozialphobie und einer

vermeidenden Persönlichkeitsstörung trat die Major Depression in einem früheren Alter erstmals auf, und es bestand eine höhere Anzahl von weiteren Achse-I-Störungen.

Bei diesen Ergebnissen muss man sich sicherlich fragen, ob es sich immer um unabhängige Störungsbilder handelt, oder ob die diagnostischen Befunde nicht in eine Störung allein integrierbar wären. Allerdings lässt sich die Forschung leichter weiter voran treiben, wenn die vielfältigen Symptome beim Auftreten einer psychischen Störung durch komorbide Diagnosen mit erfasst werden und damit Subgruppen von z.B. depressiven Patienten gebildet werden können.
Die Klassifikation komorbider Diagnosen ist vor allem forschungsrelevant, insofern dass damit eher „Typen“ erkannt und verglichen werden können, evtl. neue Störungsklassen entdeckt oder eine Vereinfachung von Störungsbildern durch eine Zusammenfassung unter eine Störung erreicht werden kann. Die prognostische Relevanz beim Vorliegen mehrerer Störungen gleichzeitig oder nachfolgend kann auch in der Prävention von Störungen wichtig sein und kann damit nach einer Bedingungsanalyse auch therapeutisch relevant sein. So könnte der Hinweis, dass z.B. einer Depression meist eine Generalisierte Angststörung vorausgeht, Präventionsarbeit hier ansetzen und das Auftreten einer Depression verhindern, damit Kosten für das Gesundheitssystem senken.

Das Kapitel „Diagnostik psychischer Störungen“ verdeutlichte die Schwierigkeiten, mit denen sich die psychiatrische Diagnostik auseinandersetzen muss. Daraus resultieren immer wieder Modifikationen in der Zuordnung von Merkmalen und Störungen. Trotz der Zweifel, ob psychische Störungen nicht besser durch dimensionale Systeme erfasst werden können, setzte sich aufgrund der besseren Operationalisierbarkeit mehr und mehr die kategoriale Diagnostik durch. Die aktuellen Klassifikationssystemen ICD-10 und DSM-IV ermöglichen eine multiaxiale und die Diagnostik komorbider Störungen.

3 Gütekriterien der Diagnostik

Das folgende Kapitel gibt Aufschluss darüber, wie Diagnoseinstrumente in ihrer Güte beurteilt werden können, welche Kriterien es hierfür gibt und wie die Gütekriterien berechnet und bewertet werden.

Wie gut und genau Diagnosen erfasst werden, wird im allgemeinen mit den „Gütekriterien“ eines diagnostischen Instruments beschrieben. Der Ausdruck „Gütekriterien“ weist auf Qualitätsanforderungen hin, die an jedes Testverfahren gestellt werden (Amelang & Zielinski, 1997).
In der klassischen Testtheorie wird davon ausgegangen, dass das Testergebnis dem wahren Ausprägungsgrad des untersuchten Merkmals entspricht. Allerdings wird zusätzlich angenommen, dass jedes Testergebnis zusätzlich von einem Messfehler überlagert wird. Als Messfehler werden beispielsweise ungeeignete Fragen, schlechte Untersuchungsbedingungen etc. angesehen.
Die klassische Testtheorie basiert auf den fünf folgenden Axiomen (Bortz & Döring, 1995):

- Das Testergebnis setzt sich additiv aus dem wahren Wert und dem Messfehler zusammen (z.B. ergibt sich die Diagnose „Major Depression“ aus der tatsächlichen Depression und evtl. aus der nur während des Tests vorhandenen negativen Sichtweise der Dinge).
- Bei wiederholten Testanwendungen kommt es zu einem Fehlerausgleich mit Reduktion des Mittelwertes des Messfehlers, so dass schließlich der wahre Wert repräsentiert wird (z.B. kann man immer stärker von einer „Major Depression“ ausgehen, je öfter der Patient die gleichen Angaben bei verschiedenen Messwiederholungen macht).
- Wahrer Wert und Messfehler sind unabhängig voneinander (Fehlereinflüsse durch eine momentan negative Sichtweise sollten bei Patienten mit und ohne „Major Depression“ möglich sein).
- Die Höhe des Messfehlers ist unabhängig vom Ausprägungsgrad anderer Persönlichkeitsmerkmale (z.B. sollten die Messfehler bei der Erfassung einer „Major Depression“ nicht von der Angst des Patienten vor dem Interview abhängig sein).

- Die Messfehler verschiedener Testwiederholungen sind voneinander unabhängig (z.B. sollte ein Patient, der bei einem Interview besonders müde war, nicht bei jedem Interview gleich müde sein).

Auf der Basis der Axiome werden die drei zentralen Testgütekriterien definiert. Dabei wird in Haupt- und Nebengütekriterien unterschieden.
Zu den *Hauptgütekriterien* zählen die Objektivität, die Reliabilität oder Zuverlässigkeit, die Validität oder Gültigkeit. Unter *Nebengütekriterien* fallen die Normierung (populationsspezifische Bezugsgrößen zur Interpretation der Ergebnisse der untersuchten Person) und die Testfairness (Ausmaß einer systematischen Diskriminierung der untersuchten Person z.B. durch die soziokulturelle Gruppenzugehörigkeit).

Im folgenden wird auf die Hauptgütekriterien näher eingegangen.

3.1 Objektivität

Die *Objektivität* steht für das Ausmaß, in dem die Ergebnisse eines Tests unabhängig von der Person des Untersuchenden sind (Durchführungsobjektivität, Auswertungsobjektivität, Interpretationsobjektivität). Durch ein Testmanual können entsprechende Vorgaben, wie der Test durchgeführt, ausgewertet und interpretiert wird, gemacht werden. Die Durchführungs- und Auswertungsobjektivität sind eine notwendige Voraussetzung für die Reliabilität eines Tests. Die Interpretationsobjektivität ist eine Voraussetzung für die Validität des Tests.

3.2 Reliabilität

Reliabilität oder Zuverlässigkeit beschreibt die Genauigkeit, mit der ein Test eine Merkmalsdimension erfasst und zwar unter der Vernachlässigung des Umstandes, ob es sich dabei auch um die Merkmalsdimension handelt, deren Erfassung intendiert ist. Aspekte der Treffsicherheit (Validität) bleiben außer Acht. Nur die Präzision der Messung an sich inte-

ressiert bei der Reliabilität (Amelang & Zielinski, 1997). Die Reliabilität ist um so höher, je geringer der Messfehler ist. Eine perfekte Reliabilität würde zeigen, dass der Test den wahren Wert ohne jeden Messfehler erfassen kann. Die Reliabilität umfasst die Unabhängigkeit der Daten vom Messinstrument bzw. die Reproduzierbarkeit der Daten.
Der Reliabilitätskoeffizient (r_{tt}) ist die Korrelation zwischen verschiedenen Messwerten, erhoben an demselben Probanden. Da das zweite Axiom vorgibt, dass sich die wahre Merkmalsausprägung auch bei wiederholten Messungen nicht ändert, dagegen der Fehler ausgeglichen wird, müsste also ein vollständig reliabler Test nach wiederholter Vorgabe bei denselben Probanden zum gleichen Ergebnis führen. Abweichungen werden auf Messfehler zurückgeführt. Auf der Grundlage der Axiome 3 bis 5 sind die Messfehler vom wahren Wert, von anderen Merkmalen und von einander unabhängig. Bei Messwiederholungen können sich somit nur unsystematische Abweichungen von den Messwerten ergeben. Diese werden als Fehlervarianz bezeichnet. Je größer die Fehlervarianz, desto mehr Messfehler beinhalten die beobachteten Werte.

$$r_{tt} = \frac{s_{wahrerWert}^2}{s_{beobachteterWert}^2}$$

Je größer der korrelative Zusammenhang zwischen beiden Messwertreihen, desto höher liegt der Anteil der Varianz der wahren Werte. Ein Reliabilitätskoeffizient von $r_{tt} = .70$ bedeutet also, dass die Varianz der beobachteten Werte zu 70 % auf wahre Unterschiede zwischen den Testpersonen und zu 30 % auf die Fehlervarianz zurückzuführen sind. Der Reliabilitätskoeffizient hat einen Wertebereich von 0 bis 1. Dabei bedeutet 0, dass sich der beobachtete Wert nur aus Messfehlern ergibt. Ein Wert von 1 gibt an, dass der beobachtete Wert identisch mit dem wahren Wert ist. Bei Intelligenztests liegt die Reliabilität meist zwischen 0,8 bis 0,95. Bei Persönlichkeitstests sind dagegen deutlich geringere Reliabilitätskoeffizienten zu erwarten, meist zwischen 0,6 bis 0,7 (Wottawa, 1981).

Zur Bestimmung der Reliabilität eines Tests wird außer der Varianz der beobachteten Werte eine Schätzung für die wahre Varianz ermittelt. Diese Schätzung kann mittels verschiedener Methoden durchgeführt werden. So besteht die Möglichkeit der Testwiederholung (Retest-Reliabilität), der Paralleltest-Reliabilität (zwei parallele Versionen eines Tests werden vorgegeben), der Testhalbierungs-Reliabilität (Items werden bei homogenen Instrumenten in zwei Hälften aufgeteilt) und der Konsistenzanalyse (Interne Konsistenz, Unterteilung in die Anzahl der Items).

Zur Reliabilitätsuntersuchung von Klassifikationen eignet sich die Testwiederholung. Hierbei wird in der Regel ein und derselbe Test zweimal derselben Stichprobe von Probanden dargeboten. Die Korrelation zwischen der ersten und der zweiten Vorgabe gibt das Ausmaß der Retestreliabilität an. Erinnerungs- und Übungseffekte sollen vermieden werden (hierdurch könnte die Reliabilität überschätzt werden), weshalb es günstig ist, einen größeren Zeitabstand für die beiden Untersuchungen zu wählen. Auf der anderen Seite soll der Zeitabstand nicht so groß sein, dass es zu realen Schwankungen des erfassten Merkmals kommt und dadurch die Messgenauigkeit des Instruments auf niedrigerem Niveau erscheint, als es tatsächlich der Fall ist. Da die Retest-Reliabilität immer abhängig ist von dem Ausmaß, in dem das zu erfassende Merkmal stabil ist, wird für den Begriff Retest-Reliabilität auch synonym der Begriff Teststabilität verwandt.
Der Einsatz der Testwiederholungsmethode empfiehlt sich nicht bei der Erfassung instabiler, zeitabhängiger Merkmale wie z.B. der Stimmung. Eine geringe Retest-Reliabilität ließe dann Zweifel offen, ob die geringe Reliabilität des Tests oder die geringe Stabilität des Merkmals für das Ergebnis verantwortlich sind.
Die Reliabilität wird von einer mangelnden Objektivität beeinträchtigt, da Diskrepanzen zwischen den Testanwendern eine Fehlervarianz erzeugen. Die Reliabilität kann deshalb nur höchstens so hoch sein wie die Objektivität.

3.3 Validität

Validität oder Gültigkeit bedeutet das Maß an Genauigkeit, mit dem der Test dasjenige Merkmal misst, das er messen soll oder vorgibt zu erfassen. D.h., ein diagnostisches Interview sollte tatsächlich die Depression messen und nicht eine momentane Müdigkeit. Die Validität eines Tests ist durch seine Korrelation mit einem Kriterium gekennzeichnet. Es handelt sich dabei um das wichtigste Gütekriterium überhaupt (Bortz & Döring, 1995; Amelang & Zielinski, 1997), da sie angibt, ob ein Test sinnvoll als Prädiktor eingesetzt werden kann. In Bezug auf die Validität sind von Relevanz die inhaltliche Validität, die kriteriumsbezogene Validität und die Konstruktvalidität.

Inhaltliche Validität (Face Validity, Logische Validität): Hierbei geht es um die Adäquatheit der Abbildung der Störung, also darum, ob die in der Testsituation erhaltenen Ergebnisse das zeigen, was außerhalb der Testsituationen an Anforderungen besteht. In Bezug auf eine depressive Störung könnte dies bedeuten, ob die Fragen das treffen, was den Patienten quält (z.B.: wird neben der Antriebslosigkeit auch alle anderen wichtigen Depressionszeichen erfragt?). Ein Interview, das zur Erfassung einer Essstörung kein Körpergewicht erfragt, hätte somit eine geringe Inhaltsvalidität. Wenn es keine Fragen oder Aufgaben gibt, die das Zielkonstrukt treffen sollen, müssen Experten beurteilen, inwieweit die Fragen dem Inhalt des Konstruktes (der Störung) entsprechen. Damit handelt es sich bei der Inhaltsvalidität nicht um ein numerisch bestimmbares Maß (Bortz & Döring, 1995), sondern lediglich um eine Zielvorgabe, die bei der Konstruktion eines Tests bedacht werden soll.

Kriteriumsbezogene Validität: Um von den Testergebnissen auf das Zielmerkmal (Kriterium) bei sich nicht in ihrer Ganzheit prüfbaren Merkmalen schließen zu können, wird mit Korrelationen zwischen Testergebnis und Kriterium geprüft, ob die Ergebnisse mit dem Kriterium übereinstimmen.

Da bei der Bestimmung von Diagnosen mehrere Kriterien denkbar sind, wobei unklar ist, welches Kriterium das optimale ist, existieren mehrere Validitäten. Bei der Auswahl von Validitätskriterien bedarf es der theoretischen Begründung in Bezug auf die jeweilige Störungsgruppe, da je Störungsgruppe andere Validitätsaspekte von Bedeutung sein können.
Hierbei widersprechen sich evtl. auch die aus verschiedenen Informationsquellen ermittelten Kriterien, so dass eine Wahl für das am relevantesten erscheinende Kriterium erfolgen muss. Wenn z.B. die Diagnose einer Bulimia nervosa kriteriumsvalidiert werden soll, könnte als Außenkriterium die Anzahl von beobachtbaren Fressattacken oder die Selbstauskunft über Fressattacken dienen. Dabei kann es jedoch sein, dass die Diagnose „Bulimia nervosa“ nicht mit der Anzahl beobachtbarer Fressattacken korreliert, da die Fressattacken heimlich durchgeführt werden oder die Selbstauskunft aufgrund von Schamgefühlen unehrlich ist. Dann hieße das nicht, dass die Validität der Diagnose hinterfragt werden müsste, sondern die des Außenkriteriums. Auch könnte es sein, dass es Moderatorvariablen gibt. Z.B. könnten bei einer Bulimia nervosa Patientinnen mit langer Krankheitsdauer ehrlicher Auskunft geben zur Anzahl der Fressattacken, als Patientinnen mit kurzer Krankheitsdauer und damit die Selbstauskunft valider die tatsächliche Störung vorhersagen.

Die Konstruktvalidität sagt aus, ob ein Test tatsächlich das psychologische Konstrukt erfasst, das er erfassen soll. Es handelt sich dabei um eine Synthese aus Inhalts- und Kriteriumsvalidität, bei der es weniger um die Validität als um die Validierung als Vorgang geht. Dabei werden Hypothesen über Zusammenhänge zwischen Testergebnissen und objektiv beobachtbarem Verhalten in einer möglichst großen Vielfalt überprüft. Wenn aus dem zu messenden Zielkonstrukt Hypothesen abgeleitet werden können, die anhand von Testwerten bestätigt werden können, gilt der Test als konstruktvalide. Dabei ist eine Konstruktvalidierung umso überzeugender, je mehr Hypothesen ihre Überprüfung bestehen.
Hierzu kann die Multitrait-Multimethod-Methode (MTMM) (Bortz & Döring, 1995) durchgeführt werden. Es handelt sich dabei um eine Über-

prüfungsmethode, bei der mehrere Konstrukte durch mehrere Erhebungsmethoden erfasst werden. Konvergente Validität liegt vor, wenn mehrere Methoden (z.B. Depressivitätsfragebogen, klinisches Urteil, Selbstauskunft des Hauptproblems) dasselbe Konstrukt (Depressivität) übereinstimmend (konvergent) messen (vgl. Richter, 1994). Diskriminante Validität liegt vor, wenn sich das Zielkonstrukt (Depressivität) von anderen Konstrukten (Substanzabhängigkeit, Somatoforme Störungen) unterscheidet, d.h. sie gibt an, wie gut verschiedene Konstrukte durch eine Methode differenziert werden.

Die Validität kann betragsmäßig insgesamt nicht größer als die Wurzel der Reliabilität sein. Tatsächlich gibt es jedoch eine Ausnahme, bei der die Validität größer als die Reliabilität sein kann, nämlich, wenn die Reliabilität des Tests niedrig, die Reliabilität des Kriteriums und die Korrelation der wahren Werte von Test und Kriterium aber sehr hoch sind (Krauth, 1995).

Eine gute Reliabilität und Validität von Diagnosen sind bzgl. der Behandlung der Störungen als Schutz vor Fehlentscheidungen von großer Relevanz. Falsche Diagnosen intendieren eine Zuführung zur falschen Behandlungsform und sind zudem nur schwer wieder zu eliminieren.

3.4 Berechnung und Bewertung der Retest-Reliabilität

Bei Klassifikationsurteilen bewegt man sich auf Nominalskalenniveau. Dementsprechend werden auch die Reliabilitätsmaße verwendet.
Zur Reliabilitätsbestimmung werden die Urteile der Interviewer in eine Matrix überführt, d.h. die Übereinstimmung bzw. Nicht- Übereinstimmung zwischen den beiden Interviewern ist hier ablesbar.

3.4.1 Prozentuale Übereinstimmung

In früheren Untersuchungen wurden oft nur die prozentualen Übereinstimmung zwischen zwei Diagnostikern betrachtet, wobei es sich um die einfachste Größe der Reliabilitätsberechnung handelt. Diese errechnet sich aus der Summe der übereinstimmenden Diagnosen im Vergleich zu allen vergebenen Diagnosen.

$$p = \frac{\sum_{j=1}^{k} f_{jj}}{n}$$

Bei der Durchführung von zwei Interviews durch zwei unabhängige Interviewer ergeben sich folgende Möglichkeiten:

- Beide Interviewer kommen zu einer übereinstimmenden Diagnose (Feld a).
- Die Interviewer kommen zu verschiedenen Diagnosen (Felder b und c).
- Beide Interviewer vergeben übereinstimmend keine Diagnose (Feld d).

		Interview 1	
		+	-
Interview 2	+	a	b
	-	c	d

Am folgenden Beispiel errechnet sich

		Interview 1	
		+	-
Interview 2	+	2	3
	-	2	184

die prozentuale Übereinstimmung aus
p = (2 + 184) / 191 = 0,97 (97%).

Der Nachteil der prozentualen Übereinstimmung liegt, besonders wenn wenige Kategorien zur Auswahl bestehen, darin, dass Zufallsübereinstimmungen in das Ergebnis mit einfließen.

3.4.2 Der Kappa-Koeffizient

Für den Vergleich von Diagnosen ist als Maß der Reliabilität daher der von Cohen (1960) (vgl. Aspendorpf & Wallbott, 1979) entwickelte Kappa-Koeffizient (κ) gebräuchlich, da hiermit die zufällige Übereinstimmung zwischen zwei Diagnostikern korrigiert werden kann.
So liegt der Kappa-Koeffizient aus dem o.g. Beispiel lediglich bei 0,40 anstatt bei 0,97.

Die Formel zur Berechnung des Kappa-Koeffizienten zur Korrektur der Zufallsübereinstimmung lautet folgendermaßen:

$$\kappa = \frac{p - p_e}{1 - p_e}$$

wobei:

$$p_e = \frac{1}{n^2} \sum_{j=1}^{k} f_{j.}\ f_{.j}$$

$f_{j.}$ = Zeilensummen

$f_{.j}$ = Spaltensummen

so dass

$$p_e = \frac{(a + b)(a + c) + (b + d)(c + d)}{(a + b + c + d)^2}$$

Ein Kappa (κ) -Wert von 0 gilt als Zufallsübereinstimmung; sind die Werte unter 0, ist die beobachtete Übereinstimmung geringer als die erwartete Zufallsübereinstimmung. Ein Wert von .40 bedeutet, dass die Diagnosenübereinstimmung 40 % über der erwarteten zufälligen Übereinstimmung liegt. Der κ -Wert von 1 gibt die perfekte Übereinstimmung zweier Beurteiler an.

Die Berechnung von κ –Werten führt bei wenigen Kategorien und bei gering gefüllten Zellen (Basisrate einer Störung) eher zu niedrigen Reliabilitätsschätzungen. Diese systematische Unterschätzung der Übereinstimmung wird von Brennan & Prediger (1981) kritisiert. Dennoch gehen Aspendorpf & Wallbott (1979) davon aus, dass κ –Werte die Reliabilität auf Nominalskalenniveau am angemessensten abbilden können.

3.4.3 Yules Y-Koeffizient

Bei Basisraten für eine Störung, die sehr gering sind, d.h. unter 10 % liegen, ist bekannt, dass der Kappa-Koeffizient die Übereinstimmung unterschätzt. In einem hypothetisch konstruierten Fall zeigen Grove et al. (1981), dass bei gleicher Sensitivität und Spezifität von diagnostischen Urteilen, aber bei Unterschieden in der Basisrate von 50 % bzw. 1 %, der Kappa-Koeffizient von .81 auf .14 gesenkt wurde. Spitznagel & Helzer (1985) schlagen daher für den Fall von geringen Basisraten die Berechnung des Y-Koeffizienten (Yule, 1912) vor. Dieser berechnet sich folgendermaßen:

$$Y = \frac{\sqrt{a\,d} - \sqrt{b\,c}}{\sqrt{a\,d} + \sqrt{b\,c}}$$

Zur Bestimmung der Reliabilität wird er oft als zusätzliches Informationsmaß verwendet (Wittchen et al., 1991; Schneider et al., 1992; Stieglitz, 2000). Bei Basisraten unter 10 % liefert Yules Y deutlich höhere Übereinstimmungswerte als κ, bei Basisraten über 10 % gleichen sich Y und κ an, wobei Y weiterhin leicht über dem Kappa-Wert liegt. Basisraten über 50 % bewirken, dass Y unzuverlässig und instabil, d.h. wenig

aussagekräftig wird (Spitznagel & Helzer, 1985). Dabei wird der Y-Wert automatisch 1, wenn in Feld b oder c eine Null steht, ist nicht berechenbar, wenn in zwei Feldern eine Null steht und wird zu –1, wenn Feld a oder b eine Null beinhaltet. Damit kann er in all diesen Fällen keine genaue Aussage über die Reliabilität machen.

Nach Bortz & Döring (1995) sollte ein guter Test eine Reliabilität von > .80 aufweisen. Danach gelten Reliabilitäten erst bei > .90 als hoch.
Interpretationsleitlinien für Maße der Retest-Reliabilität gehen eher davon aus, dass Reliabilitätskoeffizienten > .75 für eine ausgezeichnete, Werte zwischen .40 und .75 für eine ausreichende und Werte < .40 für eine ungenügende Übereinstimmung sprechen (Landis & Koch, 1977; Fleiss, 1981). Wittchen et al. (1998) geben Kappa-Werte von <.40 als geringe Übereinstimmung, Werte > .50 als mittlere, Kappa-Werte >.65 als gute und Werte > .75 als ausgezeichnete Übereinstimmung an. Nach Bortz & Döring (1995) werden Kappa-Werte über .70 als gut bezeichnet.
Für den Yule-Koeffzienten Y werden in der Literatur keine vom Kappa-Wert abweichenden Angaben zur Bewertung gemacht. Satlow (1996) verwendet jedoch erst Y-Werte > .80 als ausgezeichnete und Y-Werte zwischen .65 und .80 als ausreichende Übereinstimmung, da der Yule-Koeffizient bei Basisraten von 1 % bis 10 % relativ hohe Übereinstimmungswerte liefert und auch bei Basisraten zwischen 10 % und 50 % leicht über dem Kappa-Wert liegt.

3.5 Berechnung und Bewertung der Validität

Die Berechnung der Validität kann entweder durch die Überprüfung von Zusammenhangshypothesen (z.B. die F-DIPS-Diagnose stimmt mit der Entlassungsdiagnose überein) oder durch die Überprüfung von Unterschiedshypothesen (z.B. Patienten mit einer Agoraphobie haben eine höhere Ausprägung im Fragebogen als Patienten ohne Störung) erfolgen. Für die verschiedenen Störungen können zur Validierung Vorhersagen über Ausprägungen einer Störung in einem die Störung betreffenden

Fragebogen vorgenommen werden. Es werden Hypothesen darüber aufgestellt, wie sich eine Störung abbilden lässt (z.B. störungsspezifische Fragebögen, klinische Diagnose oder halbstrukturiertes Interview) und die Hypothesen überprüft.
Wenn durch mehrere Erhebungsmethoden das Konstrukt erfasst wird, liegt eine und sich von anderen Störungsbildern abgrenzen lässt, ist der Test valide.

Das F-DIPS soll in der vorliegenden Untersuchung auf die Hauptgütekriterien Reliabilität und Validität geprüft werden.
Im Kapitel „Gütekriterien der Diagnostik“ wurden die Gütekriterien und deren Berechnung vorgestellt. Die Bestimmung der Reliabilität bei Klassifikationen, in denen beurteilt wird, ob zwei Untersucher zu einem übereinstimmenden Ergebnis kommen, erfolgt durch die Testwiederholung, die Angabe in Kappa-Koeffizienten und Yules-Y-Koeffizienten. Die Abschätzung der Validität geschieht anhand eines Außenkriteriums und der Prüfung von Zusammenhangs- bzw. Unterschiedshypothesen.

4 Mängel bei der psychiatrischen Diagnostik und bisheriger Klassifikationssysteme

Dieses Kapitel gibt einen Überblick über verschiedene Studien zur Überprüfung der Reliabilität und Validität von Diagnosen und zu Fehlerquellen, die zur Nicht-Übereinstimmung von Diagnostikern führen.

Neben den von Spitzer et al. (1975) beschriebenen Problemen der traditionellen psychiatrischen Diagnostik, dass das medizinische Modell keine Bedeutung für psychiatrische Probleme habe, dass mehrere, auch überlappende Grundlagen für Klassifikationen bestehen, die psychiatrische Diagnostik nur einen begrenzten Wert für Behandlungen, Prognosen, Ätiologieverständnis oder für phänomenologische Beschreibungen habe, wird vor allem die mangelnde Reliabilität der psychiatrischen diagnostischen Praxis beklagt.

4.1 Studien zur Reliabilität von Diagnosen

Eine Vielzahl von Untersuchungen belegt die mangelnde Reliabilität klinisch-psychiatrischer Diagnosen, zumindest vor Einführung des DSM-III (Beck et al., 1962; Spitzer & Fleiss, 1974; Spitzer & Wilson, 1975).

Die Untersuchung von Beck et al. (1962) basiert auf der Klassifikation nach DSM-I, für dessen Gebrauch 4 Psychiater geschult wurden. Die 4 Diagnostiker wurden paarweise randomisiert und die Übereinstimmung jedes Diagnostikerpaares geprüft. Hierbei ergab sich bei der Klassifizierung von 153 ambulanten Patienten eine durchschnittliche prozentuale Übereinstimmung von 54 % in Bezug auf die Anzahl der positiven Diagnosen (siehe Tabelle), wobei die Übereinstimmung je nach Diagnostiker-Paarung stark schwankte und am geringsten war bei Paarungen mit dem relativ unerfahrensten Psychiater (6 Jahre im Vergleich zu 10-13 Jahren), der außerdem im Unterschied zu den anderen nicht mit ambulanten, sondern mit stationären Patienten arbeitete.

Außerdem wurden die Übereinstimmungen für die verschiedenen diagnostischen Kategorien berechnet. Die Ergebnisse hierfür finden sich in der Tabelle.

Kategorie	Anzahl der gewählten Diagnose durch den Untersucher	% Übereinstimmung *	κ-Koeffizient
Neurotische depressive Reaktion	92	63	.47
Angstreaktion	58	55	.45
Soziopathische Störung	11	54	.53
Schizophrene Reaktion	60	53	.42
Involutionsreaktion	11	40	.38
Persönlichkeitsstörung	26	38	.33

* Die Berechnung der prozentualen Übereinstimmung geht nicht von der Gesamtzahl der Patienten aus, sondern von der Anzahl jeder ausgewählten diagnostischen Kategorie durch die Diagnostiker. D.h., nur die Übereinstimmung bzgl. des Vergebens der Diagnose (Feld a, siehe oben) wurde berechnet und die Übereinstimmung bzgl. der Ablehnung der Diagnose (Feld d) nicht einbezogen. Somit lassen sich die relativ niedrigen prozentualen Übereinstimmungen erklären.

Wenn beide Untersucher angaben, mit ihrem diagnostischen Urteil sicher zu sein (31 Fälle), erhöhte sich die Übereinstimmung auf 81 %. Waren sich beide Untersucher unsicher (4 Fälle), bestand nur noch in einem Fall eine Übereinstimmung.
In 37 % der Fälle von Nicht-Übereinstimmung sind Inkonsistenzen oder Auslassungen im Interview mit verantwortlich.

Spitzer & Fleiss (1974) (vgl. auch Spitzer & Wilson, 1975; Margraf, 1996) zogen aus einem Literaturreview von neun damals vorliegenden Studien (Schmidt & Fonda, 1956; Kreitman et al., 1961; Beck, 1962; Sandifer et al., 1964; Cooper et al., 1969; Spitzer, 1973) zur Retest-Reliabilität den Schluss, dass die Übereinstimmung am besten bei organischen Störungen, weniger gut bei Psychosen und am schlechtesten bei Neurosen sei.

Die Übereinstimmungen wurden zur Vergleichbarkeit in Kappa-Koeffizienten umgerechnet und Durchschnittswerte aus den unterschiedlichen Kappa-Werten gebildet. Die in den Studien verwendeten Klassifikationssysteme (DSM-I, DSM-II und ICD-8) waren zwar verschieden, die Definition der Klassifikationskategorien jedoch in diesen Systemen gleich, so dass von vergleichbaren Ergebnissen ausgegangen werden kann.

Dabei ergaben sich folgende mittleren Interraterreliabilitäten (κ):

Diagnose	Studienanzahl	κ – Koeffizient
Hirnorganisches Syndrom	3	.77
Schizophrenie	8	.54
Affektive Störung	5	.37
Neurotische Depression	5	.21
Psychotische Depression	1	.19
Persönlichkeitsstörung	7	.29
Neurose	7	.36
Persönlichkeitsstörung und Neurose kombiniert	6	.39
Alkoholismus	4	.71

Für Forschungs- und Behandlungszwecke waren die damals gegebenen Übereinstimmungen sicherlich nicht ausreichend (Beck et al., 1962; Margraf, 1996).

In einer aktuelleren Studie von Way et al. (1998), in der die Einschätzung psychiatrischer Notfälle durch 8 erfahrene Psychiatern untersucht wurden, wurden klinische Interviews auf Video aufgezeichnet und den Psychiatern demonstriert. Dabei ergaben sich höhere Übereinstimmungen bzw. ähnlichere Einschätzungen (Intraclass Correlation Coefficient ICC) in Bezug auf verschiedene Diagnosen (z.B.: Psychose r = .64, Substanzmissbrauch r = .65, Depression r = .48) als hinsichtlich der Ein-

schätzung psychopathologischer Größen wie z.B. der „Fähigkeit, für sich selbst die Verantwortung zu übernehmen“ r = .28, „Impulskontrollprobleme“ r = .30, „Gefahr für sich selbst“ r = .32 und auch der Einschätzung der „Güte des am Video gesehenen Interviews“ r = .30. Bei diesen Größen handelt es sich evtl. um noch weniger klar definierte Einheiten als bei Diagnosen.

Diese Ergebnisse widersprechen Annahmen (Segal et al., 1986), dass Psychiater zur gleichen Einschätzung kämen, wenn sie über dieselben Informationen bzgl. eines Patienten verfügten. Durch die Video-Aufzeichnungen erhielten die Psychiater in der beschriebenen Studie dieselben Informationen, kamen aber dennoch zu divergierenden Einschätzungen.

4.2 Studien zur Validität von Diagnosen

Mängel bestehen auch bzgl. der Validität von Diagnosen. Aufgrund des Fehlens klarer und verlässlicher diagnostischer Kriterien gibt es in epidemiologischen Untersuchungen Schwankungen in der Einschätzung der Prävalenz psychischer Störungen in der Allgemeinbevölkerung zwischen 3 – 5 % bis zu 70 % (Margraf, 1996). In einer bundesweiten epidemiologischen Untersuchung von Wittchen et al. (1999) wird die Prävalenz für affektive Störungen, Angst- und somatoforme Störungen innerhalb eines Jahres mit 17,3 % der deutschen Bevölkerung angegeben.

Rosenman et al. (1997) verglichen in ihrer Studie an 126 Patienten einer akutpsychiatrischen Station die Übereinstimmung zwischen der mit Hilfe des standardisierten Interviews CIDI (Composite International Diagnostic Interview) ermittelten computerisierten ICD-10-Diagnose (CIDI-Auto) und der Diagnose eines Psychiaters (nach einem psychiatrischen Interview) sowie die Übereinstimmung zwischen zwei Psychiatern. Im CIDI-Auto wurden 2,3 Diagnosen pro Patient ermittelt, die Psychiater stellten im Durchschnitt 1,3 Diagnosen. Der CIDI-Auto diagnostizierte signifikant mehr Substanzgebrauchsstörungen, Affektive Störungen und neuroti-

sche Störungen als die Psychiater. Die Kappa-Koeffizienten für die verschiedenen Störungsoberklassen finden sich in der folgenden Tabelle:

	κ - Werte der Übereinstimmung Psychiater / CIDI-Auto	κ - Werte der Übereinstimmung Psychiater / Psychiater
F1x: Psychische und Verhaltensstörungen durch psychotrope Substanzen (Substanzabhängigkeiten)	.29	.53
F2x: Schizophrenie, schizotype und wahnhafte Störungen	.34	.84
F3x: Affektive Störungen	.23	.59
F4x: Neurotische, Belastungs- und somatoforme Störungen (in der Hauptsache Angst- und somatoforme Störungen)	.03	.49
F5x: Verhaltensauffälligkeiten in Verbindung mit körperlichen Störungen und Faktoren (vor allem Essstörungen)	.30	.92

Insgesamt lag die Übereinstimmung zwischen CIDI-Auto und Psychiater bzgl. der Diagnose-Oberklassen bei κ =.23, während die Übereinstimmung zwischen zwei Psychiatern bei κ =.69 lag und damit gut war. Die Autoren schließen daraus, dass die CIDI-Diagnostik mit Hilfe eines Computers im akutpsychiatrischen Bereich aufgrund der geringen Validität nicht sinnvoll eingesetzt werden kann.

Bei DSM-I, DSM-II und ICD-8 lag die erzielte Reliabilität nur für das hirnorganische Psychosyndrom und für Alkoholismus im guten bis ausgezeichneten Bereich, Schizophrenien konnten mit mittlerer Reliabilität, Affektive Störungen, Neurosen und Persönlichkeitsstörungen mit ungenügender Reliabilität erfasst werden. Insgesamt sind vorliegenden Ergebnisse sehr inkonsistent, und es wechselt je nach Studie zwischen Kritik an traditioneller Diagnostik bis hin zu sehr reliablen Ergebnissen bei psychiatrischen Diagnosen. Beim anderen Extrem des völlig standardisierten diagnostischen Vorgehens wie im computerisierten standardisierten Interview besteht die Schwierigkeit, keine validen Diagnosen mehr zu erreichen (Kappa-Werte von .03 bis .34).

4.3 Fehlerquellen bei Nicht-Übereinstimmung von diagnostischen Einschätzungen

Bei solchen Befunden ist es natürlich sinnvoll, sich Gedanken zu möglichen Ursachen für die verschiedenen Einschätzungen von geschilderten Problemen der Patienten zu machen und nach Fehlerquellen zu suchen. Dabei kann man sich vorstellen, dass bei der Erhebung von Daten für die Diagnostik nach ICD-10 oder DSM-IV folgende Fehlerquellen auf Seiten des Interviewers oder des Patienten auftreten können (Stieglitz & Freyberger, 1999b, S. 55):

- Nichtbeachtung der Symptom-, Zeit- und Verlaufskriterien der jeweiligen Störung
- Nichtberücksichtigung der Ausschlusskriterien der jeweiligen Störung
- Nichtberücksichtigung des Komorbiditätsprinzips
- Beeinflussung durch theoretische Konzepte, die nichts mit der Diagnose zu tun haben (z.B. ätiologische Konzepte)
- Einfluss eigener diagnostischer Unsicherheit bei der Entscheidung für eine Diagnose (z.B. Borderline-Störung, schizoaffektive Störung)
- Rückschluss auf eine Diagnose aufgrund eines singulären Phänomens (z.B. hysterisch = hysterische Persönlichkeitsstörung)
- falsche Schlussfolgerungen (z.B. logischer Fehler: Annahme, dass bestimmte psychopathologische Phänomene immer zusammen auftreten müssten, oder Halo-Effekt: ein besonders markantes Merkmal beeinflusst die Wahrnehmung anderer Merkmale)
- der Patient erinnert sich fehlerhaft
- der Patient gibt absichtlich falsche Informationen (Simulation, Dissimulation, Bagatellisierung, Aggravation)

Fehlerquellen, die als mögliche Ursache für die mangelnde Übereinstimmung der Diagnosen verschiedener Diagnostiker gelten, können liegen (nach Spitzer & Fleiss, 1974, vgl. Stieglitz & Freyberger, 1999b) in der:

- Subjektvarianz (z.B. durch Therapieeinfluss): Ein Patient wird zu zwei Zeitpunkten untersucht, an denen er sich in verschiedenen Krankheitszuständen befindet.

- Situationsvarianz: Ein Patient wird zu zwei Zeitpunkten untersucht, an denen er sich in verschiedenen Phasen oder Stadien einer Störung befindet.
- Informationsvarianz: Verschiedenen Untersuchern stehen unterschiedliche Informationen zum Patienten und zu seiner Erkrankung zur Verfügung.
- Beobachtungsvarianz (z.B. durch Symptomgewichtung): Verschiedene Untersucher kommen zu unterschiedlichen Urteilen und Bewertungen über Vorhandensein und Relevanz der vorliegenden Symptome.
- Kriterienvarianz: Verschiedene Untersucher verwenden unterschiedliche diagnostische Kriterien für die Diagnose derselben Störung

Außerdem kann bei Anwendung eines strukturierten oder standardisierten Interviews noch eine Rolle spielen:

- Interviewerfehler (das Interview wird fehlerhaft angewandt)
- das Klassifikationssystem oder das Interview selbst sind unzureichend oder uneindeutig, so dass es zu einem unzureichenden Informationsgewinn durch unpräzise Fragen oder Kriterien kommt.

Als Quellen der Nichtübereinstimmung beim Erstellen der Diagnosen nach DSM-I geben Ward et al. (1962) an, nachdem sie Patienten durch Psychiaterpaare diagnostisch beurteilen und danach bei einer Nichtübereinstimmung (40 Patienten) die Gründe für jede Abweichung diskutieren ließen:

- Inkonsistenzen auf Seiten des Patienten 5%
- Inkonsistenzen auf Seiten der Diagnostiker 32,5%
- Inadäquatheit der Nosologie des DSM-I 62,5%

Die Diagnostiker verwendeten teilweise unterschiedliche Interviewtechniken, durch die sie zu unterschiedlichen Informationen kamen und gewichteten die Relevanz von Symptomen verschieden.

Schneider et al. (1992) fanden bei ihrer Analyse der Fehlerquellen bei der Nichtübereinstimmung von mit dem Diagnostischen Interview bei psy-

chischen Störungen (DIPS) gewonnenen Diagnosen als häufigste Fehlerquelle die Informationsvarianz (33%). Die nächst bedeutsamen Fehlerquellen betrafen den Interviewer, zum einen die unterschiedliche Symptomgewichtung (25 %), zum anderen die unterschiedliche Durchführung des Interviews (20 %). Unklarheiten im DSM-III-R oder im DIPS waren zu 13 % verantwortlich für die Nicht-Übereinstimmung in der Diagnose.

Das Kapitel „Mängel der psychiatrischen Diagnostik“ berichtete über die Ergebnisse einer Vielzahl seit den 60er Jahren durchgeführter Studien zur Güte von Diagnosen. Bei Anwendung der früheren Klassifikationssysteme DSM-I, DSM-II und ICD-8 konnten Diagnosen nur mit mäßiger bis ungenügender Reliablilität erfasst werden. Fehlerquellen bestanden fast ausschließlich in der Inadäquatheit der Nosologie des Klassifikationssystems und in Inkonsistenzen der Diagnostiker mit unterschiedlichen Interviewtechniken. Diese Befunde deuteten auf die Notwendigkeit der Verbesserung der Klassifikationssysteme hin und sprachen für die Anwendung von Interviewleitfäden zur objektivierbaren Erfassung der Patienteninformationen.

5 Der diagnostische Prozess

Das Kapitel beschreibt, wie patientenbezogene Informationen erhoben werden können, die schließlich zu einer psychiatrischen Diagnose führen. Der Schwerpunkt liegt auf der Vorstellung diagnostischer Interviews, insbesondere des F-DIPS.

In den diagnostischen Prozess werden verschiedene Informationen mit Abweichungen in Art und Umfang einbezogen. So kann ein Interviewer verschiedene Datenquellen nutzen, wie Angaben des Patienten und fremdanamnestische Angaben (Vorbehandler, Angehörige, Beobachtungen des Pflegepersonals etc.). Außerdem kann er Informationen auf verschiedenen Datenebenen erfassen (Baumann & Stieglitz, 1994): auf der psychischen (psychologischen) Datenebene mit individuellem Erleben und Verhalten, auf der sozialen Datenebene mit dem sozialen Netz und dem Ausmaß der sozialen Unterstützung, auf der ökologischen Datenebene mit materiellen Rahmenbedingungen und der biologischen Datenebene mit der Differenzierung nach biochemischen, psychophysiologischen, neuroradiologischen und neuropsychologischen Ebenen.

5.1 Die Datenerfassung für die operationalisierte Diagnostik

Die Datenerfassung zur Klassifikation nach z.B. DSM-IV kann mit Hilfe verschiedener Interviews erfolgen, nämlich durch ein:

- freies Interview
- halbstrukturiertes Interview
- vollständig strukturiertes Interview
- standardisiertes Interview.

5.1.1 Freies Interview

Für die Diagnostik von psychischen Störungen in der klinischen Praxis stellen der psychische Befund und die Anamnese die zentralen Bausteine dar (Stieglitz & Freyberger, 1999b), wobei der psychische Befund die

Einschätzung des Psychiaters oder Psychologen zur Psychopathologie des Patienten angibt (Wachheit, Orientierung, inhaltliches und formales Denken, Ich-Störung, Psychomotorik, Affekivität, Suizidalität etc.) und in die Anamnese (Erhebung der Krankengeschichte mit Entstehungsbedingungen) psychische, soziale und ökologische Daten mit einfließen bzw. auch unterschiedliche Datenquellen genutzt werden. Im Behandlungsalltag psychiatrischer oder psychotherapeutischer Einrichtungen muss zum schnellen Behandlungsbeginn die vorläufige Diagnose ebenfalls schnell gestellt werden, so dass das freie Interview hier Anwendung findet.

Dem freien Interview kommen verschiedene Funktionen zu: Erhebung der Anamnese, Beziehungsaufnahme zum Patienten sowie Aufbau einer therapeutischen Beziehung (Stieglitz & Freyberger, 1999a). Das psychiatrische Erstinterview dauert zwischen 30 bis 60 Minuten.

Im freien Interview wird eine grobe Strukturierung, meist Gliederung in drei Teile, vorgenommen. Nach einer Eingangsfrage (Was führt Sie hierher? Was sind Ihre Beschwerden? etc.), auf die der Patient frei antworten kann, schließen sich Fragen zum Beginn der Beschwerden, zur Beeinflussbarkeit, dem Verlauf, der Familienanamnese, dem sozialen Hintergrund und biografischen Informationen an. Außerdem wird die somatische Anamnese erhoben und die aktuelle und prämorbide Persönlichkeit beurteilt. Aus diesen Informationen wird aufgrund der Erfahrung des Interviewers eine vorläufige Diagnose erstellt, die je nach Ausrichtung entweder an der ICD-9 (inzwischen selten) oder ICD-10 oder der DSM-IV-Nomenklatur angelehnt ist oder durch eine psychoanalytische Terminologie gekennzeichnet ist. Die formalen Diagnosekriterien aus einem Klassifikationssystem können dabei, müssen aber nicht vollständig erfragt werden, um die vorläufige Diagnose und differenzialdiagnostische Erwägungen zu formulieren. Oft zählen der Eindruck vom Patienten, die Erwartungen des Untersuchers oder dessen „Prototyp“-Erfahrungen von einer bestimmten Störung. Sobald eine subjektive Sicherheit bzgl. der Diagnose besteht, werden evtl. noch Validierungsfragen gestellt. Im dritten Teil des Gesprächs werden weitere diagnostische Schritte oder therapeutische Maßnahmen und Therapieziele besprochen.

Die im Erstinterview gestellte Diagnose wird bis zum Ende einer Behandlung oft mehrfach modifiziert, indem neue Informationen einbezogen werden, so dass Aufnahme- und Entlassungsdiagnose oft nicht identisch sind (Stieglitz & Freyberger, 1999a).

5.1.2 Halbstrukturierte Interviews

Eine Form halbstrukturierter Interviews sind Checklisten. Diese beinhalten Auflistungen von Diagnosekriterien, bzgl. derer der Patient befragt wird, wobei die Art der Frageformulierung zur Ermittlung des Kriteriums dem Untersucher vorbehalten bleibt. Es folgen Entscheidungsbäume mit konkreten Angaben dazu, wann die Diagnose vergeben wird. In der täglichen Routinediagnostik sind die Checklisten als Leitfaden für das Explorationsgespräch als Unterstützung möglich, da sie flexibel angewendet werden können. Die Informationserhebung erfordert klinische Erfahrung und variiert je nach Art und Strategie, die Symptome zu erfragen (Stieglitz, 2000). Wenn nicht alle Checklisten bzw. alle Kriterien überprüft werden sollen, ist ein hypothesengesteuertes Vorgehen notwendig, wodurch sich eine weitere Fehlerquelle ergibt.
Eine etwas stärker strukturierte Form eines halbstrukturierten Interviews stellt das SCAN dar. Hier sind die Fragen schon anformuliert, jedoch kann der Interviewer seine Formulierungen benutzen. (Beispiele: SCAN *Schedules for Assessment in Neuropsychiatry,* dt. van Gülick-Bailer et al., 1995 / IDCL *Internationale Diagnosen Checkliste für DSM-IV,* Hiller et al., 1995).

5.1.3 Vollständig strukturierte Interviews

In strukturierten Interviews werden Fragen in der vorformulierten Weise gestellt und Störungen in einer festgelegten Reihenfolge systematisch erfasst. Es existieren Sprungregeln zum Auslassen von Fragen für den Fall, dass Eingangsfragen negativ beantwortet werden. Der Interviewer

hat Freiheiten in bezug auf Umformulierungen der Fragen, falls der Patient sie nicht versteht oder die Antwort nicht eindeutig ist. Er kann Patientenangaben selbst bewerten, so dass es auch möglich ist, Beobachtungen, die den Angaben des Patienten widersprechen, einfließen zu lassen. (Beispiele: DIPS *Diagnostisches Interview bei psychischen Störungen,* Margraf et al., 1991, 1994 / SKID *Strukturiertes Klinisches Interview für DSM-IV,* Wittchen et al.,1997 / IPDE *International Personality Disorder Examination,* Loranger, dt. Mombour et al., 1996 / CIS-R *Clinician Interview Schedule-Revised,* Lewis et al., 1992)

5.1.4 Standardisierte Interviews

In standardisierten Interviews sind alle Schritte des diagnostischen Vorgehens (Fragestellung und Auswertung) genau festgelegt. Der Interviewer verfügt über keinerlei Freiheiten, kann bei Widersprüchen nicht selbst bewerten, sondern muss sich ausschließlich an den Antworten des Patienten orientieren. Damit ist eine klinische Erfahrung des Untersuchers keine Voraussetzung für die Durchführung des Interviews. Da jeder Patient genau die gleichen Vorgaben erhält und auch die Auswertung der Antworten standardisiert erfolgt, ist die Objektivität gewährleistet, und die Reliabilität wird maximiert. Ob sich jede Störung gleich gut auf diese Weise erfassen lässt, die Validität also gewährleistet ist, stellt sich jedoch als Frage.
(Beispiele: CIDI *Composite International Diagnostic Interview,* dt. Wittchen & Rupp, 1991 / M-CIDI *Munich-Composite International Diagnostic Interview* Wittchen et al., 1998/ DIS *Diagnostic Interview Schedule,* Robins, 1981).

Die folgende Tabelle gibt einen Überblick über einige gebräuchliche diagnostische Interviews für psychische Störungen (Achse I des DSM-IV):

Kurzbezeichnung	Ausführlicher Name	Diagnosesystem	Autoren
DIS	Diagnostic Interview Schedule	DSM-III, Feighner Kriterien, RDC (Research Diagnostic Criteria)	NIMH National Institute of Mental Health; Robins et al., 1981
SKID	Strukturiertes Klinisches Interview für DSM-IV	DSM-IV	Wittchen et al., 1997
SCAN	Schedules for Clinical Assessment in Neuropsychiatry	PSE (ICD-10)- kompatible Klassen und DSM-III-R	WHO, Wing, 1988, dt.: van Gülick-Bailer et al., 1995
DIPS	Diagnostisches Interview bei psychischen Störunen	DSM-III-R und therapierelevante Informationen	Margraf et al., 1991, 1994
CIDI	Composite International Diagnostic Interview	ICD-10 und DSM-IV	Wittchen et al., 1998
ADIS-IV	Anxiety Disorders Interview Schedule	DSM-IV	Brown et al., 1994
IDCL	Internationale Diagnosen Checklisten	DSM-IV und ICD-10	Hiller et al., 1997

Für die Durchführung der meisten Verfahren sind klinisch erfahrende und in den Verfahren trainierte Rater vorgesehen. Nicht nur Möller (1998) sieht das Einsatzgebiet dieser Instrumente wegen des hohen Aufwandes weniger in der Routinepraxis als im Rahmen aufwendiger Forschungsprojekte.

Die Reliabilität lässt sich jedoch durch die strukturierte oder standardisierte Befunderhebung noch weiter verbessern und wird sowohl von der American Psychiatric Association als auch von der Division for Mental Health der Weltgesundheitsbehörde empfohlen, um die psychopathologischen, taxonomischen und nosologischen Entscheidungsregeln der Klassifikationssysteme besser erlernen und anwenden zu können.

5.2 Studien zur Qualität diagnostischer Interviews

5.2.1 zur Retest-Reliabilität

Von Wittchen et al. (1998) wurde die Retestreliabilität des standardisierten Interviews M-CIDI (Münchener Composite International Diagnostic Interview) in der computerisierten DSM-IV-Fassung erfasst. Dazu wurden 60 Personen innerhalb eines durchschnittlichen Zeitraums von 38 Tagen ein zweites Mal durch trainierte Interviewer betragt. Es waren dabei geringere Kappa-Werte für Bulimia nervosa (.55) und Generalisierte Angststörung (.45) sowie gute Kappa-Werte (.65 und mehr) für alle anderen Störungen sowie ausgezeichnete für Angststörungen insgesamt (.81) und Alkoholabhängigkeit und -missbrauch (.78) zu verzeichnen. Die relativ geringen Übereinstimmungen bei der Generalisierten Angststörung und der Bulimia nervosa führen Wittchen et al. (1998) auf die Zeitdauer- und Häufigkeitskriterien bei den beiden Störungsbildern zurück. Thornton et al. (1996) sehen für die Essstörungen eher den ich-syntonen Charakter, die Heimlichkeit und Verleugnung als Ursache für die schwierige Erfassung der Störungen mit Hilfe des CIDI.

Für das ADIS und das SCID für DSM-IV sowie das SCAN (ICD-10 und DSM-III-R) fanden sich bei der Literaturrecherche keine Retest-Reliabilitätsstudien. Jedoch wird das SCID in vielen Studien eingesetzt und innerhalb der Studien die Interraterreliabilität trainiert und anhand kleiner Stichproben, meist nur in Bezug auf bestimmte Störungen untersucht. Für Alkohol- und Substanzmissbrauch geben Martin et al. (2000) eine Interraterreliabilität von Kappa-Werten zwischen 0,82 und 1,0 an. Das SCAN wird vielfach als Kriterium in Validitäts-Studien verwendet, z.B. im Vergleich mit dem CIDI (Wittchen, 1994b) oder mit der DIS (Eaton et al., 2000). Für den Bereich Alkohol- und Drogenabhängigkeit existiert eine gute bis ausgezeichnete Reliabilität (Easton et al., 1997).
Ventura et al. (1998) fanden, dass nach einem Training der Interraterreliabilität klinisch erfahrene und neu angelernte Interviewer eine hohe,

nicht signifikant voneinander unterschiedliche Reliabilität erreichen und halten konnten.

In der folgenden Abbildung werden als Übersicht die Retest-Reliabilitäten (κ –Koeffizienten) verschiedener Interviews für die Diagnostik nach DSM-III-R aufgeführt.

Diagnose	DIS (1992)	CIDI (1987)	ADIS (1983)	SCID (1991)	DIPS (1992) lifetime	DIPS (1992) derzeit	Durchschnittlicher κ
Major Depression	.49	.66	.57	.69	.69	.52	.60
Bipolare Störung		.47		.70	-.01		.59 *
Dysthyme Störung	.16	.47	.57	.80	.51	.50	.50
Panikstörung	.45	.84	.69	.27	.70	.81	.63
Agoraphobie	.49	.65	.85	.36	.68	.84	.65
Phobische Störungen	.51	.57			.68	.97	.68
Alkoholabhängigkeit	.63	.79		.75			.72
Drogen-/ Medikamenten-abhängigkeit		.73		.75			.74

* ohne Einbeziehung des Kappa-Wertes des DIPS

Diagnostische Instrumente, die für DSM-IV oder ICD-10 entwickelt wurden und bereits auf ihre Retest-Reliabilität untersucht sind, werden in der folgenden Tabelle zusammenfassend aufgeführt.

Diagnose	M-CIDI (1998)	IDCL (1997)	Durchschnittlicher κ
Major Depression	.68	.73	.71
Bipolare Störung	.64	.85	.75
Dysthyme Störung	.70	.50	.60
Panikstörung	1.0	.88	.94
Agoraphobie	.84	.52	.68
Phobische Störungen	.77	.67	.72
Alkoholabhängigkeit	.78	.80	.79
Drogen-/ Medikamenten-bhängigkeit	.64	.77	.71

Wie an dieser Auflistung der Reliabilitätskoeffizienten zu ersehen ist, gab es, soweit man bei den Daten zu lediglich zwei neueren diagnostischen Interviews schon eine Aussage treffen kann, zwischen Interviews, die nach dem alten Klassifikationssystem DSM-III-R diagnostizierten und solchen, die nach DSM-IV und ICD-10 klassifizieren, keine entscheidenden Veränderungen bzgl. der Reliabilität. Lediglich bei der Major Depression und bei der Panikstörung könnte sich ein positiver Effekt in Richtung Steigerung der Reliabilität abzeichnen.

5.2.2 zur Validität

Der Vergleich zwischen dem standardisierten Interview CIDI– Psychose Modul und einer Diagnose-Checkliste, durchgeführt von erfahrenen Psychiatern (Cooper et al., 1998) zeigte zwar sehr gute Ergebnisse für die Diagnose der Schizophrenie (κ =.82 für DSM-IV und κ = .71 für ICD-10), jedoch deutlich geringere Übereinstimmungen für Subkriterien wie negative Symptome, Katatonie oder Neologismen bzw. dann, wenn für Symptome ein klinisches Urteil gefragt war, so dass insgesamt eine geringe Übereinstimmung zwischen CIDI und klinischer Checkliste entstand.
Das computerisierte Version des Composite International Diagnostic Interview (CIDI-Auto) wurde außer in der oben genannten Studie (Kap. 4.2) auch von Peters & Andrews (1995) auf seine Validität, ebenfalls durch den Vergleich mit klinischen Diagnosen, untersucht. Insgesamt gab es eine Übereinstimmung zwischen Kliniker und CIDI-Auto von κ =.40 mit der Spanne von einem zufälligen Übereinstimmungswert von κ =.02 für die Generalisierte Angststörung bis zu einer ausgezeichneten Übereinstimmung für die Zwangsstörung (κ =.81).

Die Überprüfung der Validität des DIPS (Diagnostisches Interview bei psychischen Störungen) (Margraf et al., 1991) erfolgte, indem die DIPS-Diagnosen mit den Ergebnissen einer Fragebogenbatterie und mit der Größe und dem Gewicht der Patientinnen bei Essstörungen verglichen

wurden. Es zeigte sich hierbei für alle Angststörungen, für Depressionen, somatoforme Störungen und Essstörungen eine gute Validität. Die Angststörungen konnten außerdem voneinander anhand der Fragebogenergebnisse differenziert werden. Lediglich dysthyme Störung und Major Depression ließen sich schlecht unterscheiden.

Für das SCID (Structured Clinical Interview for DSM-IV) ergaben sich in einer Validitätsstudie von Kranzler et al. (1996) im Vergleich zu Expertenurteilen eine gute Validität für den Substanzmissbrauch, eine mittlere für die Major Depression und eine ungenügende Validität für Angststörungen.

Bei der Kreuzvalidierung von dem voll strukturierten Interview CIS-R (Clinician Interview Schedule-Revised) mit dem halbstrukturierten SCAN (Brugha et al., 1999) zeigte sich eine ungenügende Konkordanz für die ICD-10-Kategorie „Neurotische Störungen“ (κ =.25) und für die Depressive Störung (κ =.23). Insgesamt wird davon ausgegangen, dass das SCAN ein valides Instrument ist (vgl. Hesselbrock et al., 1999), weshalb es gerne zur Validierung verwendet wird, eine Studie zur Untermauerung dieser Aussage ließ sich jedoch bei der Literaturrecherche nicht finden.

Die ICDL für DSM-III-R, von erfahrenen Psychiatern durchgeführt, wurde mit dem CIDI verglichen (Janca et al., 1992). Hierbei gab es ausgezeichnete Übereinstimmungen für die Depressive Störung (κ =.84), Substanzmissbrauch (κ =.83) und Angststörungen (κ =.76).

5.3 Das F-DIPS

Das F-DIPS (Diagnostisches Interview bei psychischen Störungen in der Forschungsversion, Margraf et al., 1996, follow-up 1997) ist als strukturiertes Interview konzipiert. Es beschränkt sich auf die Diagnostik von einigen, ausgewählten DSM-IV-Achse-I-Störungen (Klinische Störungen und andere klinisch relevante Probleme), die in der psychologischen Praxis häufig sind und auf die globale Beurteilung des psychosozialen Funktionsniveaus (Achse V). Die diagnostische Einschätzung auf Achse II (Persönlichkeitsstörungen und Geistige Behinderung), Achse III (medizinische Krankheitsfaktoren) und Achse IV (psychosoziale und Umgebungsprobleme) kann nicht mittels F-DIPS-Interview erfolgen. Die Entwicklung des F-DIPS basiert zum einen auf dem Anxiety Disorders Schedule for DSM-IV (ADIS-IV, Brown et al., 1994), zum anderen auf dem Vorläufer DIPS (Margraf et al. 1991, 1994), das den Kriterien des DSM-III-R folgt. Auf Achse I ermöglicht es die diagnostische Einschätzung von

A Angststörungen

- Panikstörung ohne Agoraphobie (300.01)
- Panikstörung mit Agoraphobie (300.21)
- Agoraphobie (300.22)
- soziale Phobie (300.23)
- spezifische Phobie (300.29)
- generalisierte Angststörung (300.02)
- Zwangsstörung (300.3)
- Posttraumatische (309.81) und Akute Belastungsstörung (308.3)

B Affektiven Störungen

- schwere depressive Störung (Major Depression, 296.2x, 296.3x)
- dysthyme Störung (300.4)
- Bipolar-I und -II- Störung (296.0x, 296.4x, 296.6x, 296.5x, 296.89)
- zyklothyme Störung (301.13)
- (Anpassungsstörung mit Angst und depressiver Stimmung gemischt (309.28))

C Somatoformen Störungen
- Hypochondrie (300.7)
- Somatisierungsstörung (300.81)
- Konversionsstörung (300.11)
- somatoforme Schmerzstörung (307.80)

D Substanzmissbrauch und -abhängigkeit
- Alkohol (303.90, 305.00)
- Medikamente (304.10, 305.40)
- Drogen (304.x0, 305.x0)

E Essstörungen
- Anorexia nervosa (307.1)
- Bulimia nervosa (307.51)

Auch die Diagnostik von Störungen des Kindes- und Jugendalters ist mit dem F-DIPS (1996) möglich, nicht jedoch mit der follow-up-Version (1997). Mit Hilfe einiger Screening-Fragen soll es möglich sein, Psychosen zu erkennen, die Klassifikation ist aber nicht möglich.

Bisher wurde das F-DIPS vor allem innerhalb der Studie „Prädiktoren psychischer Gesundheit / Krankheit bei jungen Frauen“ in Dresden angewendet. In dieser Stichprobe sind Frauen im Alter von 18 bis 24 Jahren. D.h., die Fragen sind im soziodemografischen Bereich und bei Körpersymptomen auf junge Frauen bezogen.
Da innerhalb der o.g. Studie auch subklinische Beschwerden interessieren, dürfen Abschnitte *nur* übersprungen werden, wenn es gar keine Auffälligkeiten gibt. So werden neben den Eingangsfragen auch die weiteren störungsbezogenen Fragen gestellt, bevor Abschnitte übersprungen werden können. Die gemischte Angst-Depressions-Störung wurde im F-DIPS als einzige Anpassungsstörung zu Forschungszwecken mit aufgenommen, ohne dass im F-DIPS die für Anpassungsstörungen spezifischen Kriterien erfragt werden. Hiermit soll untersucht werden, ob Per-

sonen, die die Kriterien einer Angst- oder affektiven Störung nur in subklinischem Maß erfüllen, in die Kategorie „Angst-Depression-gemischt“ fallen.

Da das F-DIPS speziell für die Forschung entwickelt wurde, sind die therapierelevanten Fragen, die noch im DIPS zum DSM-III-R vorhanden waren, zum größten Teil nicht mehr enthalten. Dagegen ist es möglich, eine Lebenszeit-Diagnose (die gesamte Vorgeschichte des Patienten wird dazu mit berücksichtigt, aktuell muss die Symptomatik dazu nicht vorhanden sein) zu stellen, während sich die Diagnosen im DIPS ausschließlich auf den present state (Querschnittsdiagnose) bezogen.

Tritt eine Störung mehrmals im Leben auf, wie z.B. eine rezidivierende depressive Störung, wird laut Manual immer die aktuelle Depressivität eruiert. Falls diese Episode gleichzeitig die „schlimmste“ Episode ist, brauchen die früheren Episoden nicht mehr überprüft zu werden. Ist eine frühere Episode die „Schlimmste“, muss diese vollständig befragt werden.

Das Manual ist gegliedert in einen soziodemografischen Teil, den Frageteil zu den Störungen auf Achse I, eine medizinische Anamnese (Behandlungen, körperliche Erkrankungen, Medikamente, Familienanamnese), in eine Einschätzung des Funktionsniveaus (Achse V) und in den Teil der Auflistung der DSM-IV-Kriterien für die einzelnen Störungsbilder. Außerdem werden die Patienten innerhalb des Interviews (eine der letzten Fragen) danach gefragt, was sie selbst für ihr Hauptproblem halten. Die Frage erscheint im Gegensatz zu einem klinischen-psychiatrischen Interview (dort ist es die Eingangsfrage) erst so spät, damit der Interviewer während des Interviews nicht hypothesengeleitet, sondern objektiv vorgeht.

Im Frageteil zu den Störungen ist die Abfolge der Fragen an den Störungsbildern orientiert, so dass jede Störung mit all ihren Symptomen vollständig erfragt wird bzw., falls keine positiven Anzeichen für die Störung vorliegen, wird mit Hilfe der Sprungregel der Rest des Störungkapitels hin zur nächsten Störung übersprungen.

Damit ist die Dauer des Interviews abhängig von der Anzahl der angegebenen auffälligen Informationen des Patienten, die es trotz vielleicht im Endeffekt nur wenigen tatsächlichen Störungen notwendig machen, den Störungsbereich vollständig zu erfragen.
Die Durchführungsdauer beträgt zwischen 40 und 180 Minuten, die Auswertungsdauer zusätzlich 20 – 30 Minuten, ebenfalls abhängig von der Anzahl der Störungen.

Die Auswertung der Antworten erfolgt anhand der im Anhang angegebenen DSM-Kriterien. Während aus Forschungsgründen und klinischem Interesse beim Einschätzen der Symptome eine Abstufung im Schweregrad von 0 – 8 möglich ist, erfolgt bei der Verwendung dieses Schweregrad-Ratings zur Diagnosestellung eine Umformulierung in eine dichotome Bewertung: ein Rating <4 gilt als klinisch unauffällig, ein Rating ≥ 4 spricht für ein klinisch relevantes Ausmaß der Symptomatik. D.h. um das Rating durchzuführen, muss ein klinisches Urteil gefällt werden.
Gleichzeitig wird mit dem Rating ein dimensionales System angewendet, das dann in ein klassifikatorisches überführt wird.

Beispielhafte Darstellung einer gekürzten Stelle aus dem diagnostischen Interview F-DIPS. Die möglichst wörtlich zu stellenden Fragen sind fett, der Erläuterungstext für den Interviewer ist kursiv gedruckt. Ergänzend zu dem Interviewleitfaden wird ein Protokollbogen verwandt.

GENERALISIERTE ANGSTSTÖRUNG

1. Eingangsfragen

A379/380 *Die Fragen dieses Abschnittes dienen zur Feststellung von Anspannung bzw. Angst ohne einen für die Patienten ersichtlichen Grund, oder von Angst in Zusammenhang mit exzessiver Sorge über familiäre, berufliche, finanzielle und ähnliche Angelegenheiten sowie Sorgen über geringfügige Anlässe. Diese Anspannung oder Angst ist nicht Teil bzw. ensteht nicht in Antizipation von Panikanfällen oder phobischen Ängsten. Die Sorgen und die Anspannung sind auch nicht Teil einer affektiven oder psychotischen Störung.*

Haben Sie sich im letzten Jahr über mehrere Dinge sehr viele Sorgen gemacht oder sich deswegen ängstlich gefühlt? *A379 J/N*

Falls JA:
Welche Sorgen waren das?
(Versichern Sie sich, daß die Sorgen der Probandin nicht in der Erwartung von Panikanfällen oder der Konfrontation mit phobischen Situationen begründet sind. Fragen Sie gegebenenfalls nach, ob Ängstlichkeit oder Sorgen auch bezüglich anderer Dinge auftreten.) *A380 (Text)*

Falls JA bei A379, weiter mit A388/417.

2. Symptom-Ratings

A388/417 *Falls die Probandin nicht von einer Phase anhaltender Sorgen berichtet (NEIN bei A379), dann befragen Sie sie dennoch bezüglich der aktuellen Sorgen.*
Spezifizieren Sie jeweils den Gegenstand der Sorgen (Text), auch wenn die Information bereits in A380 erhoben wurde.
Zusätzliche Fragen können nötig sein, um zu bestimmen, ob die berichteten Sorgen nicht in Zusammenhang mit einer eventuell gleichzeitig bestehenden Störung der Achse I stehen. Wenn eindeutig geklärt ist, daß ein Bereich, auf den sich Sorgen beziehen, ganz auf eine unter Achse I diagnostizierte Störung zurückgeführt werden kann, dann schätzen Sie diesen Bereich mit "0" ein.

Aktuell:
Ich werde Ihnen jetzt einige Fragen über aktuelle Sorgen bezüglich verschiedener Lebensbereiche stellen.

Beurteilen Sie jeden Bereich, auf den sich die Sorgen beziehen, getrennt nach der Ausprägung der Sorgen (Häufigkeit/Schweregrad) und der wahrgenommenen Unkontrollierbarkeit, indem Sie die nachfolgenden Skalen und Fragen benutzen.

Ausprägung der Sorgen:

0	1	2	3	4	5	6	7	8
nie besorgt/ keine Anspannung		selten besorgt/ leichte Anspannung		gelegentlich besorgt/ mäßige Anspannung		häufig besorgt/ schwere Anspannung		immer besorgt/ sehr schwere Anspannung

Wie häufig sorgen Sie sich über_____? Sorgen Sie sich auch dann über _____, wenn alles gut zu laufen scheint? Wieviel Anspannung und Angst erzeugen die Sorgen über_____bei Ihnen?
Unkontrollierbarkeit:

0	1	2	3	4	5	6	7	8
nie/ keine Schwierigkeiten		selten/ leichte Schwierigkeiten		gelegentlich/ mäßige Schwierigkeiten		häufig/ starke Schwierigkeiten		immer/ sehr starke Schwierigkeiten

Ist es schwer für Sie, die Sorgen über_____ zu kontrollieren, d.h. haben Sie Schwierigkeiten, die Sorgen zu beenden, oder ist es schwer für Sie, die Sorgen über_____ zu kontrollieren, so daß sie sich Ihnen aufdrängen, wenn Sie versuchen, sich auf etwas anderes zu konzentrieren?

- **Kleinere Angelegenheiten, alltägliche Probleme (z.B. Pünktlichkeit oder kleinere Reparaturen)**		*A388*	*(Text)*
	Häufigkeit/Schweregrad:	*A389*	*(0-8)*
	Unkontrollierbarkeit:	*A390*	*(0-8)*
- **Arbeit oder Ausbildung**		*A391*	*(Text)*
	Häufigkeit/Schweregrad:	*A392*	*(0-8)*
	Unkontrollierbarkeit:	*A393*	*(0-8)*
- **Familie**		*A394*	*(Text)*
	Häufigkeit/Schweregrad:	*A395*	*(0-8)*
	Unkontrollierbarkeit:	*A396*	*(0-8)*
- **Finanzen**		*A397*	*(Text)*
	Häufigkeit/Schweregrad:	*A398*	*(0-8)*
	Unkontrollierbarkeit:	*A399*	*(0-8)*
- **Sozial/Zwischenmenschlich**		*A400*	*(Text)*
	Häufigkeit/Schweregrad:	*A401*	*(0-8)*
	Unkontrollierbarkeit:	*A402*	*(0-8)*
- **Eigene Gesundheit**		*A403*	*(Text)*
	Häufigkeit/Schweregrad:	*A404*	*(0-8)*
	Unkontrollierbarkeit:	*A405*	*(0-8)*
- **Gesundheit nahestehender Personen**		*A406*	*(Text)*
	Häufigkeit/Schweregrad:	*A407*	*(0-8)*
	Unkontrollierbarkeit:	*A408*	*(0-8)*
- **Gesellschaft bzw. Weltgeschehen**		*A409*	*(Text)*
	Häufigkeit/Schweregrad:	*A410*	*(0-8)*
	Unkontrollierbarkeit:	*A411*	*(0-8)*
- **Sonstiges**		*A412*	*(Text)*
	Häufigkeit/Schweregrad:	*A413*	*(0-8)*
	Unkontrollierbarkeit:	*A414*	*(0-8)*

Falls Hinweise für mehrere Bereiche vorliegen, auf die sich die Sorgen beziehen:
Welche dieser Sorgen haben die größten Auswirkungen auf Ihr Leben? *A418* *(Text)*

Falls keine Hinweise auf aktuelle (A338/417) exzessive/unkontrollierbare Sorgen vorliegen, (alle Ratings < 4), weiter mit A500 (Zwangsstörung).

3. Aktuelle Episode

A450 **Ich möchte Ihnen jetzt eine Reihe von Fragen über die vergangenen Monate stellen, in denen die Sorgen aufgetreten sind.**

Führen Sie die wichtigsten Sorgen auf (vgl. A418): *(Text)*

Wurden Sie in den vergangenen 6 Monaten durch diese Symptome an der Mehrzahl der Tage beeinträchtigt?
A450 *J/N*

A451 **Wieviel Stunden am Tag sind Sie innerhalb der letzten Monate an einem durchschnittlichen Tag besorgt?**
A451 (Stunden)

Was befürchten Sie, könnte infolge von _____ passieren? *(Text)*

A452/463 **Ich werde mit Ihnen nun einige Dinge durchgehen, die zusammen mit den Sorgen auftreten können. Haben Sie während der vergangenen 6 Monate häufig_____ erlebt, wenn Sie sich Sorgen machten? Kam _____ an mehr als der Hälfte der Tage während der letzten 6 Monate vor?** *(Berücksichtigen Sie keine Symptome, die im Zusammenhang mit anderen Störungen auftreten, wie z.B. Panikanfälle, Sozialphobie usw.)*

Beurteilen Sie die Häufigkeit bzw. den Schweregrad der Symptome anhand der folgenden Skala (ein kombiniertes Rating):

0	1	2	3	4	5	6	7	8
nie besorgt/ keine Anspannung		selten besorgt/ leichte Anspannung		gelegentlich besorgt/ mäßige Anspannung		häufig besorgt/ schwere Anspannung		immer besorgt/ sehr schwere Anspannung

-Ruhelosigkeit, Nervosität
Häufigkeit/Schweregrad: *A452* *(0-8)*
Mehr als die Hälfte der Zeit? *A453* *J/N*

-Leichte Ermüdbarkeit
Häufigkeit/Schweregrad: *A454* *(0-8)*
Mehr als die Hälfte der Zeit? *A455* *J/N*

-Konzentrationsschwierigkeiten oder plötzlich etwas vergessen
Häufigkeit/Schweregrad: *A456* *(0-8)*
Mehr als die Hälfte der Zeit? *A457* *J/N*

-Reizbarkeit
Häufigkeit/Schweregrad: *A458* *(0-8)*
Mehr als die Hälfte der Zeit? *A459* *J/N*

-Muskelspannung
Häufigkeit/Schweregrad: *A460* *(0-8)*
Mehr als die Hälfte der Zeit? *A461* *J/N*

-Probleme einzuschlafen oder durchzuschlafen, unruhiger oder oberflächlicher Schlaf
Häufigkeit/Schweregrad: *A462* *(0-8)*
Mehr als die Hälfte der Zeit? *A463* *J/N*

A464/465 **Werden Sie durch die Sorgen und die damit in Zusammenhang stehende Anspannung und Angst in Ihrem Leben beeinträchtigt? Wie stark beeinträchtigt bzw. belastet Sie das in Ihrem Leben (z.B. Tagesablauf, Arbeit, soziale Aktivitäten)?**

Beurteilen Sie die Beeinträchtigung und die Belastung anhand der folgenden Skala:

0 ..... 1 ..... 2 ..... 3 ..... 4 ..... 5 ..... 6 ..... 7 8
gar nicht ein wenig mäßig schwer sehr schwer

Beeinträchtigung: A464 (0-8)
Belastung: A465 (0-8)

A466 **Haben Sie in der Zeit, in der Sie sich Sorgen machten und angespannt und ängstlich waren, regelmäßig irgendeine Art von Medikamenten oder Drogen zu sich genommen?**
(Diese Frage dann bejahen, wenn die Medikamente/Drogen oder deren Absetzen Ursache der Sorgen sein könnten.) *A466 J/N*
Spezifizieren Sie Art, Dosis, Zeitpunkte der Einnahme: (Text)

A467 **Lagen in dieser Zeit besondere körperliche Erkrankungen wie z.B. Schilddrüsenüberfunktion (Hyperthyreose) vor?** *A467 J/N*
Spezifizieren Sie Art, Beginn und Ende: (Text)

A468 **Wann begannen die Sorgen und Symptome der Angst/Anspannung in dem Sinn ein Problem zu werden, daß sie anhaltend auftraten, Sie sich durch die Symptome oder Sorgen belastet fühlten und sie Schwierigkeiten hatten, die Sorgen zu kontrollieren?**
(Falls die Probandin keinen genauen Zeitpunkt angeben kann, versuchen Sie, genauere Informationen zu erhalten, indem Sie z.B. den Beginn mit objektiven Lebensumständen verbinden.)

Beginn: A468 (M/J)

5.3.1 Bisherige Untersuchungen zu Gütekriterien des F-DIPS

Bisher liegt zur Interrater-Reliabilität des F-DIPS eine Untersuchung innerhalb einer Diplomarbeit (Satlow, 1996) vor. Dazu wurde zum einen die Übereinstimmung von Diagnosen am Ende einer Interviewerschulung betrachtet, dann die Übereinstimmung von Diagnosen und Störungsoberklassen anhand von Tonbandaufnahmen, die durch die Diplomandin ein zweites Mal geratet wurden, und zum dritten wurde die Übereinstimmung bzgl. der Einschätzung der Beeinträchtigung und Belastung eruiert. Beim letzten Punkt ging es um die Frage, inwieweit klinisch unerfahrene Interviewer Störungsbeeinträchtigungen ähnlich einschätzen, sprich erkennen können.

In Bezug auf die Übereinstimmung von Diagnosen einer professionellen Interviewerin mit denen der 60 Schulungsteilnehmer kam es zu folgenden Ergebnissen bei der Beurteilung von vier Testvideos:
Auf Video wurde die Diagnoseerhebung einer Panikstörung mit Agoraphobie (.69), einer Agoraphobie ohne Panikstörung (.69), einer Sozialphobie (.95), einer spezifischen Phobie (.43), einer Major Depression (.48) und einer Enuresis (.78) demonstriert. In Klammern befinden sich die Kappa-Koeffizienten (κ).

Die Ergebnisse der Übereinstimmung zwischen Interwiewern und der gegenkodierenden Diplomandin anhand von Tonbandaufnahmen lauten folgendermaßen:
Aufgrund zu niedriger Basisraten konnten nur Kappa-Werte für Störungsoberklassen sowie für die Sozialphobie (.78), die spezifische Phobie (.93), die Major Depression frühere Episode (.66) und die Anorexia nervosa (1) berechnet werden. In Bezug auf Angststörungen insgesamt ergab sich ein κ = .79 und bei Affektiven Störungen von .58.

Ob die Beeinträchtigung oder die Belastung in ihrer klinischen Relevanz übereinstimmend beurteilt wurden, wurde für die Sozialphobie, die spezifische Phobie und die Major Depression untersucht. Es ergaben sich

dabei gute Übereinstimmungen (κ zwischen .71 und .89) für Sozialphobie und spezifische Phobie sowie eine mittlere bis starke Übereinstimmung für die Major Depression (.56 und .65).

Die Aufnahme patientenbezogener Informationen soll durch diagnostische Interviewleitfäden erleichtert, systematisiert und besser operationalisiert werden, um somit die Reliabilität von Diagnosen zu erhöhen.
Das Kapitel „Der diagnostische Prozess“ stellte eine Reihe solcher diagnostischer Interviews vor, halb strukturierte, vollständig strukturierte und standardisierte Interviews. Beim F-DIPS handelt es sich um ein voll strukturiertes Interview, das vorformulierte Fragen und festgelegte Sprungregeln enthält, dem Interviewer jedoch die Freiheit lässt, Widersprüche oder Unklarheiten zu klären.

6 Fragestellung

Die Komplexität psychischer Störungen in ihrem Erscheinungsbild mit fließenden Übergängen zwischen verschiedenen Störungen und einem unzureichenden Wissen über Entstehungsbedingungen (Möller, 1998) macht einen Großteil der Probleme mit Klassifikationen psychischer Störungen deutlich.
Es gibt, nachdem in den letzten 25 Jahren verstärkt die Forschungsenergie in die Entwicklung syndromatologischer Klassifikationssysteme geflossen ist, aktuell zwei Klassifikationssysteme, die ICD-10 (International Statistical Classification of Diseases, WHO, 1991) sowie das DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, APA, 1994) mit besonders forschungsspezifischen Vorteilen für das letztere System.
Klassifikationssysteme sollen sowohl in der Forschung als auch in der klinischen Praxis helfen, Patientengruppen zu charakterisieren und vergleichbar zu machen, die Kommunikation zu vereinheitlichen und zu vereinfachen, Behandlungsmaßnahmen abzuleiten und zu überprüfen. Trotz der kontinuierlichen Verbesserung der Klassifikationssysteme wird auch an ihnen Kritik in der Form geübt, dass die diagnostisch relevanten Merkmale zu unspezifisch beschrieben und quantifiziert werden, so dass dadurch einzelne Diagnosen zu breit definiert sind und sich nur unscharf voneinander abgrenzen lassen. Auch werden die Systeme oft nicht sorgfältig gebraucht, sondern es fließen traditionelle Modelle und klinische Gepflogenheiten mit in die Diagnostik ein.
Um dies auszuschließen, wurden eine Reihe von diagnostischen Interviewleitfäden entwickelt.

Um die Güte der Diagnostik psychischer Störungen mittels verschiedener Klassifikationssysteme bzw. verschiedener Interviews zu beurteilen, wird in zahlreichen Studien die Retestreliabilität und die Validität der Diagnosen untersucht. Die Reliabilität gibt die Genauigkeit an, mit der in einem Interview eine Diagnose erfasst werden kann, wobei bei Klassifikationen zur Ermittlung der Reliabilität ausschließlich die Testwiederholungsmethode eingesetzt wird, so dass betrachtet wird, inwieweit sich

die Diagnose des ersten Interviews mit der Diagnose aus der zweiten Durchführung des Interviews deckt.

Das neu entwickelte strukturierte „Diagnostische Interview bei Psychischen Störungen in der Forschungsversion“ (F-DIPS: Margraf et al., 1996) wurde bisher vor allem in der epidemiologischen Studie zu „Prädiktoren psychischer Gesundheit/Krankheit bei jungen Frauen“ eingesetzt und ist im Stile des für das DSM-III-R gültigen DIPS (Margraf et al., 1991) konstruiert. Mit seiner Hilfe ist es möglich, eine Auswahl an Achse I- Störungen (Angststörungen, Affektive Störungen, Somatoforme Störungen, Substanzmissbrauch und –abhängigkeit, Essstörungen und Störungen des Kindes- und Jugendalters) zu diagnostizieren, eine soziale und medizinische Anamnese zu erheben und eine Einschätzung des allgemeinen Funktionsniveaus (Achse V des DSM-IV) vorzunehmen.
Ob das F-DIPS reliable und valide Diagnosen treffen kann, soll hier überprüft werden.

Dazu werden die folgenden Fragen beantwortet:

1. Wie hoch ist die Übereinstimmung der F-DIPS-Diagnosen von zwei im Abstand von 2 Wochen (±1 Woche) durchgeführten Interviews durch unabhängige Rater am selben Patienten?

2. Inwieweit stimmen die Interviewer auch in der Vergabe komorbider Störungen überein?

3. Spielen die klinisch-psychiatrische Erfahrung oder die Erfahrung in der Anwendung des F-DIPS eine Rolle für die Reliabilität der Diagnosen?

4. Gibt es eine höhere Übereinstimmung der Diagnosen, wenn eine hohe Sicherheit bei der Vergabe der Diagnose vorlag?

5. Existieren konfundierende Variablen, die die Genauigkeit der Messung mit dem F-DIPS beeinflussen?

6. Gibt es besonders häufige Fehlerquellen bei der Anwendung des F-DIPS?

7. Inwiefern diagnostiziert das F-DIPS die Störung, die der Patient tatsächlich hat?

   Es bestehen folgenden Hypothesen zur Validität des F-DIPS:

   - Patienten mit einer depressiven Störung im F-DIPS zeigen höhere Werte im BDI und in der SCL-90-R-Skala „Depressivität“ als Patienten ohne Depression. Sie erhalten auch eine Entlassungsdiagnose „Depressive Störung“, „Depressive Reaktion“ oder „Neurotische Depression“.
   - Patienten mit einer Angststörung im F-DIPS zeigen im BAI und auf den SCL-90-R-Skalen „Unsicherheit im Sozialkontakt“, „Ängstlichkeit“ und “Phobische Angst“ höhere Werte als Patienten ohne Angststörungen. Außerdem erhalten sie als Entlassungsdiagnose eine Angststörung (Soziale Phobie, Panikstörung mit/ohne Agoraphobie, Generalisierte Angststörung, Phobie).
   - Patienten mit einer Somatoformen Störung im F-DIPS haben höhere Werte als andere Patienten im Whiteley-Index und in der SCL-90-R-Skala „Somatisierung“ und erhalten eine entsprechende Entlassungsdiagnose (Somatoforme Schmerzstörung, Somatisierungsstörung, Konversionsstörung, Hypochondrie).
   - Patienten mit einer Substanzabhängigkeit bzw. –missbrauch im F-DIPS erhalten diese Diagnose auch in der Entlassungsdiagnose.
   - Patienten mit einer Essstörung im F-DIPS erhalten diese Diagnose (Anorexia nervosa, Bulimia nervosa) auch in der Entlassungsdiagnose.

8. Wie zuverlässig und valide misst das F-DIPS psychische Störungen im Vergleich zu anderen diagnostischen Interviews?

7 Empirischer Teil

7.1 Durchführung der Untersuchung

In der im folgenden beschriebenen Untersuchung wird sowohl die Reliabilität des strukturierten „Diagnostischen Interviews bei Psychischen Störungen – Forschungsversion (F-DIPS)“ als auch die Validität ermittelt.

Zur Erfassung der Reliabilität wird die Testwiederholungsmethode eingesetzt, und das F-DIPS in einem zeitlichen Messabstand von ca. 2 Wochen von zwei unabhängigen Ratern durchgeführt. Dieses Vorgehen ist bei entsprechenden Überprüfungen der Reliabilität von Klassifikationen üblich (Wittchen et al., 1991; Wittchen et al., 1998; Schneider et al., 1992; Hiller et al., 1997; DiNardo et al., 1983; Semler et al., 1987), wobei der zeitliche Abstand zwischen den beiden Interviews groß genug sein sollte, dass Erinnerungseffekte gering gehalten werden können, und gleichzeitig ist der Abstand nicht so groß, dass sich die Einschätzung der Beschwerden allzu sehr verändert haben kann.

Außerdem wird die Jahresprävalenz der Störungen erfasst – mit Ausnahme bei den affektiven Störungen, bei denen es um die Lebenszeitprävalenz geht. Gründe für die Erfassung der Jahresprävalenz liegen zum einen darin, dass in der Studie zu „Prädiktoren der Gesundheit junger Frauen“ das F-DIPS zu zwei Zeitpunkten durchgeführt, wobei zum ersten Zeitpunkt die aktuelle und die Lebenszeit-Diagnose eruiert wurde, beim zweiten Zeitpunkt dann nur noch die Zeit seit dem letzten Interview (zur Ermittlung der Inzidenz). Dieser Zeitraum belief sich auf ein bis 1 ½ Jahre. Da für die „Retest-Reliabilitäts-Untersuchung“, auf die Studien-Stichprobe aus der zweiten Messung zurückgegriffen wird, wird der Befragungszeitraum für eine Störung daran angelehnt und die Prävalenz für das vergangene Jahr erfasst. Dies hat zudem den Vorteil, dass sich damit der Befragungszeitraum bei beiden Interviews deckt. (Bei Fragen nach den letzten vier oder zwei Wochen, wie sie bei vielen Störungen üb-

lich sind, würden sonst beide Interviewer einen etwas voneinander abweichenden Zeitraum befragen und könnten deshalb zu unterschiedlichen Resultaten (Diagnosen) kommen.

Die Bestimmung der Validität der im F-DIPS gestellten Diagnosen wird durch den Vergleich mit verschiedenen Selbstauskunfts-Fragebögen und der Entlassungsdiagnose der behandelnden Ärzte und Psychologen ermittelt. Zudem kann die Diagnose mit der Auskunft zum Hauptproblem durch den Patienten verglichen werden.

Die Patientinnen und Patienten erhalten gleichzeitig mit dem ersten Interview eine Fragenbogenbatterie mit den folgenden Verfahren:

- SCL-90-R (Symptom-Checkliste)
- BDI (Beck Depressionsinventar)
- BAI (Beck Angstinventar)
- WHITELEY-Index
- SKID-II (Strukturiertes Klinisches Interview für DSM-IV, Achse II: Persönlichkeitsstörungen)

Mit dem Strukturierten klinischen Interview für DSM-IV – Persönlichkeitsstörungen (SKID-II) sollen Hinweise auf Persönlichkeitsstörungen bei der Untersuchungsstichprobe erfasst werden, um eine Aussage darüber treffen zu können, ob das Vorliegen einer Persönlichkeitsstörung einen Einfluss auf die Übereinstimmung der Interviewer in ihren Achse-I-Diagnosen hat.
Das SKID-II ist ein halbstrukturiertes Interview mit einem zweistufigen Verfahren, das aus einem Fragebogen und einem anschließenden verbalen Überprüfen der „ja“-Antworten zur Erfassung von 12 Persönlichkeitsstörungen (selbstunsichere, dependente, zwanghafte, negativistische, depressive, paranoide, schizotypische, schizoide, histrionische, narzisstische, Borderline, antisoziale) besteht. Die gewonnenen Werte können zum einen für eine kategoriale, aber auch für eine dimensionale Diagnostik hinsichtlich entsprechender Persönlichkeitsmerkmale verwendet werden.

Hier soll lediglich aus ökonomischen Gründen ein Screening zum Verdacht auf eine Persönlichkeitsstörung durchgeführt werden, indem ausschließlich der Fragebogen vorgelegt und auf das ergänzende Interview, das im SKID-II eigentlich verlangt wird, verzichtet wird.

Außerdem wird das Hauptproblem, das die Patienten für sich selbst sehen, innerhalb des Interviews erfasst.

Die Entlassungsdiagnosen, die meistens in der ICD-10-Nomenklatur verfasst sind, werden aus den Abschlussberichten zu den entsprechenden Patienten herausgesucht.

Um Stimmungsveränderungen vom 1. zum 2. Interview kontrollieren zu können sowie den eventuellen Einfluss auf die Beantwortung der Interviewfragen, wird den Patienten auch beim zweiten Interview das BDI vorgelegt.

Zudem wird ein Teil der Patienten dazu befragt, ob sie Unterschiede im Ablauf der beiden Interviews wahrgenommen haben bzw. selbst in den beiden Interviews unterschiedliche Auskünfte gegeben haben und ob sie Vertrauen zum Interviewer fassen konnten. Diese Fragen werden mit Hilfe eines kurzen Fragebogens erhoben:

	Stimmt überhaupt nicht	Eher nicht	Teils teils	Stimmt in etwa	Stimmt völlig	Mittelwert	Standardabweichung
	0	1	2	3	4		
1. Ich habe mich beim 2. Interview noch gut an die Fragen aus dem 1. Interview erinnert.						2,5	1
2. Ich fand das 1. Interview angenehmer.						1,3	1,1
3. Ich fand das 2. Interview angenehmer.						2,3	1,3
4. Ich habe den Eindruck, ich habe in beiden Interviews die gleichen Auskünfte gegeben.						2,9	0,7

	Stimmt überhaupt nicht	Eher nicht	Teils teils	Stimmt in etwa	Stimmt völlig	Mittelwert	Standardabweichung
	0	1	2	3	4		
5. Ich habe in den beiden Interviews verschieden geantwortet, weil						1,2	0,9
▪ ich mir nach dem 1. Interview noch Gedanken dazu gemacht habe und zu anderen Schlüssen kam.						1,3	1,2
▪ ich in einer anderen Stimmung war.						1,8	1,4
▪ die Fragen anders waren.						1	0,9
▪ der Interviewer als Person anders war.						1,9	1,5
6. Ich habe zu dem 1. Interviewer Vertrauen gehabt.						2,9	1,1
7. Ich habe zu dem 2. Interviewer Vertrauen gehabt.						3,2	0,9

Damit außerdem Fehlerquellen genauer untersucht werden können, werden so häufig wie möglich Videoaufnahmen durchgeführt.

7.2 Beschreibung der Instrumente zur Validierung

Die Validität des F-DIPS soll wie die Validität des DIPS (1991) anhand einer Fragebogenbatterie, jedoch zusätzlich anhand der Entlassungsdiagnose bestimmt werden („Expertenurteil“). Die Validierung anhand von Expertenurteilen wird zwar noch im DIPS (1991) als problematisch angegeben, da von einer begrenzten Reliabilität der klinischen Diagnosen ausgegangen wird (vgl. Margraf et al., 1991, 1994; Margraf, 1996) und damit als Außenkriterium zur Validierung von Diagnosen anfechtbar sind. Auf der anderen Seite gibt es eine Reihe von Studien, die klinische Diagnosen als Kriterium verwenden (Peters & Andrews, 1995; Kranzler et al., 1996; Rosenman et al., 1997). Zudem zeigt die Studie von Rosenman et al. (1997, s.o.) eine gute Reliabilität von klinischen Diagnosen auf der Basis der ICD-10-Verschlüsselung. Klinische Diagnosen sind außerdem in der Praxis die, aufgrund derer eine Behandlung durchgeführt wird. Damit sind sie von klinischer Relevanz.

Die Auswahl der Fragebögen ist durch die geplante Durchführung dieser Untersuchung in Anlehnung an die Studie „Prädiktoren psychischer Gesundheit bei jungen Frauen“ zu erklären. Die oben vorgestellten Instrumente wurden auch in dieser Studie vorgegeben, so dass die Studienprobandinnen nicht noch durch weitere, eventuell für eine Validitätsstudie besser geeignete Instrumente belastet werden sollen.
Es werden im einzelnen folgende Validierungsinstrumente verwendet:

7.2.1 Beck-Depressions-Inventar (BDI)

Das Beck-Depressionsinventar (BDI) zählt zu den am meisten verwendeten Selbstbeurteilungsverfahren zur Messung von Depression. Zur Messung der Validität anderer Verfahren dient das BDI häufig als Kriterium (Richter et al., 1994). Jedoch ist das BDI äußerst änderungssensitiv, wobei sich auch von Therapiemaßnahmen unabhängige signifikante Mittelwertsunterschiede der BDI-Summenwerte im Abstand von einem Tag ergeben (Richter, 1991). Dies ist natürlich unerwünscht, will man Therapieeffekte kontrollieren. Es zeigte sich jedoch in vielen Studien, dass das BDI zwischen psychiatrischen Stichproben und klinisch unauffälligen Kontrollpersonen gut differenziert (vgl. Richter, 1994). Nach Jacobson & Revenstorf (1988) gelten Werte unter 9 als nicht-depressiv, zwischen 10 und 15 als Hinweis für dysphorisch, zwischen 16 und 19 als Hinweis auf eine depressive Verstimmung, zwischen 20 und 30 als Hinweis für eine mittelschwere Depression und über 30 als schwere Depression.

7.2.2 Beck-Angst-Inventar (BAI)

Das Beck-Angst-Inventar (Margraf & Ehlers, 1998) gibt Auskunft über die Schwere von Angstsymptomen und hat sich zur Erfassung klinisch relevanter Ängste als geeignet erwiesen. Es trennt Angst dabei gut von depressiven Symptomen. Als auffällig bei einer ostdeutschen Normver-

gleichsstichprobe können bei Männern Werte über 9, bei Frauen Werte über 13 gesehen werden (Abweichung > 1 Standardabweichung).

7.2.3 Symptom-Checkliste (SCL-90-R)

Bei der Symptom-Checkliste von Derogatis (Franke, 1995) handelt es sich um ein international verbreitetes Instrument zur Erfassung unterschiedlicher Grade psychischer Beeinträchtigung. Die Checkliste besteht aus 90 Items und 9 Skalen: Somatisierung, Zwanghaftigkeit, Unsicherheit im Sozialkontakt, Depressivität, Ängstlichkeit, Aggressivität/Feindseligkeit, Phobische Angst, Paranoides Denken und Psychotizismus. Außerdem gibt es drei globale Kennwerte, wobei der GSI die grundsätzliche psychische Belastung misst, der PSDI die Intensität der Antworten und der PST Auskunft gibt über die Anzahl der Symptome, bei denen eine Belastung vorliegt. Die Einzelitems werden von den Patienten in Bezug auf die vergangenen 7 Tage hinsichtlich ihrer Intensität (5-stufiges Rating von „überhaupt nicht“ bis „sehr stark“) beurteilt und besitzen eine klinische Relevanz. Das Verfahren ist zur Veränderungsmessung geeignet. In verschiedenen Studien (Rief et al., 1991; Stieglitz, 2000) ergaben sich hohe Interkorrelationen zwischen den postulierten Skalen. Es ergeben sich über viele Studien hinweg Hinweise auf einen allgemeinen Faktor psychischer Beeinträchtigung. Das Verfahren ist damit zwar zur globalen Veränderungsmessung gut geeignet, erlaubt jedoch kaum eine Diskriminierung zwischen verschiedenen Krankheitsgruppen (Rief et al., 1991). Obwohl die von Derogatis & Cleary (1977) postulierte Faktorenstruktur nicht vollständig repliziert werden kann, weisen die Skalen hohe Werte für die innere Konsistenz auf und können unterschiedliche diagnostische Gruppen befriedigend beschreiben und differenzieren (Stieglitz, 2000).

7.2.4 Whiteley-Index

Der Whiteley Index ist ein kurzes Screening Instrument für hypochondrische und somatoforme Bescherden. Es werden nicht einzelne Symptome, sondern hypochondrische Einstellungen gegenüber den eigenen Körperfunktionen und körperlichen Krankheiten erfragt. Ein Summenwert von mindestens 8 im Whiteley-Index stellt ein Kriterium für Hypochondrie dar.
Drei Faktoren konnten extrahiert werden: Krankheitsängste, Somatische Bescherden und Krankheitsüberzeugung (Rief et al., 1994)
Auf den Dimensionen *Krankheitsängste* und *somatische Beschwerden* konnten durch ein verhaltensmedizinisches Vorgehen signifikante Verbesserungen erreicht werden, während der Bereich *Krankheitsüberzeugung* veränderungsresistenter erschien.

7.2.5 Selbsteinschätzung des Hauptproblems

Hier gibt der Patient frei an, was aus seiner Sicht sein Hauptproblem darstellt, ohne dass dies einen diagnostischen Namen beinhalten muss. Dies bedeutet, dass die Angaben des Patienten, z.B. „das Essen ist mein Hauptproblem“ in eine diagnostische Kategorie überführt werden muss, um die Vergleichbarkeit mit einer Diagnose zu gewährleisten. Die Angaben sind damit nur bedingt verwendbar. Es ist vorgesehen, nur Problemoberklassen zu formulieren (Affektive Störung, Angststörung, Somatoforme Störung, Substanzmissbrauch und –abhängigkeit, Essstörung, andere Probleme). Als Außenkriterium dient es auch deshalb nur begrenzt, da die F-DIPS-Diagnosen nicht in Haupt- und sekundäre Störungen unterteilt werden. Somit können sich F-DIPS-Diagnose und Selbsteinschätzung nur decken, wenn der Patient mehrere Problembereiche angibt oder nur eine F-DIPS-Diagnose vergeben wird. Dennoch ist interessant, ob der Patient schon in dieser kurzen Frage im Sinne einer Selbstdiagnose antworten kann.

7.2.6 Entlassungdiagnose

Diese unterscheidet sich von einer Diagnose aus einem klinischen Erstgespräch dadurch, dass ein längerer Beobachtungszeitraum durch die Therapie besteht, in dem die Diagnosen teilweise mehrfach diskutiert und verändert werden. Informationen von mehreren Therapeuten aus verschiedenen Beobachtungssituationen (Gruppentherapien, Einzeltherapie, Alltag zwischen den Therapien), der Selbstbericht des Patienten und fremdanamnestischen Daten durch Familienangehörige und Vordiagnosen anderer Behandler, sowie die F-DIPS-Diagnose (alle Behandler werden über das F-DIPS-Ergebnis informiert) können mit einfließen, so dass ein umfassenderes Bild des Patienten entstehen kann. Dies macht die Diagnose sicherer.

Zur Bestimmung der Validität des Interviewergebnisses werden in Bezug auf die Depression das BDI und die SCL-90-R-Skala Depressivität, in Bezug auf eine Angststörung das BAI und die SCL-90-R- Sklalen „phobische Angst, Ängstlichkeit, soziale Unsicherheit“, in Bezug auf die Somatoforme Störungen / Hypochondrie / Körperbezogene Ängste der Whiteley-Index und die Skala SCL-90-R-Somatisierung bestimmt. Alle Diagnosen werden mit der Entlassungsdiagnose verglichen. Essstörungen sowie Substanzabhängigkeiten werden lediglich anhand der Entlassungsdiagnosen validiert.

7.3 Stichprobe

Zur Durchführung der Untersuchung wurde eine Stichprobe aus 191 Personen zusammengestellt, die sich folgendermaßen rekrutieren ließ:

- 73 Patientinnen und Patienten aus einer psychosomatischen Rehabilitationsklinik (Median-Klinik Berggießhübel),
- 105 Patientinnen und Patienten aus der Universitätsklinik für Psychotherapie und Psychosomatik Dresden, davon 25 ambulant untersuchte Patienten und 80 stationär oder teilstationär behandelte Patienten,
- 11 Patientinnen und Patienten der Ambulanz der psychiatrischen Universitätsklinik Heidelberg und
- 2 Teilnehmerinnen aus der Studie „Prädiktoren psychischer Gesundheit junger Frauen“ in Dresden (Margraf & Becker).

Zusätzlich nahmen 24 Personen zwar am ersten Interview teil, lehnten die Teilnahme am 2. Interview jedoch ab oder waren oder waren nicht mehr erreichbar.

7.4 Die F-DIPS-Interviewerinnen und -Interviewer

Die Interviewer setzten sich zum großen Teil aus Psychologie-Studentinnen nach dem Vordiplom mit der Erfahrung eines klinischen Praktikums und eines „DIPS-Seminars“ an der Universität zusammen. Das Durchschnittsalter der Interviewer lag bei 27,3 Jahren (22-37 J.). Es handelte sich um 10 Frauen und 3 Männer. Einige (n=4) hatten bereits F-DIPS-Interviews im Rahmen der Studie „Psychische Gesundheit junger Frauen“ durchgeführt und nach einer Schulung in diesem Kontext schon über 50 Interviews geführt. Diese bereits aus der Studie erfahrenen Interviewer erhielten keine weitere Schulung durch mich, sondern nahmen nur noch an den Supervisionsbesprechungen teil.

Die von mir durchgeführte Schulung für die zukünftigen Interviewer bestand aus einer

- theoretischen Einführung in die Handhabung des Interviews
- selbständigen Durchführung mit einem Probanden aus dem Bekanntenkreis jedes Schulungsteilnehmers
- Beobachtung eines am Patienten durchgeführten Interviews mit gleichzeitiger Kodierung der Schulungsteilnehmer und
- Diskussion der Unklarheiten und Ratingmodalitäten
- zwei weiteren Interviews am Patienten durch die Schulungsteilnehmer (eine Gruppe mit 3-4 Personen), selbständig im Wechsel ausgeführt, mit ebenfalls einer Diskussion von auftretenden Problemen und Unklarheiten in Anwesenheit des Patienten
- diagnostische Einschätzung von zwei Videos (Gegenkodierung)
- vollständig alleine durchgeführten Interview mit einem Patienten unter Supervision
- weiteren Gegenkodierung von Video-Interviews.

Nach der Schulung wurden sämtliche geführten Interviews nachbesprochen, Unklarheiten und Diagnosen diskutiert in der Kleingruppe bzw. mit Einzelsupervision und Rückmeldungen zu den Interview-Videos.
Insgesamt soll durch die Schulung erreicht werden, dass die Interviewerfehler minimiert werden und ein einheitlicher Interviewstil mit hoher Interrater-Reliabilität etabliert wird.

Auch nach der Schulung, d.h. zum Zeitpunkt des ersten für die Reliabilitätsuntersuchung durchgeführten Interviews, unterschieden sich die Interviewer noch in ihrer Erfahrung. Der unterschiedliche Erfahrungsstand sowie die Anzahl der geführten Interviews von Interviewern dieses Erfahrungsstandes sind in der Tabelle aufgeführt.

Gruppe	Erfahrung	Anzahl der Interviewer	Anzahl der geführten Interviews
1	< 10 F-DIPS-Interviews (direkt nach der F-DIPS-Schulung) und klinisch-psychiatrische Erfahrung > 7 Jahre	2	53
2	> 50 F-DIPS-Interviews, ohne klinische Erfahrung (Schulung innerhalb des Projektes „Gesundheit junger Frauen“	4	106
3	< 10 F-DIPS-Interviews (direkt nach der F-DIPS-Schulung), ohne klinische Erfahrung	7	223

Die Paarung der 13 Interviewer erfolgte nach pragmatischen Kriterien, wobei, da die Untersuchung sich über zwei Jahre erstreckte, einige Paarungen nicht zustande kamen.

7.5 Supervision der durchgeführten Interviews

Die Interviewer führten nach der Schulung die F-DIPS-Interviews in eigener Verantwortung und werteten sie selbständig aus. Jedes geführte Interview wurde nach Abschluss der Auswertung anhand des Kodierungsbogens, des Berichts des Interviewers zum Patienten und dem allgemeinen Eindruck sowie, soweit vorhanden, anhand der Videoaufzeichnung mit mir, teils in Anwesenheit der nicht am 2. Interview beteiligen Interviewer, durchgesprochen. Dabei kamen besonders Kodierungszweifel und Fragen bzgl. der Komorbidität zur Sprache. Da beide Interviews am selben Patienten durch mich supervidiert wurden, versuchte ich, um das Interviewergebnis nicht zu beeinflussen, auch in den Fällen, in denen ich mich an das erste Interview erinnern konnte, mich streng an die erhaltenen Informationen zu halten und diese nicht mit dem anderen Interview zu vermischen. Teilweise wurden die Diagnosen jedoch korrigiert, wenn sich die Diagnosen nicht klar aus den Informationen ableiten ließen oder einzelne Symptome überbewertet wurden.

In einigen Fällen führte ich selbst eines der beiden Interviews. In diesen Fällen veränderte ich nichts an der Einschätzung des anderen Interviewers, da sich dadurch eine künstliche Erhöhung der Übereinstimmung ergeben hätte.

7.6 Fehleranalyse

Bei Nicht-Übereinstimmung werden die Fehlerquellen näher untersucht, indem ich die Kodierungsbögen auf die Antworten der Patienten hin, auf die Symptomgewichtungen und die Ableitungen der Diagnosen, auf die Vollständigkeit der Daten, die Vollständigkeit der gestellten Fragen anschaue. Außerdem existieren aus der Supervision der Interviews freie Notizen zu den Schwierigkeiten bei der Diagnosestellung und Notizen über beobachtete Schwierigkeiten aus den Videos. Die Interviewer selbst machten neben der Diagnosesicherheit auch Angaben zu den Schwierigkeiten, die sie beim Interview hatten, so dass auch diese Informationen einen Hinweis auf die Fehlerquellen geben können.
Die Fehler werden ausgezählt für jede Diagnose, die bei den beiden Interviews nicht übereinstimmt, so dass auch mehrere Fehler bei einem Patienten auftreten können.

7.7 Statistische Auswertung

Um zu sehen, wie zuverlässig eine Störung erfasst werden kann, wird die Übereinstimmung der in beiden Interviews ermittelten Diagnosen in einer 4-Felder-Tafel festgehalten, die prozentuale Übereinstimmung errechnet und sowohl der Kappa-Koeffizient (Cohen, 1960) als auch der Yule-Koeffizient (Spitznagel & Helzer, 1985; Shrout, Spitzer & Fleiss, 1987) berechnet, wobei letzterer unabhängig von der beobachteten Basisrate ist (Spitznagel & Helzer, 1985; Stieglitz, 2000).

Nach Fleiss (1981), Landis & Koch (1977) und Wittchen et al. (1998) wird von einer ausgezeichneten Übereinstimmung ausgegangen, wenn die Reliabilitätskoeffizienten > .75 liegen. Die anderen Bereiche werden von den verschiedenen Autoren leicht unterschiedlich bewertet. Im folgenden wird davon ausgegangen, dass Reliabilitätskoeffizienten <.40 eine ungenügende Übereinstimmung anzeigen. Zwischen .40 und .65 handelt es sich um eine mäßig bis mittlere Übereinstimmung und Werte >.65 zeigen nach Wittchen et al. (1998) eine gute Übereinstimmung an.

Die folgende Tabelle zeigt die in der durchgeführten Untersuchung verwendete Bewertung der Reliabilitätskoeffizienten im Überblick an.

Übereinstimmung	κ	Y
ungenügend	< .40	
mäßige bis mittlere	.40 - .65	.65 - .80
gut	.65 - .75	
ausgezeichnet	> .75	> .80

Um zu überprüfen, welche Einflussfaktoren auf die Nichtübereinstimmung einer Diagnose vorhanden sind, wird die Stichprobe unterteilt in die Gruppe der Patienten, bei denen die Interviewer zu einer übereinstimmenden diagnostischen Einschätzung kam und in die Gruppe, bei der die Interviewer zu divergierenden Einschätzungen kam. Als Einflussfaktoren werden untersucht:

- Sicherheit beim Erstellen einer Diagnose
- Abstand der beiden Interviews voneinander
- das Funktionsniveau der Patienten
- das Vorhandensein einer Persönlichkeitsstörung
- die Anzahl der vergebenen F-DIPS-Diagnosen
- die Dauer des Interviews
- die Depressivität (BDI)

- die Auskünfte der Patienten zu ihren F-DIPS-Interviews (Erinnerung an das erste Interview beim zweiten, Wohlfühlen, Vertrauen zum Interviewer, gleiche Antworten gegeben, verschiedene Antworten gegeben wegen neuer Sicht der Dinge, anderer Stimmung, Person des Interviewers).

Die beiden Gruppen (abhängige Variable) werden je nach Skalenniveau des untersuchten Merkmals mit einem t-Test (intervallskaliert) oder einem Mann-Whitney-U-Test (Ordinalskalierung) für unabhängige Stichproben verglichen.

Zur Ermittlung der Validität werden t-Tests für abhängige Stichproben durchgeführt, da aufgrund der Stichprobengröße von einer Normalverteilung ausgegangen werden kann. Die Unterschiede bzgl. des Medians werden zur besseren Anschaulichkeit in Boxplots dargestellt, in denen die Standardabweichung angegeben wird.

7.8 Beurteilung des Vergleichs mit anderen Instrumenten

Um das F-DIPS mit anderen diagnostischen Interviews vergleichen zu können, wird mittels der Literaturdatenbanksysteme „Medline“ und „PsycLit“ unter den Stichworten „diagnostic interview“, „psychiatric interview“, „reliability“, „validity“, „DSM-IV“, „ICD-10“, „SCID“, „SKID“, „SCAN“, „ADIS“, „CIDI“, „DIS“, „IDCL“ nach Angaben gesucht und die Reliabilitäts- und Validitätsdaten anhand der gefundenen Veröffentlichungen miteinander verglichen.

8 Ergebnisse

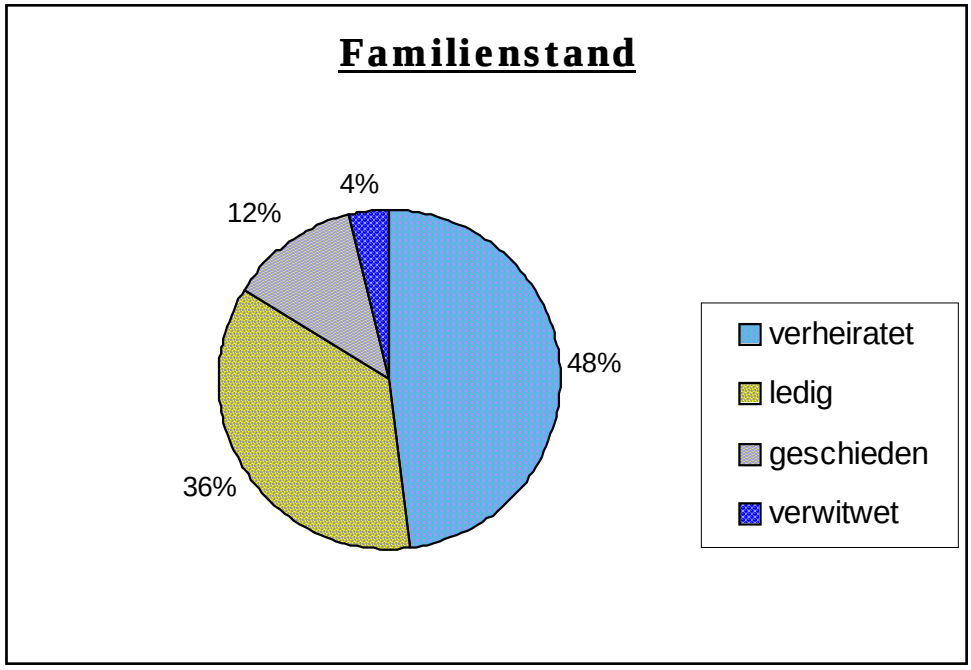
8.1 Stichprobenbeschreibung

Die Untersuchungsstichprobe setzt sich folgendermaßen zusammen: 191 Patientinnen und Patienten nahmen an der Untersuchung teil, davon kamen

- 73 aus einer psychosomatischen Rehabilitationsklinik,
- 80 waren stationäre oder teilstationäre Patienten aus einer Akutklinik und
- 36 waren Patienten aus zwei Ambulanzen von Akutkliniken.
- Außerdem nahmen zwei Probandinnen aus dem Dresdner Projekt teil.

78 % waren Frauen, 22 % Männer.

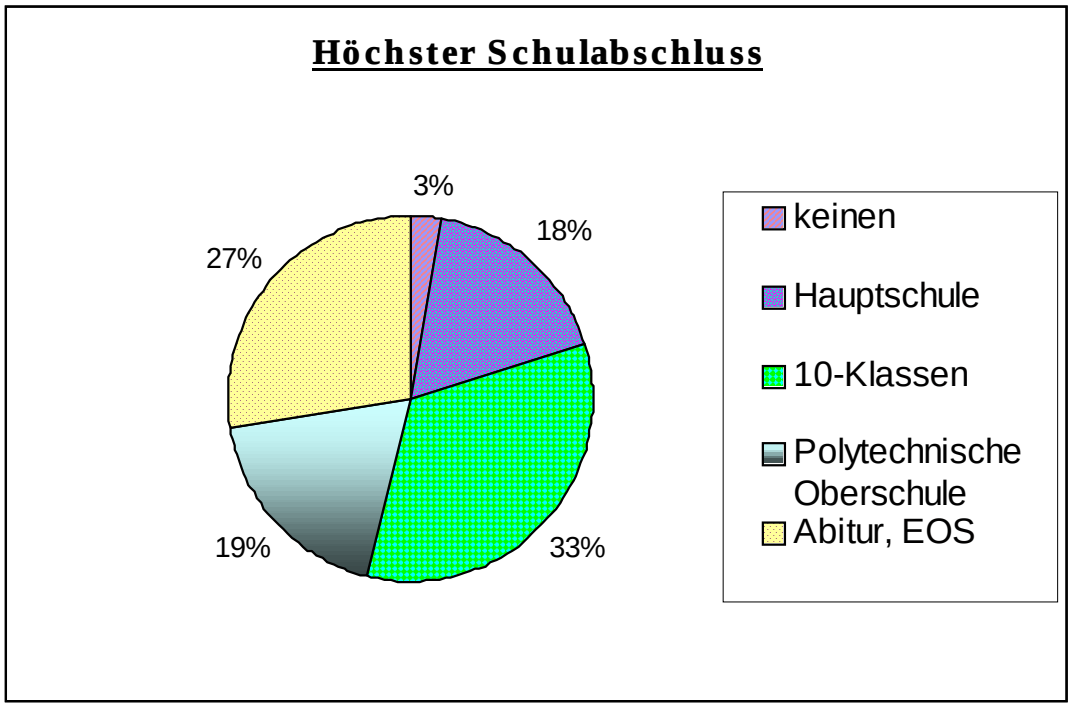
Das Durchschnittsalter betrug 40 Jahre (± 13,4 Jahre) mit einem Minimum von 18 und einem maximalen Alter von 79 Jahren.



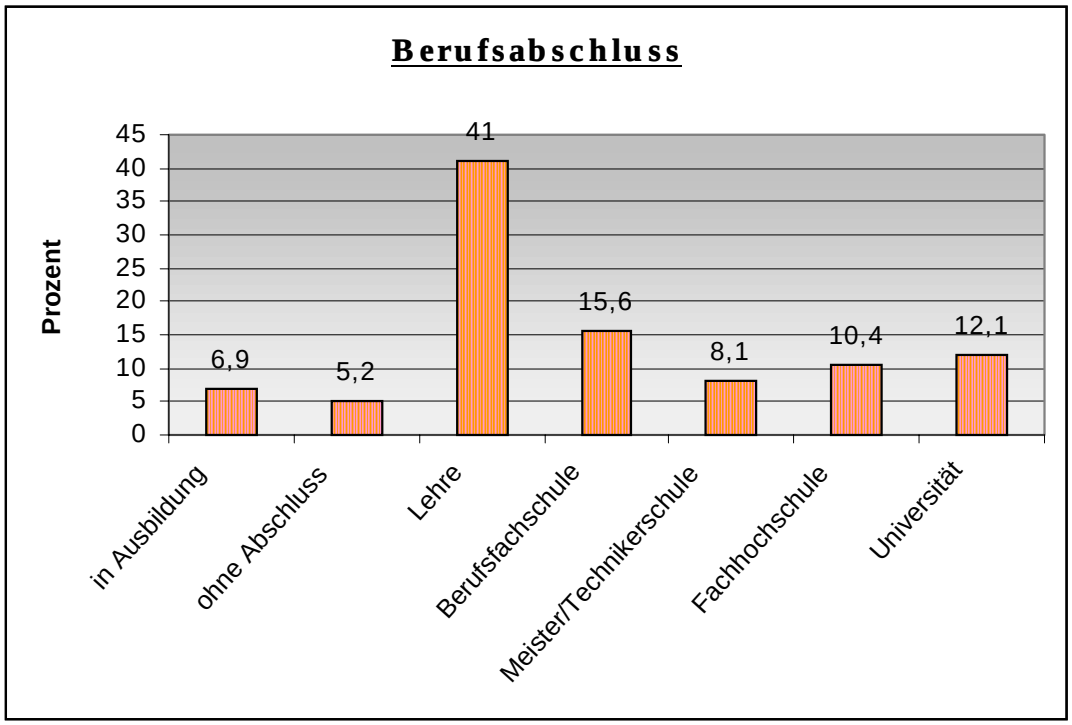
48 % waren verheiratet, 36 % ledig, 16 % geschieden oder verwitwet (siehe Grafik).

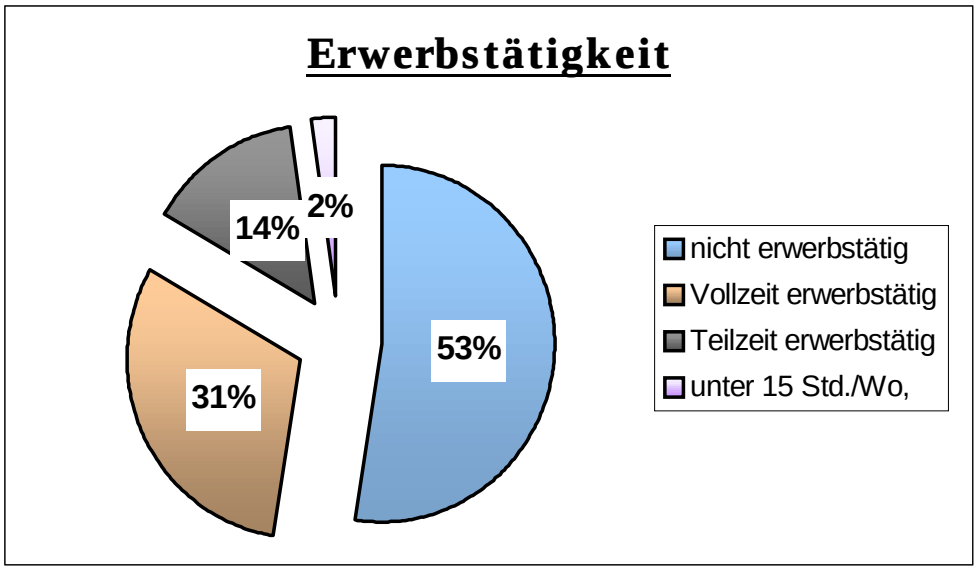
Das monatliche Nettoeinkommen des Haushalts betrug im Mittel 3000 DM.

33 % der Patienten verfügte über einen 10-Klassen-Schulabschluss, 27 % hatte die erweiterte Oberschule (EOS) besucht mit Abitur, 19 % die polytechnische Oberschule, 18 % die Hauptschule, und 3 % hatten keinen Schulabschluss (Grafik).



Die meisten Patienten (41 %) hatten eine Lehre absolviert, 16 % die Berufsfachhochschule, 12 % waren Akademiker, 10 % Fachhochschulabsolventen und 5 % blieben ohne Berufsabschluss (Grafik).





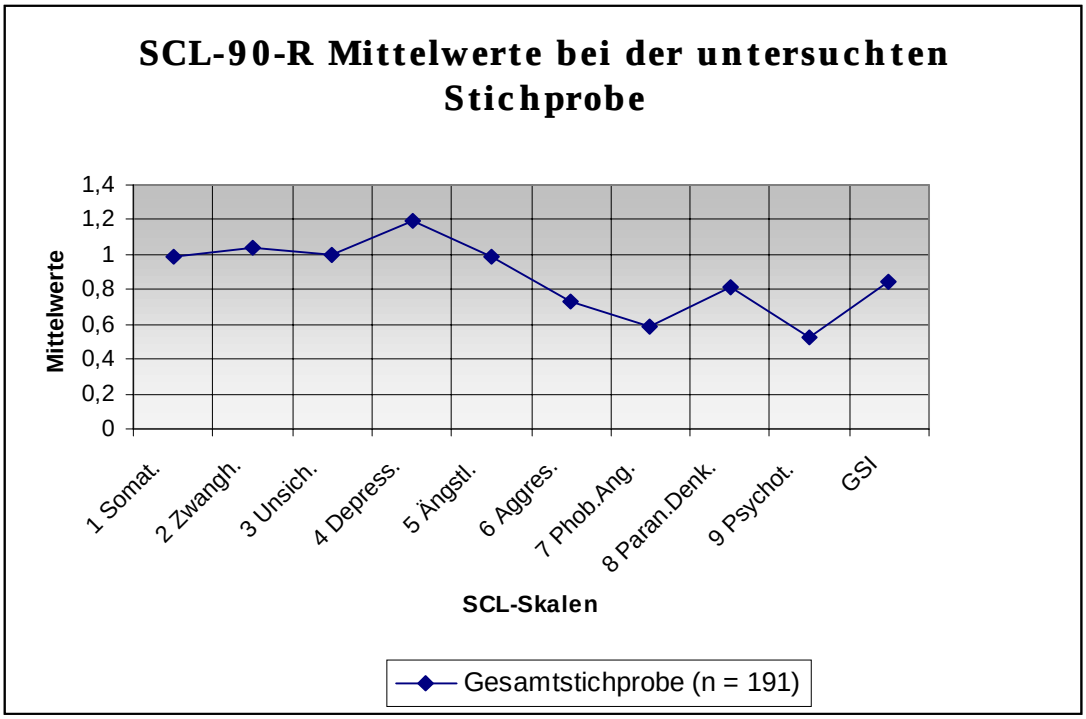
Von den Patienten war über die Hälfte (53 %) nicht erwerbstätig, 31 % standen in einem Vollzeit-Arbeitsverhältnis, 16 % arbeiteten in einer Teilzeitstelle (Grafik: Erwerbstätigkeit).
Von den zur Zeit nicht erwerbstätigen Patienten sind 46 % schon über 1¼ Jahr arbeitsunfähig geschrieben bzw. in Rente, 36,4 % sind arbeitslos und 10 % noch in Ausbildung.

Im Durchschnitt dauerten die F-DIPS-Interviews 97 Minuten (± 41,4 Min.), wobei die in der Rehabilitationsklinik geführten Interviews (83 ± 34 Minuten) im Vergleich zu den Akutkliniken (109 ± 43,2 Minuten) signifikant schneller durchgeführt wurden.

8.2 Fragebogenergebnisse der Gesamtstichprobe

Bei der untersuchten Stichprobe liegt der BDI-Gesamtwert bei einem Durchschnitt von 16,5 mit einer Standardabweichung von 11 (Minimum = 0, Maximum = 43). Der BAI-Mittelwert ist 15 mit einer Standardabweichung von 10,9 (Minimum = 0, Maximum = 46).
Der Mittelwert im Whiteley-Fragebogen liegt bei 4,84 mit einer Standardabweichung von 3,6 (Minimum = 0, Maximum = 14).

Die SCL-90-R-Mittelwerte können der Grafik entnommen werden.



Die Standardabweichungen für die einzelnen Skalen sind für die Skala Somatisierung: 0,77 (Max.: 3,6), Zwanghaftigkeit: 0,77 (Max.: 3,6), Unsicherheit im Sozialkontakt: 0,86 (Max.:3,3), Depressivität: 0,86 (Max.: 3,3), Ängstlichkeit: 0,79 (Max.: 2,8), Aggressivität / Feindseligkeit: 0,7 (Max.: 3,8), Phobische Angst: 0,8 (Max.: 4), Paranoides Denken: 0,79 (Max.: 3,3), Psychotizismus: 0,5 (Max.: 2,3), Globaler Kennwert GSI: 0,57 (Max.: 2,43). Die Minima liegen bei allen Skalen bei 0.

8.3 Ergebnisse aus den F-DIPS-Interviews

8.3.1 Bestimmung der Reliabilität des F-DIPS (Beantwortung der 1. Frage: Wie hoch ist die Übereinstimmung der F-DIPS-Diagnosen von zwei im Abstand von 2 Wochen (± 1 Woche) durchgeführten Interviews durch unabhängige Rater am selben Patienten?)

Die Daten in der folgenden Tabelle zeigen zum einen die Häufigkeitsverteilung an bei der Vergabe einer Diagnose durch beide Interviewer (Feld a) oder durch einen der beiden Interviewer und durch den anderen nicht (Feld b und c). Feld d gibt an, wenn beide Interviewer keine Störung diagnostizierten.

		Interview 1 +	Interview 1 -
Interview 2	+	a	b
	-	c	d

Zum anderen sind in der Tabelle die Basisraten enthalten. Die Basisrate errechnet sich aus den vergebenen Diagnosen in Relation auf die insgesamt geführten Interviews, d.h. n_{ges} = 382. Für die 4-Felder Matrix bedeutet dies, dass die Basisrate folgendermaßen errechnet wird: (a + a + b + c) : n_{ges} · 100 = Basisrate. Beispielsweise liegt die Basisrate für die Panikstörung ohne Agoraphobie bei 7 + 7 + 5 + 8 = 27. 27 Mal eine vergebene Diagnose bei 382 Interviews sind 7,1 %.

Bei Basisraten unter 10 % unterschätzt der Kappa-Wert die Übereinstimmung. Aus diesem Grund wird zusätzlich der Y- Koeffizient bestimmt. In der folgenden Tabelle ist zur besseren Übersicht bei Basisraten unter 10% der Y-Wert fett markiert, bei Basisraten über 10% der Kappa-Wert.

Des weiteren werden in der Tabelle die prozentualen Übereinstimmungen angegeben. Diese liegen aufgrund der hohen Fallzahlen in Feld d relativ hoch.

Diagnose	Häufigkeit +/+	+/-	-/+	-/-	Basisrate	Prozentuale Übereinstimmung	κ	Y
Panikstörung ohne Agoraphobie	7	5	8	171	7,1	93 %	.46	.69
Panikstörung mit Agoraphobie	19	5	6	161	12,8	94 %	.73	.82
Agoraphobie ohne Panikstörung	1	1	4	185	1,8	97 %	.25	.74
Sozialphobie	24	14	14	139	19,9	85 %	.53	.61
Spezifische Phobie	10	10	4	167	8,9	93 %	.56	.73
Generalisierte Angststörung	4	7	6	174	5,5	93 %	.36	.61
Zwangsstörung	4	2	1	184	2,9	98 %	.67	.90
PTSD	2	5	2	182	2,9	96 %	.33	.72
Dysthyme Störung	4	6	12	169	6,8	91 %	.31	.51
Major Depression, einzelne Episode	35	13	17	126	26,2	84 %	.65	.63
Major Depression, rezidivierend	45	9	7	130	27,7	92 %	.80	.81
Major Depression, gesamt	89	12	14	76	53,4	86 %	.72	.73
Bipolar I,II, Zyklothyme Störung	5	1	0	185	2,9	99 %	.83	(1)
Angst-Depressions-Störung	0	2	0	185	0,5	97 %	.40	-
Hypochondrie	3	0	1	187	1,8	99 %	.75	(1)
Somatisierungsstörung	2	3	2	184	2,4	97 %	.40	.77
Konversionsstörung	2	2	5	182	2,9	96 %	.30	.72
Somatoforme Schmerzstörung	23	13	10	145	18,1	88 %	.59	.67
Alkoholmissbrauch und –abhängigkeit	3	4	1	183	2,9	97 %	.40	.84
Medikamentenmissbrauch und –abhängigkeit	2	2	3	184	2,4	97 %	.40	.77
Drogenmissbrauch und –abhängigkeit	2	1	0	188	1,3	99 %	.67	(1)
Anorexia nervosa	7	1	3	180	4,7	98 %	.78	.91
Bulimia nervosa	9	1	1	180	5,2	99 %	.90	.95

Ausgezeichnete Übereinstimmungen (κ > .75 oder Y > .80) ergaben sich demzufolge bei der Klassifikation der folgenden Störungen:

- *Zwangsstörung*
- *Major Depression, rezidivierend*
- *Bipolar I, II-Störung, zyklothyme Störung*
- *Hypochondrie*
- *Alkoholmissbrauch und -abhängigkeit*
- *Drogenmissbrauch und –abhängigkeit*
- *Anorexia nervosa*
- *Bulimia nervosa*

Gute Übereinstimmungen (κ zwischen .65 - .75 oder Y < .80 und > .70) zeigten sich bzgl. der Störungsbilder:

- *Panikstörung mit Agoraphobie*
- *Agoraphobie ohne Panikstörung*
- *Spezifische Phobie*
- *Posttraumatische Belastungsstörung*
- *Major Depression, einzelne Episode*
- *Somatisierungsstörung*
- *Konversionsstörung*
- *Medikamentenmissbrauch und –abhängigkeit*

Mäßige bis mittlere Übereinstimmungen (κ zwischen .40 und .65 oder Y <.70 und >.65) waren zu finden bei:

- *Panikstörung ohne Agoraphobie*
- *Sozialphobie*
- *Somatoforme Schmerzstörung*

Ungenügende Übereinstimmungen (κ < .40 oder Y < .65) fanden sich bei den Diagnosen:

- *Generalisierte Angststörung*
- *Dysthyme Störung*
- *Angst-Depressions-Störung*

In der folgenden Tabelle sind die Übereinstimmungen für die Störungsoberklassen aufgeführt.

Störungsoberklasse	Häufigkeit +/+ +/- -/+ -/-	Basisrate	Prozentuale Übereinstimmung	κ	Y
Angststörungen	70 13 21 87	45,5	82 %	.64	.65
Affektive Störungen	106 11 15 59	62,3	86 %	.71	.72
Somatoforme Störungen	33 10 12 136	23,0	88 %	.66	.72
Substanzmissbrauch und –abhängigkeit	7 4 2 178	5,2	97 %	.70	.85
Essstörungen	17 1 3 170	9,9	98 %	.89	.94
keine Achse I - Störung	17 5 10 159	12,8	92 %	.65	.76

Bei den verschiedenen Störungsgruppen ergaben sich ausgezeichnete (κ > .75 oder Y > .80) Übereinstimmungen in Bezug auf:

- *Substanzmissbrauch und –abhängigkeit und*
- *Essstörungen.*

Gute Übereinstimmungen (κ > .65) ergaben sich bei:

- *Affektiven Störungen und*
- *Somatoformen Störungen.*

Eine mäßige bis mittlere Übereinstimmung (κ = .64) zeigte sich bei den

- *Angststörungen.*

Auch die Übereinstimmung im Urteil, dass keine Achse I-Störung vorliegt, liegt im mittleren Bereich.

Fasst man die Angststörungen bzgl. der Trait- (Sozialphobie, Generalisierte Angststörung) und der State-Ängste (Agoraphobie, Panikstörung mit und ohne Agoraphobie, Spezifische Phobie) zusammen und misst hierfür die Retestreliabilität, ergaben sich folgende Werte:

Angst-Störungsgruppe	Häufigkeit +/+ +/- -/+ -/-	Basisrate	Prozentuale Übereinstimmung	κ	Y
State-Ängste	40 16 13 122	28,5	85 %	.63	.66
Trait-Ängste	28 17 16 130	23,3	83 %	.52	.57

Sowohl die State-Ängste als auch die Trait-Ängste lassen sich mit mittlerer Reliabilität erfassen, State-Ängste im Vergleich etwas besser.

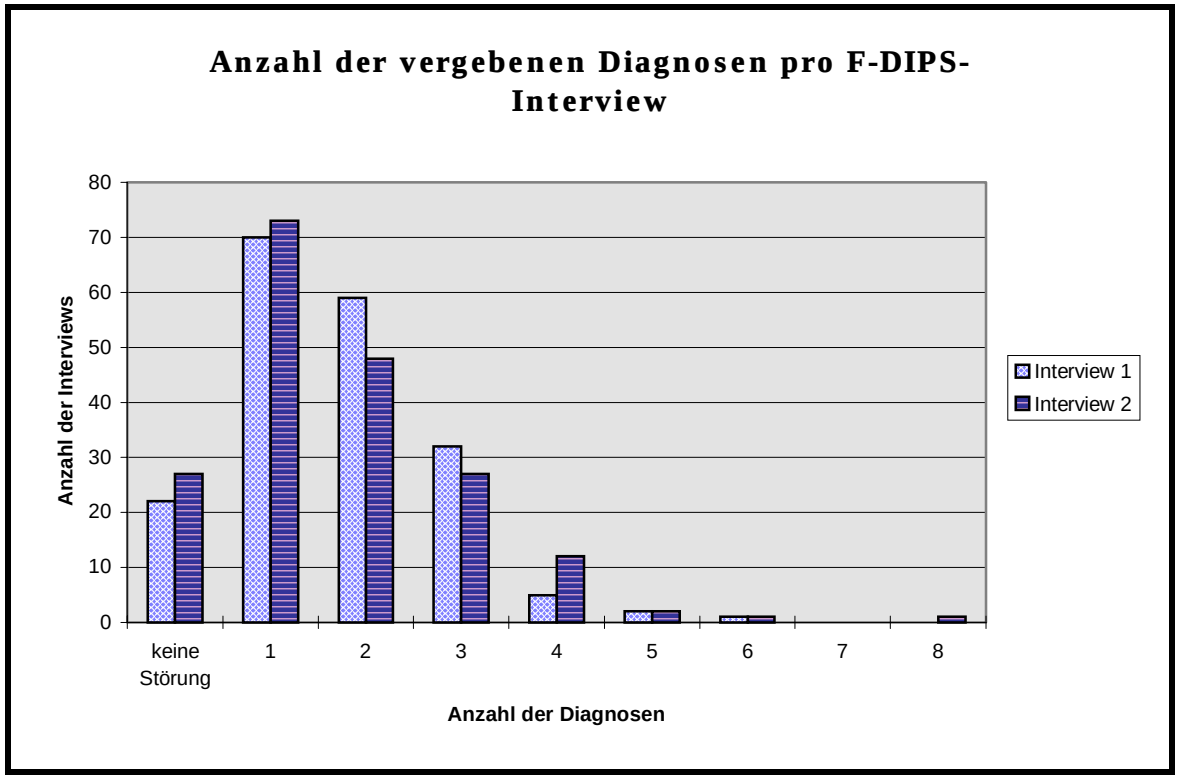
Eine andere Gruppierung könnte in *Panikstörung* (Panikstörung mit und ohne Agoraphobie), *Agoraphobie* (Panikstörung mit Agoraphobie, Agoraphobie ohne Panikstörung), *Phobische Störungen* (Agoraphobie, Sozialphobie, Spezifische Phobie, Panikstörung mit Agoraphobie) erfolgen.
Hierbei ergibt sich folgende Reliabilität:

Angst-Störungsgruppe	Häufigkeit +/+ +/- -/+ -/-	Basisrate	Prozentuale Übereinstimmung	κ	Y
Panikstörung	29 11 8 143	20,2	90 %	.69	.75
Agoraphobie	22 5 8 156	14,9	93 %	.73	.81
Phobische Störungen	52 20 15 104	36,4	82 %	.60	.62

Die Retest-Reliabilität des F-DIPS kann bei 16 Störungen als gut bis ausgezeichnet bezeichnet werden. Bei 3 Störungen fand sich eine mäßige bis mittlere und bei 3 Störungen eine ungenügende Übereinstimmung. Auf die Störungsoberklassen bezogen, gab es gute bis ausgezeichnete Übereinstimmung zwischen den beiden Interviewern mit Ausnahme der Angststörungen, wo die Übereinstimmung nur eine mittlere war.

8.3.2 Übereinstimmung in komorbiden Störungen (Beantwortung der 2. Frage: Inwieweit stimmen die Interviewer auch in der Vergabe komorbider Störungen überein?)

Durchschnittlich wurden sowohl im ersten als auch im zweiten Interview 1,7 Diagnosen vergeben. Es ergab sich ein hoch signifikanter Zusammenhang zwischen der Anzahl der Diagnosen im ersten und der im zweiten Interview (t-Test für gepaarte Stichproben) und keine signifikanten Unterschiede bzgl. der Diagnosenanzahl.



Im ersten Interview wurde 22 Mal (11,5 %) keine Diagnose gestellt, im zweiten 27 Mal (14,1 %). Ausschließlich eine Diagnose wurde in 70 ersten Interviews (36,6 %) erfasst und in 73 zweiten Interviews (38,2 %). Zwei Störungen wurden in 59 Fällen im ersten Interview diagnostiziert (30,9 %) und in 48 Fällen (25,1 %) im zweiten. 32 Mal (16,8 %) wurden 3 Störungen im ersten Interview gefunden und 27 Mal (14,1 %) im zweiten. Vier Störungen wurden diagnostiziert in 5 Fällen (2,6 %) beim ersten In-

terview und in 12 Fällen (6,3 %) im zweiten Interview. Mehr als 4 Störungen wurden im ersten Interview 3 Mal (1,5 %), im zweiten Interview 4 Mal (2 %) vergeben.
Bei 116 (60,7 %) von 191 Patienten stimmte die diagnostische Einschätzung der beiden Interviewer in allen Diagnosen überein.

Zu Frage 2, inwieweit die Interviewer auch in der Vergabe komorbider Störungen übereinstimmten, lässt sich sagen, dass sie bzgl. der Anzahl der Diagnosen sehr gut übereinstimmten. Insgesamt wurden 61 % der Patienten auch hinsichtlich ihrer komorbiden Störungen von den Interviewern gleich bewertet.

8.3.3 Die Rolle der Erfahrung auf die Reliabilität der Diagnosen (Beantwortung der 3. Frage: Spielen die klinisch-psychiatrische Erfahrung oder die Erfahrung in der Anwendung des F-DIPS eine Rolle für die Reliabilität der Diagnosen?)

Hierzu wurden die Interviewer je nach Erfahrung in verschiedene Gruppen (1,2,3) eingeteilt (vgl. auch Kap. 7.4) und die Übereinstimmung miteinander verglichen. Zu Gruppe 1 (klinisch erfahrene Interviewer) gehörten zwei Interviewer mit klinisch-psychiatrischer Erfahrung, ohne Erfahrung mit dem F-DIPS, zu Gruppe 2 (F-DIPS-erfahrene Interviewer) vier Interviewer, die bereits aus der Studie „Gesundheit junger Frauen“ viel Erfahrung in der Anwendung des F-DIPS, jedoch keine psychiatrische Erfahrung mitbrachten, und in Gruppe 3 (unerfahrene Interviewer) sind sieben Interviewer zusammengefasst, die weder Vorerfahrung im F-DIPS noch klinisch-psychiatrische Erfahrung aufwiesen.

Die folgende Tabelle zeigt auf, wie viele Interviews von den einzelnen Interviewer-Paarungen durchgeführt wurden:

	Gruppe 1	Gruppe 2	Gruppe 3	SUMME
Gruppe 1	11 *	22	9	53
Gruppe 2	22	31 *	22	106
Gruppe 3	9	22	96 *	223
Summe	53	106	223	382

* Die Anzahl der Interviews aus diesen Feldern verdoppelt sich, da von jedem Interviewer ein Interview geführt wurde.

Um zu erfassen, inwieweit sich klinische Erfahrung bzw. F-DIPS-Erfahrung auf die Reliabilität auswirken, wurde die Gruppe 3 einmal mit Gruppe 1 zur Gruppe der Interviewer ohne F-DIPS-Erfahrung verbunden und dann mit der Gruppe 2 zur gemeinsamen Gruppe der Interviewer ohne klinische Erfahrung zusammengefasst. Da die Basisraten in den einzelnen Gruppen sehr verschieden sind und sich dies auf die Kappa-Werte auswirkt, werden, falls der Kappa-Wert niedrig ist und eine geringe Basisrate vorhanden ist, auch zur genaueren Information die Yule-Koeffizienten als Ergänzung zu den Kappa-Werten in den folgenden Tabellen angegeben.

Es ergaben sich folgende Resultate:

Vergleich klinisch erfahrene vs. unerfahrene Interviewer

a) in Bezug auf Angststörungen

	Interviewer ohne klinische Erfahrung			Interviewer mit klinischer Erfahrung		
	Basisrate	κ	Y	Basisrate	κ	Y
Ohne klinische Erfahrung	43,6	.64				
Mit klinischer Erfahrung	66,1	.50		31,8	.79	

Die gepaarten Interviewer mit klinischer Erfahrung konnten die Angststörungen ausgezeichnet reliabel erfassen.
Die Übereinstimmung in der Störungsoberklasse Angststörungen war am geringsten zwischen klinisch erfahrenen und klinisch unerfahrenen Interviewern.

b) in Bezug auf affektive Störungen

	Interviewer ohne klinische Erfahrung			Interviewer mit klinischer Erfahrung		
	Basisrate	κ	Y	Basisrate	κ	Y
Ohne klinische Erfahrung	59,4	.68				
Mit klinischer Erfahrung	69,4	.77		81,8	1.0	

Affektive Störungen wurden mit einer ausgezeichneten Reliabilität von der Paarung Interviewer mit klinischer Erfahrung, aber auch zwischen Interviewern mit klinischer Erfahrung und Interviewern ohne klinische Erfahrung diagnostiziert.

c) in Bezug auf somatoforme Störungen

	Interviewer ohne klinische Erfahrung			Interviewer mit klinischer Erfahrung		
	Basisrate	κ	Y	Basisrate	κ	Y
Ohne klinische Erfahrung	24,5	. 69				
Mit klinischer Erfahrung	24,2	.56		0		

Hinsichtlich der somatoformen Störungen gab es eine etwas größere Übereinstimmung zwischen den Interviewern ohne klinische Erfahrung als

zwischen den Interviewern mit klinischer Erfahrung vs. ohne klinische Erfahrung.

d) in Bezug auf Substanzmissbrauch und -abhängigkeit

	Interviewer ohne klinische Erfahrung			Interviewer mit klinischer Erfahrung		
	Basisrate	κ	Y	Basisrate	κ	Y
Ohne klinische Erfahrung	5,0	.65	.83			
Mit klinischer Erfahrung	8,1	.78	(1)	0		

Bei den Störungen zum Substanzmissbrauch und –abhängigkeit war die Übereinstimmung zwischen Interviewern mit klinischer Erfahrung und solchen ohne klinische Erfahrung ausgezeichnet, aber auch die unter den Interviewern ohne klinische Erfahrung gut bis ausgezeichnet.

e) in Bezug auf Essstörungen

	Interviewer ohne klinische Erfahrung			Interviewer mit klinischer Erfahrung		
	Basisrate	κ	Y	Basisrate	κ	Y
Ohne klinische Erfahrung	9,7	.89				
Mit klinischer Erfahrung	14,5	.87		0		

Essstörungen konnten sowohl von Interviewern mit als auch ohne klinische Erfahrung mit ausgezeichneter Übereinstimmung klassifiziert werden.

Anwendungserfahrung im F-DIPS

a) in Bezug auf Angststörungen

	Interviewer ohne F-Dips-Erfahrung			Interviewer mit F-Dips-Erfahrung		
	Basis-rate	κ	Y	Basis-rate	κ	Y
Ohne F-Dips-Erfahrung	46,6	.72				
Mit F-Dips-Erfahrung	47,7	.64		38,7	.33	

Die Angststörungen konnten von Interviewern, die das F-DIPS noch nicht innerhalb der Studie „Gesundheit junger Frauen“ vielfach angewendet hatten, mit guter Übereinstimmung klassifiziert werden. Die erfahrenen F-DIPS-Interviewer erzielten untereinander nur eine ungenügende Übereinstimmung.

b) in Bezug auf affektive Störungen

	Interviewer ohne F-Dips-Erfahrung			Interviewer mit F-Dips-Erfahrung		
	Basis-rate	κ	Y	Basis-rate	κ	Y
Ohne F-Dips-Erfahrung	63,4	.76				
Mit F-Dips-Erfahrung	63,6	.71		56,5	.55	

Auch in Bezug auf affektive Störungen zeigten sich bessere Übereinstimmungen bei den Interviewern ohne F-DIPS-Erfahrung.

c) in Bezug auf somatoforme Störungen

	Interviewer ohne F-Dips-Erfahrung			Interviewer mit F-Dips-Erfahrung		
	Basis-rate	κ	Y	Basis-rate	κ	Y
Ohne F-Dips-Erfahrung	21,1	.77				
Mit F-Dips-Erfahrung	27,3	.66		24,2	.39	

Die Interviewer mit F-DIPS-Erfahrung diagnostizierten mit deutlich geringerer Übereinstimmung eine somatoforme Störung als die Interviewer ohne F-DIPS-Erfahrung.

d) in Bezug auf Substanzmissbrauch und -abhängigkeit

	Interviewer ohne F-Dips-Erfahrung			Interviewer mit F-Dips-Erfahrung		
	Basis-rate	κ	Y	Basis-rate	κ	Y
Ohne F-Dips-Erfahrung	6,0	.54				
Mit F-Dips-Erfahrung	5,7	.79		3,2	1.0	

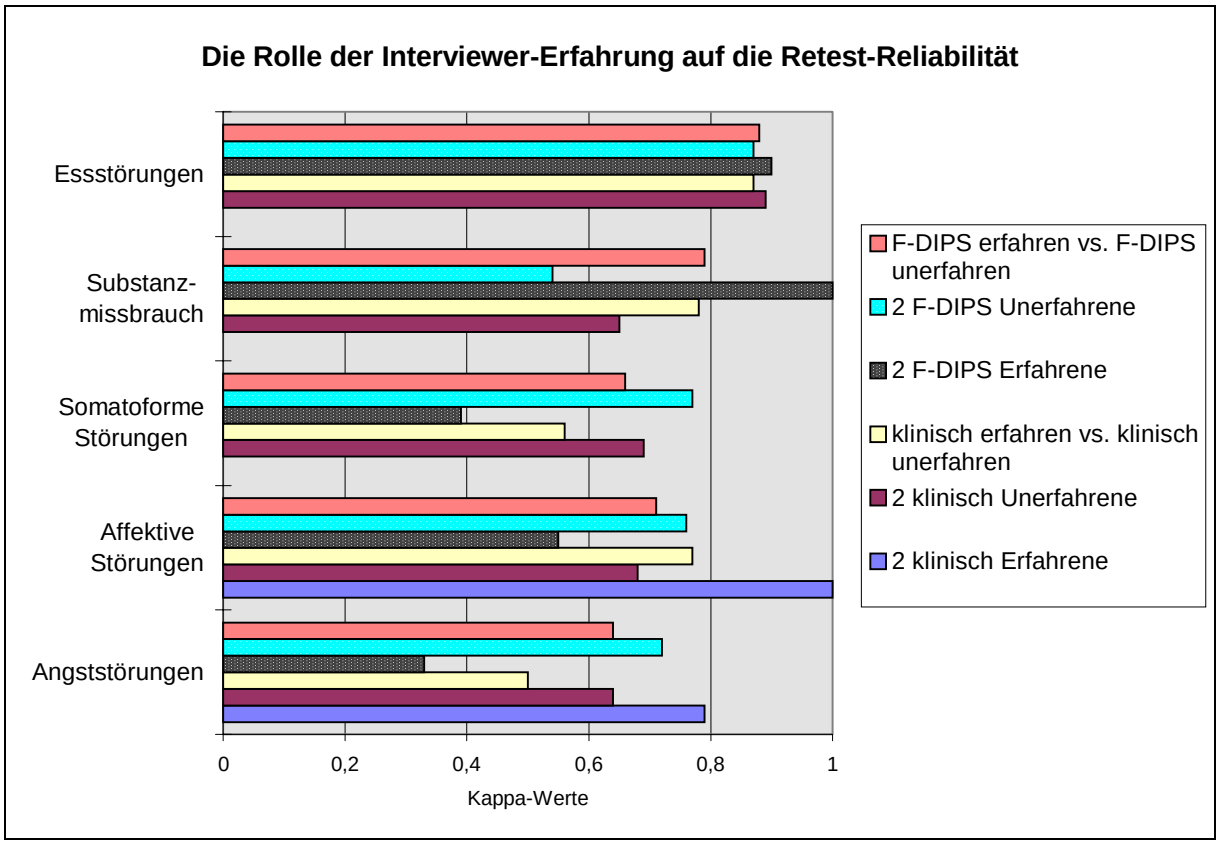
Hinsichtlich des Substanzmissbrauchs und –abhängigkeit sind die Basisraten in allen Gruppen niedrig, deshalb die Kappa-Werte miteinander vergleichbar, weshalb auf das Angeben der Yules-Koeffizienten verzichtet wurde. Interviewer mit F-DIPS-Erfahrung konnten die Störungsgruppe mit ausgezeichneter Reliabilität erfassen. Interviewer ohne F-DIPS-Erfahrung erreichten nur eine mäßige Übereinstimmung.

e) in Bezug auf Essstörungen

	Interviewer ohne F-Dips-Erfahrung			Interviewer mit F-Dips-Erfahrung		
	Basis-rate	κ	Y	Basis-rate	κ	Y
Ohne F-Dips-Erfahrung	6,9	.87				
Mit F-Dips-Erfahrung	10,2	.88		21,0	.90	

Essstörungen konnten von Interviewern ohne und mit F-DIPS-Erfahrung gleich reliabel diagnostiziert werden.

Zusammenfassend lassen sich die Ergebnisse in der folgenden Grafik darstellen.



Wird außerdem die Übereinstimmung der Rater für alle außer den ersten 5 Interviews überprüft (n=148), ergeben sich (siehe Tabelle) quasi keine Veränderungen.

Störungsoberklasse	κ (unter Einbeziehung aller Interviews, s.o., n=191)	κ (Auswahl der „erfahreneren Interviews“, nach den ersten 5, n=148)
Angststörungen	.64	.60
Affektive Störungen	.71	.70
Somatoforme Störungen	.66	.63
Substanzmissbrauch und –abhängigkeit	.70	.73
Essstörungen	.89	.89

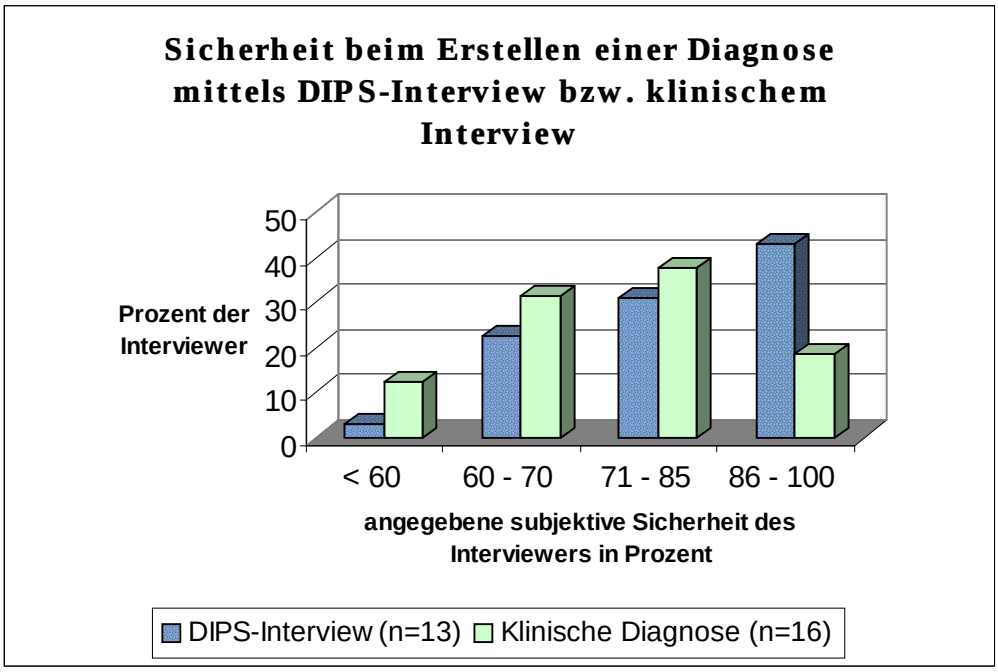
Insgesamt kann die Frage 3, ob klinisch-psychiatrische Erfahrung bzw. Erfahrung in der Anwendung des F-DIPS einen Einfluss auf die Retest-Reliabilität haben, folgendermaßen beantwortet werden:

- Aufgrund der kleinen Stichprobe (n=11) der gepaarten Interviewer mit klinischer Erfahrung lässt sich keine Aussage zum Einfluss der klinischen Erfahrung machen. Es dürfte jedoch keine größeren Unterschiede geben, da es keine deutlichen oder gleichbleibenden Tendenzen im Vergleich der Gruppierung „ohne kl. Erfahrung / ohne kl. Erfahrung“ mit der Gruppierung „ohne kl. Erfahrung / mit kl. Erfahrung“ abzeichnen.
- Dagegen scheint die Erfahrung in der Anwendung des F-DIPS einen negativen Effekt auf die Reliabilität zu zeigen. Lediglich Essstörungen und Substanzmissbrauch und –abhängigkeit konnten die erfahrenen Rater übereinstimmend diagnostizieren.

Dieses eher überraschende Ergebnis könnte evtl. damit zusammenhängen, dass der Effekt nicht durch die Erfahrung zustande kommt, sondern durch die schon länger zurückliegende Schulung der Interviewer

bzw. damit, dass die Rater aus der Studie ihre Erfahrung an meist Gesunden (Allgemeinbevölkerung) sammelten und deshalb in der Einschätzung klinischer Störungsbilder mit multiplen Auffälligkeiten Schwierigkeiten hatten.

8.3.4 Die Rolle des Gefühls der Sicherheit auf die Reliabilität (Beantwortung der 4. Frage: Gibt es eine höhere Übereinstimmung der Diagnosen, wenn eine hohe Sicherheit bei der Vergabe der Diagnose vorlag?)



Es wurden sowohl die Interviewer, die ihre Diagnose mit Hilfe des F-DIPS als auch Psychologen und Psychiater, die Diagnosen nach einem klinischen Interview formulierten, befragt, wie sicher sie sich mit ihren Diagnosen gefühlt haben. Dabei wurde nach jedem F-DIPS-Interview eine Schätzung der subjektiven Sicherheit vermerkt. Die Sicherheit für die klinischen Diagnosen wurde retrospektiv über alle im letzten Jahr vergebenen Diagnosen erfragt.

Es ergab sich eine durchschnittliche diagnostische Sicherheit von 82 % bei den F-DIPS Interviews. Dabei fühlten Interviewer lediglich in 3 % aller Interviews eine unter 60 %ige Sicherheit, 23 % fühlten sich zwischen 60 – 70 % sicher, 31 % zwischen 71 – 85 %, und die meisten (43 %) fühlten sich über 86 % sicher in ihrer diagnostischen Einschätzung.
Die klinischen Diagnosen der Behandler der Einrichtungen, in denen die F-DIPS-Interviews durchgeführt wurden, werden mit deutlich geringerer Sicherheit vergeben. Hier fühlen 13 % unter 60 % Sicherheit beim Vergeben einer Diagnose, 31 % zwischen 60 – 70 %ige Sicherheit, 38 % 71 – 85 %ige und nur 19 % über 86 %ige Sicherheit.

Berechnet man die Kappa-Koeffizienten in Abhängigkeit von der Sicherheit bei der Vergabe der F-DIPS-Diagnose, ergibt sich folgendes:

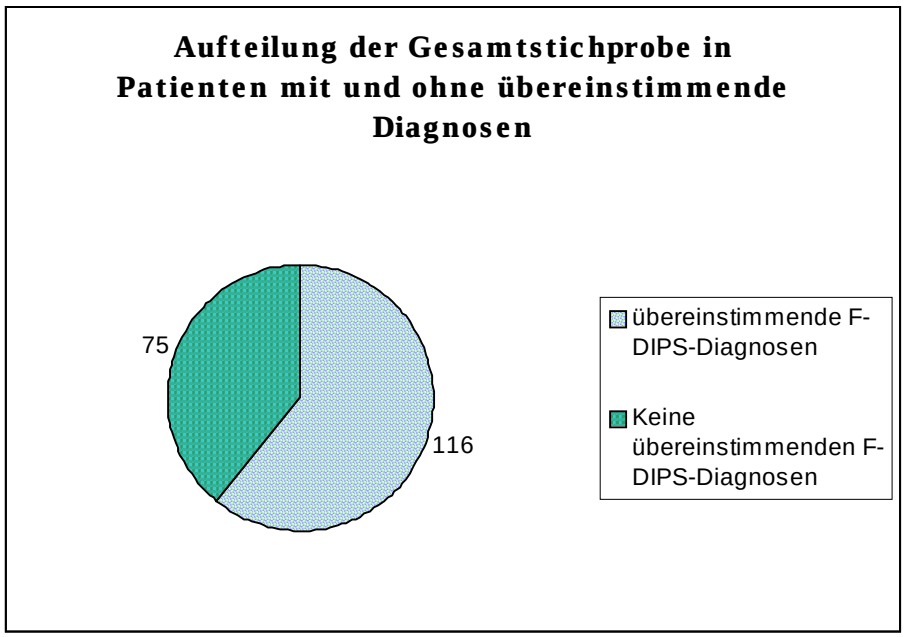
	Sicherheit		
Störungsoberklasse	**≤ 70 %** (n=44) κ	71-85 % (n=67) κ	**≥ 86 %** (n=76) κ
Angststörungen	.64	.64	.65
Affektive Störungen	.71	.69	.70
Somatoforme Störungen	.64	.77	.59
Substanzmissbrauch und -abhängigkeit	.79	.85	.41
Essstörungen	1.0	.70	.95

Unabhängig davon, wie die eigene Sicherheit beim Vergeben der Diagnose ist, unterscheiden sich die Kappa-Werte nicht in einer systematischen Weise voneinander, so dass von keinem Effekt der Diagnosesicherheit auf die Reliabilität, mit der eine Störung gemessen wird, ausgegangen werden kann. Die beiden schlechtesten Reliabilitätswerte sind im Gegenteil die von den Interviewern, die sich am sichersten mit ihrer Diagnose gefühlt hatten (bei Substanzmissbrauch und –abhängigkeit sowie somatoformer Störung).

Frage 4, ob es höhere Übereinstimmungen der Diagnosen gibt, wenn sich die Interviewer sicher mit ihrer Diagnose fühlten, kann verneint werden. Ein zu sicheres Gefühl sprach sogar bei der Substanzabhängigkeit gegen eine gute Reliabilität. Behandler fühlen sich im Vergleich mit F-DIPS-Interviewern deutlich seltener sehr sicher in ihrer Diagnose.

8.3.5 Konfundierende Variablen (Beantwortung der 5. Frage: Existieren konfundierende Variablen, die die Genauigkeit der Messung mit dem F-DIPS beeinflussen?)

Für die Beantwortung dieser Frage wurde die Gesamtstichprobe in die Gruppe derjenigen, die in den beiden Interviews voneinander abweichende F-DIPS- Diagnosen erhielten (Nicht- Übereinstimmungs-Stichprobe) und in die Gruppe der Patienten, die eine übereinstimmende F-DIPS-Diagnose erhielten, unterteilt und miteinander mit Hilfe des Mann-Whitney-U-Tests für zwei unabhängige Stichproben bzw. bei Intervallskalenniveau mit einem t-Test für zwei unabhängige Stichproben verglichen.



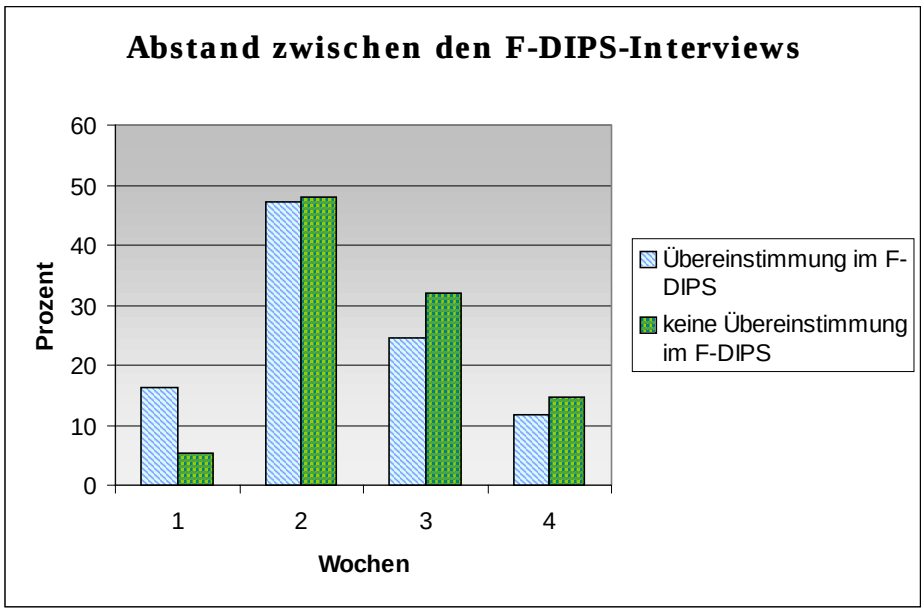
8.3.5.1 Abstand der beiden F-DIPS-Interviews voneinander

Im Durchschnitt sollte das F-DIPS in einem Abstand von 2 Wochen zweimal durchgeführt werden. Aufgrund von Zeit- und Koordinierungsproblemen bei Patient oder Interviewer gab es jedoch Abweichungen von den angestrebten 2 Wochen. Ob diese Abweichungen einen Einfluss auf das Ergebnis haben, sollte mit dem Vergleich der beiden Teilstichproben untersucht werden.

Der Abstand zwischen den Interviews war (knapp) nicht signifikant verschieden (Mann-Whitney-U-Test: p= 0,06) in der Gruppe der Interviews mit übereinstimmender Diagnose (x = 2,32; s= 0,89) und der ohne übereinstimmende Diagnose (x=2,56; s=0,81).

Die Tendenz spricht jedoch für einen Einfluss des Abstandes zwischen den Interviews auf die Übereinstimmung in der Diagnose.

Dabei ist besonders die kleine Gruppe, bei der das Interview bereits in



einer Woche wiederholt wurde, interessant, da es hier verhältnismäßig zu einem höheren Anteil an übereinstimmenden Diagnosen kam als bei den Patienten, bei denen erst später das zweite Interview durchgeführt wurde. Bei einem Interview-Abstand von 3 oder 4 Wochen ging die Tendenz dagegen eher dahin, dass es mehr Nichtübereinstimmungen gibt.

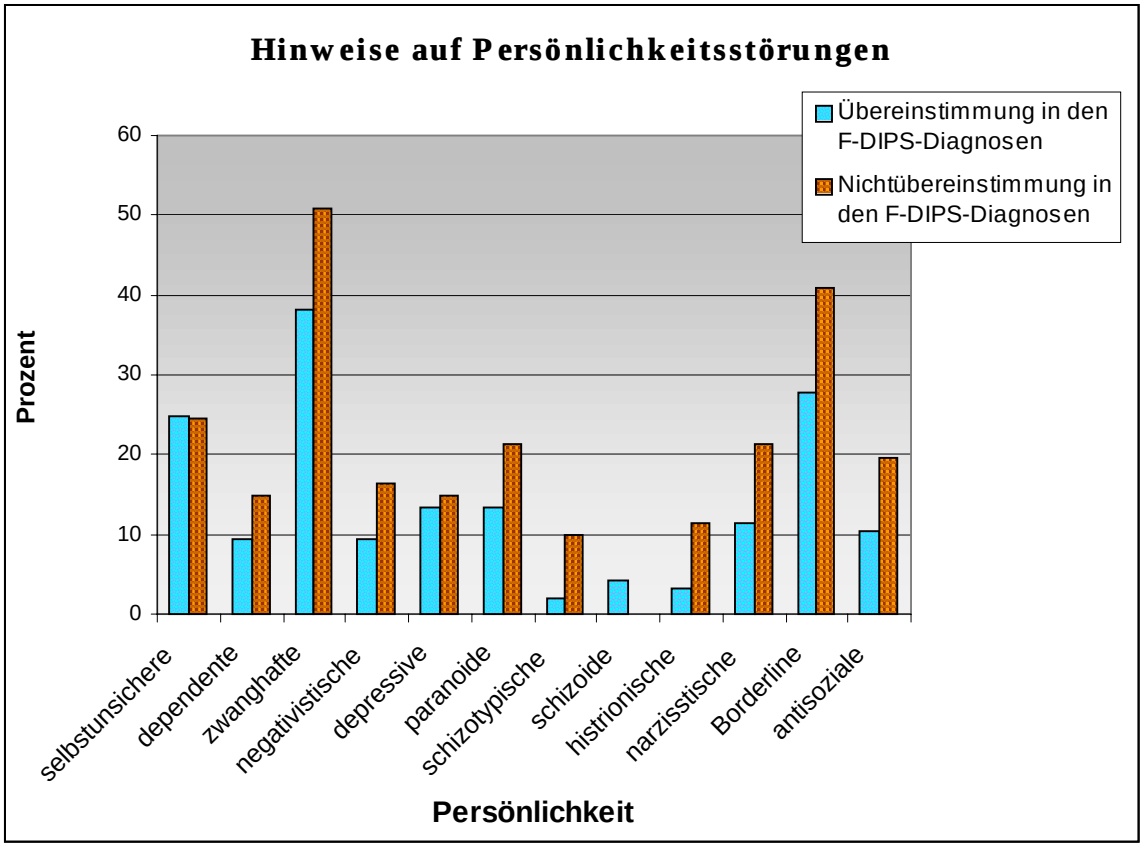
8.3.5.2 Globale Erfassung des Funktionsniveaus (GAF)

Auf einem hypothetischen Kontinuum von psychischer Gesundheit bis Krankheit wurden psychische, soziale und berufliche Funktionsbeeinträchtigungen von 1 (ständige Gefahr, sich oder andere schwer zu verletzen) bis 100 (hervorragende Leistungsfähigkeit) von den Interviewern eingeschätzt. Durchschnittlich ergab sich daraus ein aktuelles Funktionsniveau von 63 und ein Funktionsniveau für das letzte Jahr von 59. Die ermittelten Funktionsniveaus unterscheiden sich weder von einem Interview zum nächsten, noch von der Übereinstimmungsstichprobe zur Nicht-Übereinstimmungsstichprobe (t-Test für gepaarte bzw. für unabhängige Stichproben).

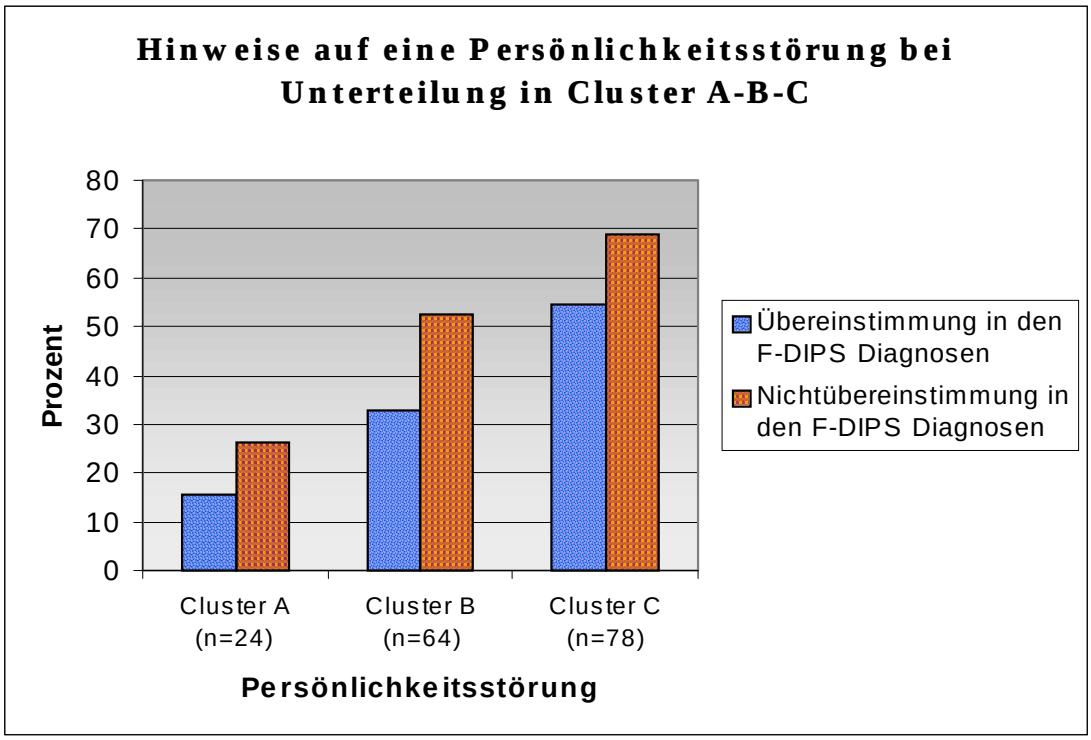
8.3.5.3 Persönlichkeitsstörung

Aus dem SKID-II ergaben sich in der untersuchten Stichprobe (n=158) 31 % ohne einen Hinweis auf eine Persönlichkeitsstörung. Bei 28 % besteht der Verdacht auf eine Persönlichkeitsstörung, bei 11 % auf 2 Persönlichkeitsstörungen und bei weiteren 30 % auf 3 oder mehr Persönlichkeitsstörungen. Im einzelnen waren dies zu 43 % Hinweise auf eine zwanghafte und zu 33 % auf eine Borderline-Persönlichkeitsstörung. Andere Persönlichkeitsstörungen mit einem hohen prozentualen Anteil sind die selbstunsichere (25 %), die paranoide (17 %) und die narzisstische (15 %).

Lediglich bei Patienten mit Hinweis auf eine schizotypische ($p<0,05$) und bei solchen mit Hinweis auf eine histrionische Persönlichkeitsstörung ($p<0,05$) gibt es signifikante Unterschiede (Mann-Whitney-U-Test) zwischen der Gruppe der Patienten mit nichtübereinstimmenden Diagnosen und denen mit übereinstimmender F-DIPS-Diagnose. Beide Persönlichkeitsstörungen kamen jedoch nur bei einem sehr geringen Teil der Patientenstichprobe vor, weshalb sie irrelevant sind. Insgesamt sind jedoch Tendenzen für mehr Persönlichkeitsstörungen bei der Nichtübereinstimmungsgruppe in der Grafik zur Häufigkeitsverteilung sichtbar.



Fasst man die Persönlichkeitsstörungen als Cluster, wie im DSM unterteilt, aus Cluster A (paranoide, schizoide, schizotypische Persönlichkeitsstörung), Cluster B (antisoziale, Borderline, histrionische und narzisstische Persönlichkeitsstörung) und Cluster C (selbstunsichere, dependente, zwanghafte, negativistische und depressive Persönlichkeitsstörung) zusammen, werden die in den Einzeldiagnosen nur angedeuteten Effekte signifikant in Bezug auf die Cluster B-Persönlichkeitsstörungen.



[Die prozentualen Prävalenz-Unterschiede für die einzelnen Cluster, die in der Grafik auffallen, kommen durch die unterschiedliche Anzahl an Persönlichkeitsstörungen, die in ein Cluster einfließen, zustande – bei Cluster A drei Störungen, bei Cluster B vier und bei Cluster C fünf Störungen].

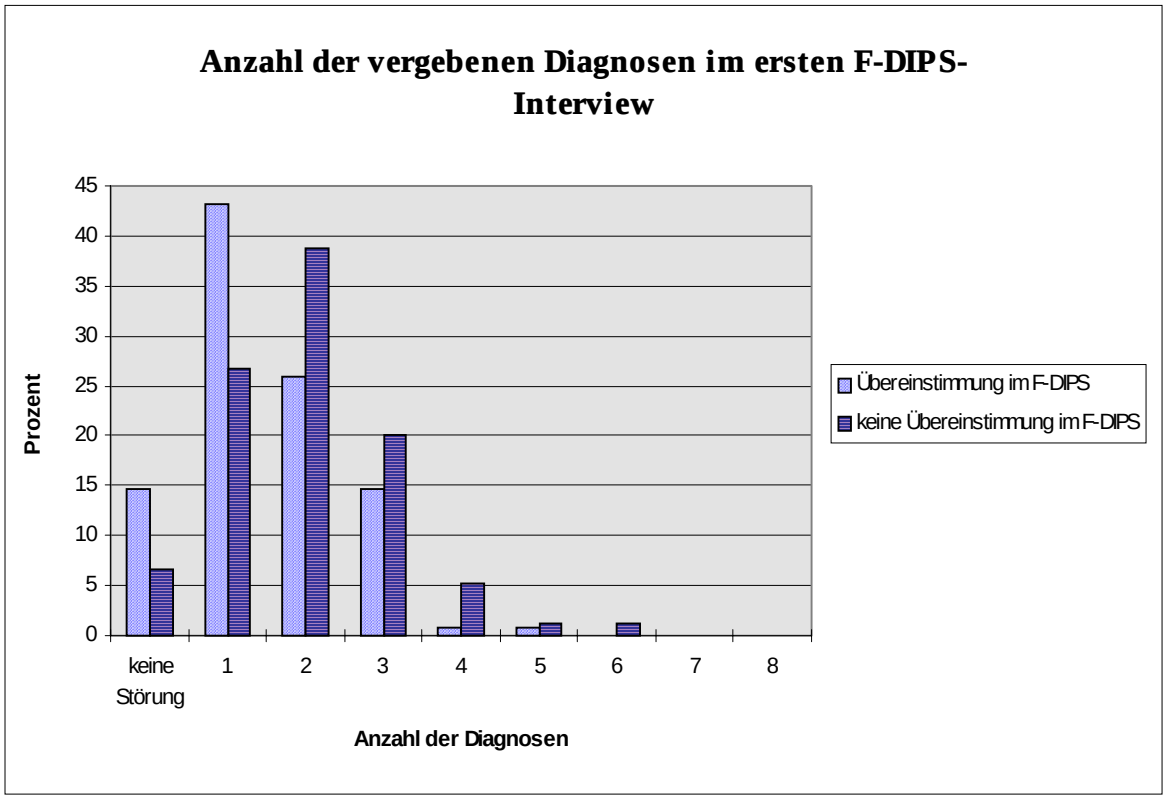
Die Unterschiede zwischen der Gruppe mit übereinstimmender F-DIPS-Diagnose und der ohne übereinstimmender Diagnose sind bei Cluster B-Persönlichkeitsstörungen bei Berechnung des Mann-Whitney-U-Tests hoch signifikant ($p < 0,01$), wenn man Mehrfachdiagnosen (z.B. Borderline und antisozial) aufaddiert und signifikant ($p < 0,05$), wenn wie in der Grafik nur das Vorhandensein oder Nicht-Vorhandensein eines Clusters bewertet wird. In der Gruppe der Patienten, bei denen es zu keiner Übereinstimmung in der F-DIPS-Diagnose kam, finden sich also insgesamt häufiger Hinweise auf antisoziale, Borderline-, histrionische und narzisstische Persönlichkeitsstörungen als in der Gruppe der Patienten, die die Interviewer im F-DIPS gleich beurteilten.

Untersucht man die Cluster-B-Patienten (n=64) genauer, kann man feststellen, dass diese Patienten zu 50 % unterschiedlich in ihren Achse-I-Diagnosen (Gesamtstichprobe: 39 %) eingeschätzt wurden, sie im

Durchschnitt 2,1 Diagnosen erhielten (Non-Cluster-B-Patienten: 1,3), die ersten Interviews im Durchschnitt 120 Minuten andauerten (Non-Cluster-B-Patienten: 103). Der BDI aus der ersten Messung (Mittelwert = 25; s = 9,4) ist zudem in dieser Teilstichprobe höchst signifikant ($p < 0{,}001$) verschieden von dem in der zweiten Messung (Mittelwert = 17; s = 10,2).

8.3.5.4 Anzahl der F-DIPS-Diagnosen

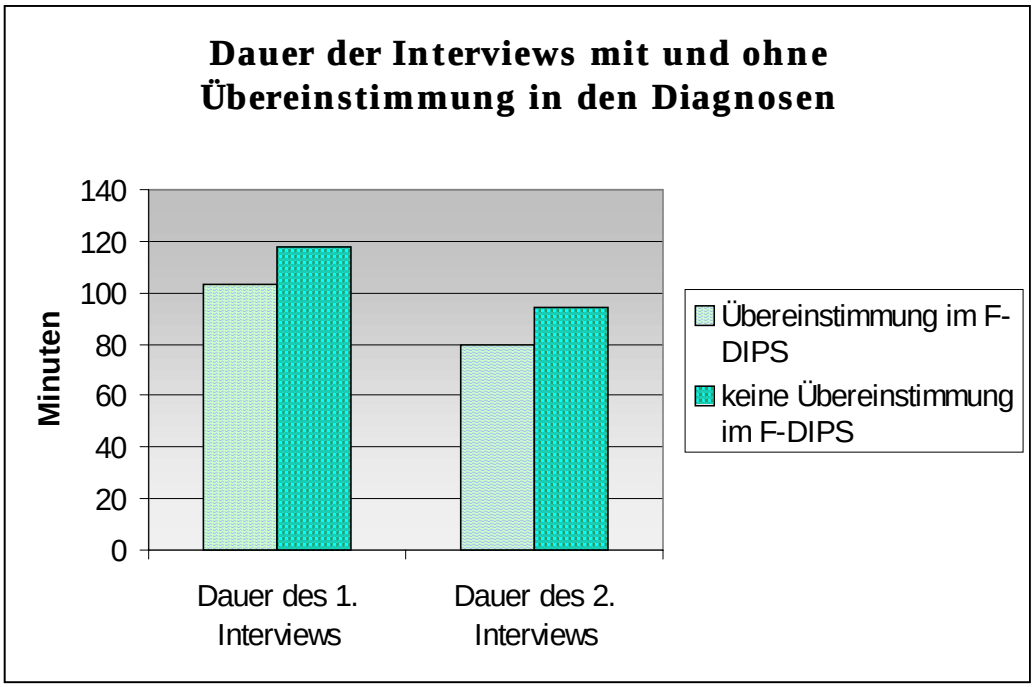
Bei Patienten mit nicht übereinstimmenden Diagnosen im F-DIPS wurden im t-Test höchst signifikant ($p=0{,}001$) mehr Diagnosen gestellt als bei Patienten mit übereinstimmenden Einschätzungen. Der Anzahl-Mittelwert für Patienten, bei denen keine übereinstimmende Diagnose zustande kam, lag bei 2 (s=1,1) Störungen, der für Patienten mit übereinstimmender Diagnose bei 1,5 (s=1) Störungen.



8.3.5.5 Dauer des Interviews

Ein weiterer Punkt, der zu Fehlern in der Beurteilung führen kann, ist die Dauer des Interviews. Ist ein Interview nicht lange genug, werden Fragen evtl. nur flüchtig beantwortet und Störungen übersehen. Dauert es zu lange, können die Konzentrationsfähigkeit und die Motivation nachlassen.

Die Interviews ohne übereinstimmende Diagnose dauerten (t-Test) signifikant länger ($p < 0,05$) als Interviews mit Übereinstimmung in den diagnostischen Urteilen der beiden Rater.

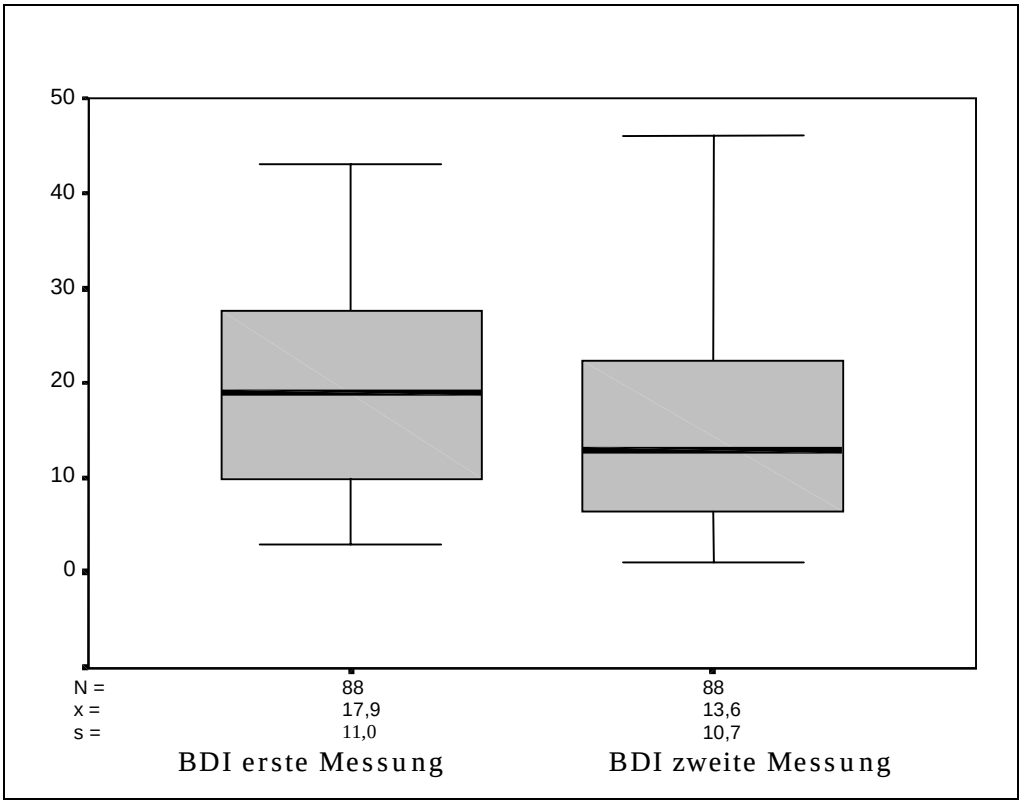


8.3.5.6 Depressivität

Das BDI wurde zu beiden Messzeitpunkten vorgelegt. Die Stichprobe, die beide Tests ausfüllte, liegt bei $n = 88$. Die Stichprobe ist kleiner als die Gesamtstichprobe, da das BDI erst zu einem späteren Zeitpunkt der Studie in zweifacher Weise vorgegeben wurde. Beim BDI, das zum ersten

F-DIPS-Interview vorgelegt wurde, ergab sich ein Mittelwert von 17,9 (s=11,1), beim zweiten BDI ein Mittelwert von 13,6 (s=10,7). Der t-Test bei gepaarten Stichproben ergab eine höchst signifikante Änderung des BDI-Wertes vom ersten bis zum zweiten Interview ($p < 0{,}001$).

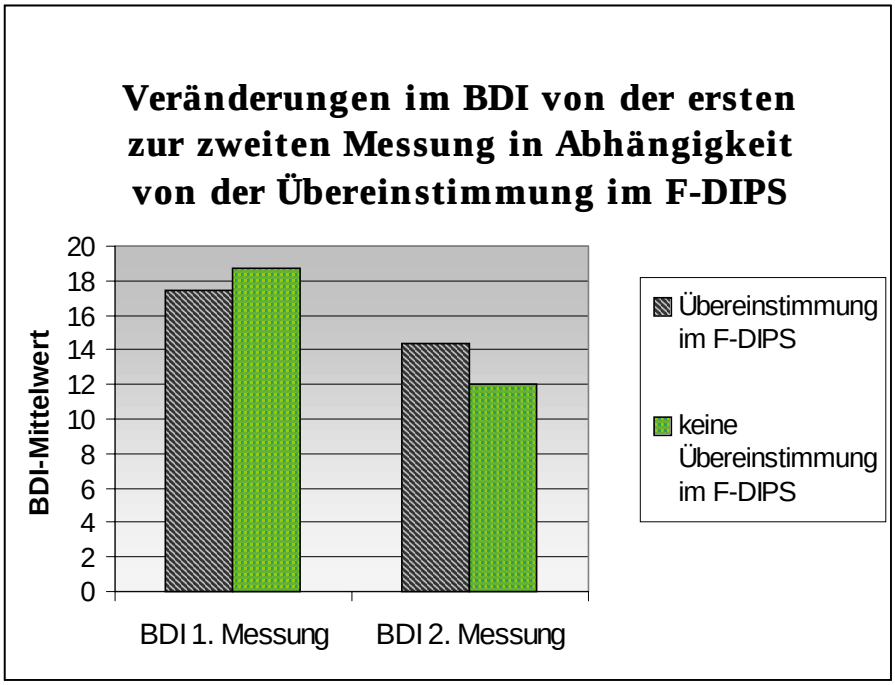
BDI – Werte im Vergleich vom ersten zum zweiten Interview



Bei Unterteilung der Stichprobe in Patienten, bei denen die Diagnosen im F-DIPS übereinstimmten und in Patienten, bei denen die Diagnosen nicht übereinstimmten, zeigen sich dann aber nur noch kleinere Unterschiede zwischen den beiden Stichproben:

- der erste BDI-Mittelwert bei der Nichtübereinstimmungs-Stichprobe liegt bei 18,7, der der Übereinstimmungsstichprobe bei 17,5

- der zweite BDI-Mittelwert der Nichtübereinstimmungs-Stichprobe liegt bei 12, der der Übereinstimmungsstichprobe bei 14,4
- diese Differenz ist bei der Nichtübereinstimmungsstichprobe (Mittelwert: 6,8) höchst signifikant ($p<0,001$), bei der Übereinstimmungsstichprobe (Mittelwert: 3,1) hoch signifikant ($p<0,01$).

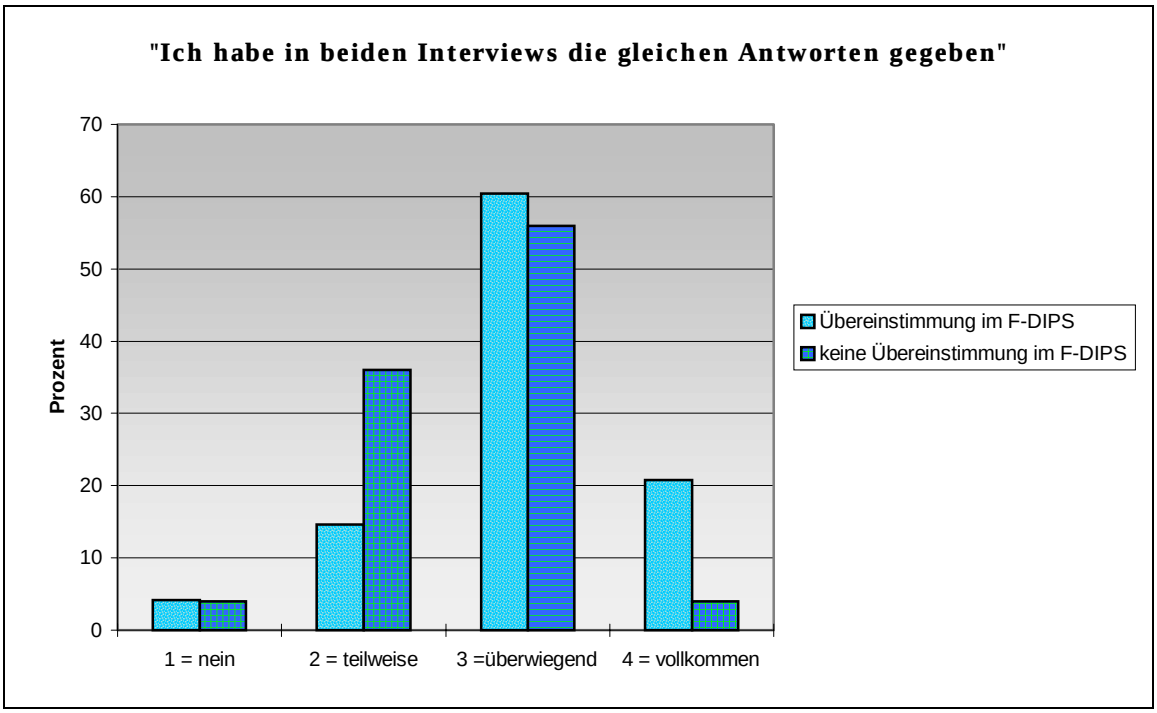


- Die Unterschiede zwischen den beiden Gruppen sind nicht signifikant.

8.3.5.7 Einschätzung des eigenen Antwortverhaltens im F-DIPS-Interview

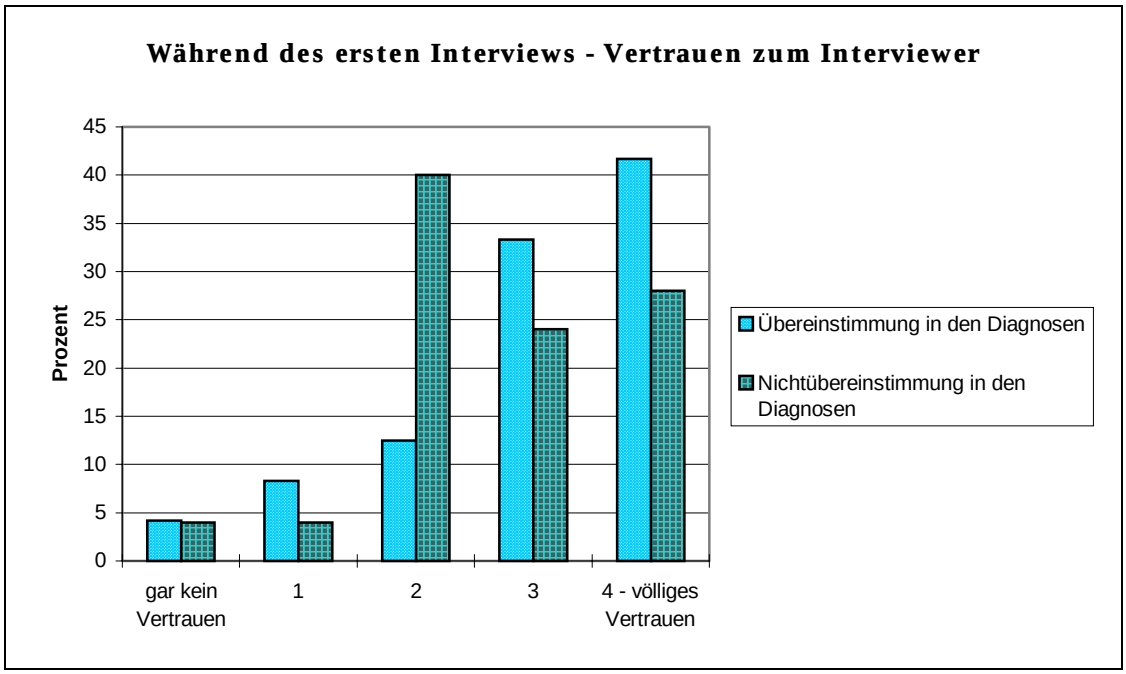
Die Einschätzungen der Patienten zum Erleben der beiden Interviews wurden mit Hilfe des Mann-Whitney-U-Tests verglichen.
Lediglich eine Frage erwies sich als signifikantes Unterscheidungsmerkmal zwischen der Gruppe der Patienten ohne übereinstimmender Diagnose im F-DIPS und der mit übereinstimmender Diagnose.

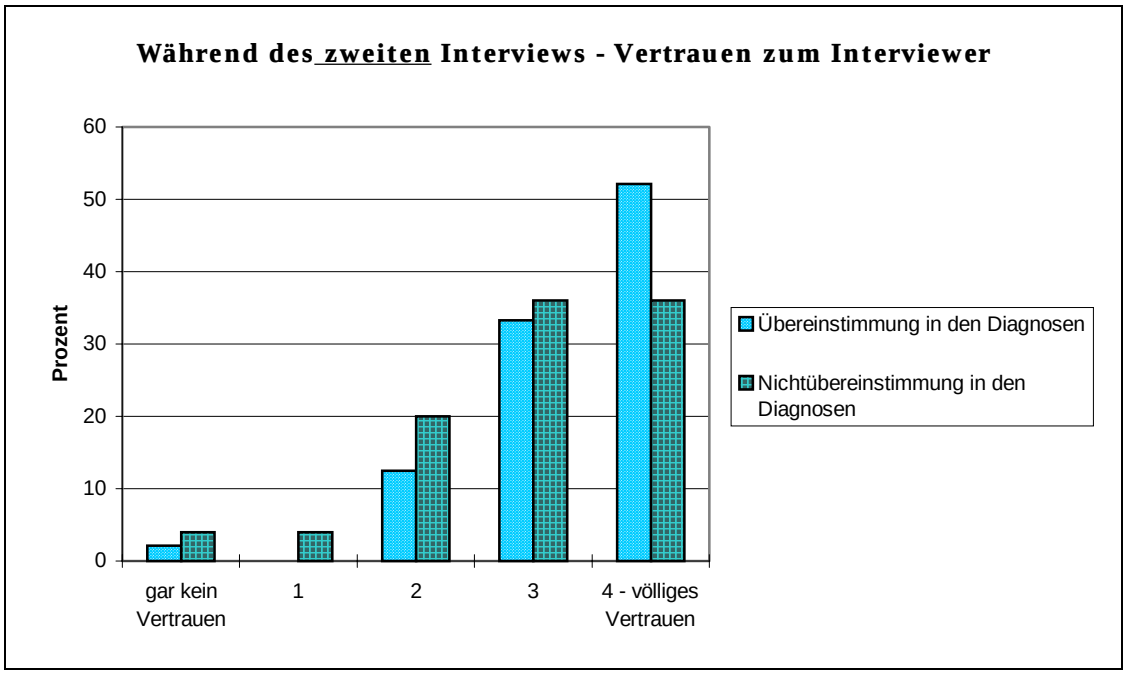
Dabei handelt es sich um die Auskunft, ob sie in beiden Interviews die gleichen Antworten gegeben haben. Die Patienten, die keine übereinstimmende Diagnose erhalten haben, sind signifikant weniger davon überzeugt, in beiden Interviews die gleichen Antworten gegeben zu haben.



Die Frage nach dem Vertrauen zu den beiden Interviewern fiel so aus, dass die Patienten zum ersten Interviewer in beiden Stichproben (Übereinstimmung / Nichtübereinstimmung im F-DIPS) ein geringeres Vertrauen zeigten als zum zweiten Interviewer (t-Test bei gepaarten Stichproben).

Dieser Unterschied im Vertrauen zwischen erstem und zweitem Interview war in der Übereinstimmungsstichprobe signifikant ($p<0,05$), in der Nichtübereinstimmungsstichprobe nicht signifikant.





Die Unterschiede bzgl. des Vertrauens von Patienten zu ihren Interviewern, bei denen die Diagnosen übereinstimmend gegeben wurden, zu jenen, die keine übereinstimmenden Diagnosen erhalten haben, sind jedoch nicht signifikant. Das Ergebnis ist unabhängig von der Person des Interviewers, da jeder Interviewer abwechselnd das erste oder das zweite Interview durchführte.

Auch die anderen Fragen danach, ob die Patienten

- aufgrund ihrer anderen Stimmung in beiden Interviews unterschiedliche Antworten gegeben haben
- ein Interview als angenehmer als das andere erlebt haben
- oder sich zwischen den Interviews Gedanken gemacht haben und zu Neubewertungen gekommen sind,

wurden in der Übereinstimmungs- und Nichtübereinstimmungsstichprobe nicht verschieden beantwortet.

Im einzelnen lässt sich bzgl. der Frage 5, ob es konfundierende Variablen gibt, die die Genauigkeit der Messung beeinflussen, sagen, dass unter den Patienten, bei denen keine übereinstimmende Diagnose gegeben wurde,

- häufiger ein Hinweis auf eine Cluster B- Persönlichkeitsstörung gegeben war
- häufiger komorbide Störungen zu finden waren
- die Interviews länger dauerten und
- die Patienten häufiger angaben, dass sie nur teilweise die gleichen Antworten in beiden Interviews gegeben hatten.

Als beinahe signifikant verschieden erwiesen sich die Patienten mit übereinstimmender Diagnose von denen ohne übereinstimmende Diagnose in Bezug auf die Abstände der beiden Interviews voneinander.
Hinsichtlich des Funktionsniveaus (GAF, Achse V) oder der Depressivität existieren keine signifikanten Gruppenunterschiede zwischen Patienten mit und ohne Übereinstimmung in der F-DIPS-Diagnose.

8.3.6 Fehlerquellen bei der Anwendung des F-DIPS (Beantwortung der 6. Frage: Gibt es besonders häufige Fehlerquellen bei der Anwendung des F-DIPS?)

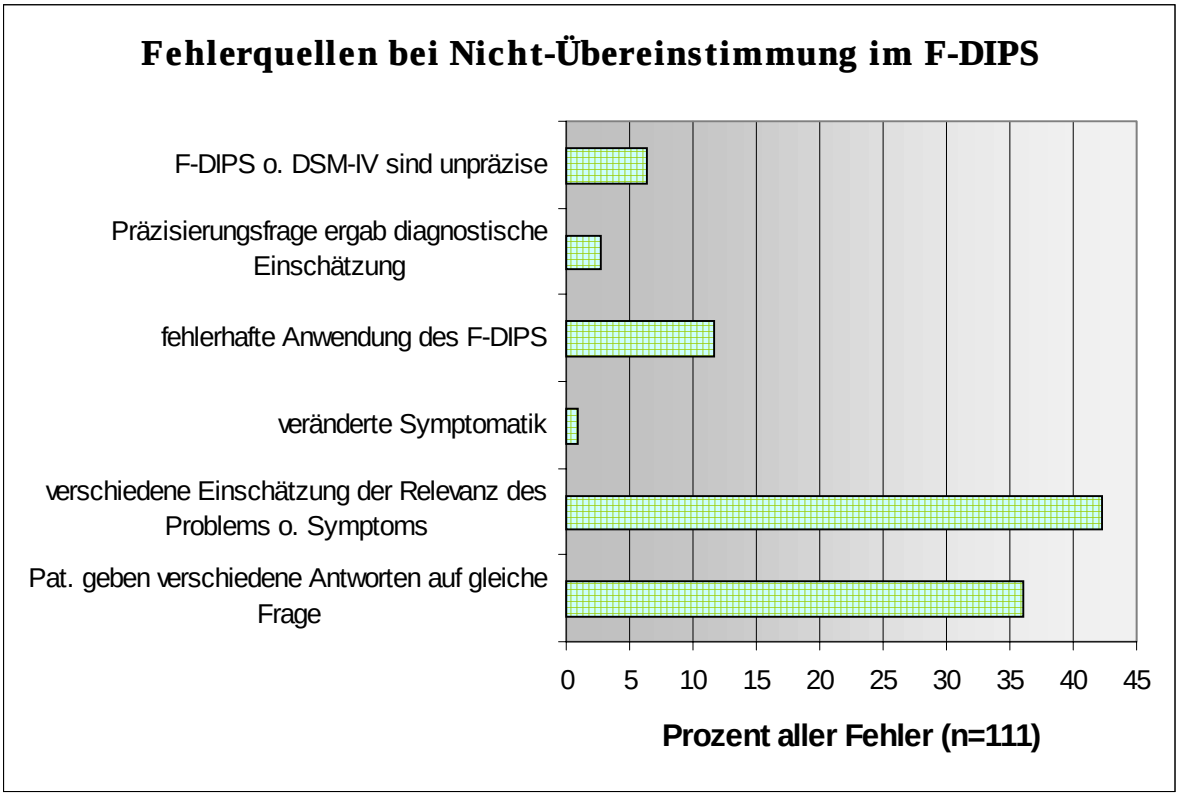
Für eine Auswahl an Interviews, bei denen die Interviewer keine Übereinstimmung der Diagnosen erzielten, wurde eine genauere Fehleranalyse durchgeführt.
Insgesamt ist bei 75 Interviews bzgl. mindestens einer Diagnose eine fehlende Übereinstimmung vorhanden. Von diesen 75 Interviews existiert zu 34 ein Video, das innerhalb der Supervision angeschaut und Auffälligkeiten daraus notiert wurden. Außerdem existieren zu den meisten Interviews Aussagen der Interviewer zu Schwierigkeiten mit der Diagnostik.

Fehlerquellen wurden gesucht in der:

- Informationsvarianz (z.B. wenn Patienten auf gleiche Fragen verschiedene Antworten geben)
- Subjektvarianz (z.B. durch Therapieeinfluss ändert sich die Symptomatik beim Patienten)
- Situationsvarianz (der Patient ist in verschiedenen Phasen einer Störung)
- Beobachtungsvarianz (z.B. Symptomgewichtung, Relevanz der vorliegenden Symptome wird verschieden eingeschätzt)
- Interviewerfehler (F-DIPS wird fehlerhaft angewandt)
- Diagnostische Einschätzung wird durch Präzisierungsfrage erreicht, nicht allein durch F-DIPS- Fragen
- DSM-IV oder F-DIPS selbst unzureichend (unzureichender Informationsgewinn durch unpräzise Fragen oder Kriterien)

Die Kriterienvarianz wurde als Fehlerquelle nicht genauer untersucht, da sie aufgrund der Anwendung eines strukturierten Interviewleitfadens und festgelegten Auswertungskriterien nicht auftreten dürfte, sonst wäre das F-DIPS falsch angewendet worden (Interviewerfehler).

Insgesamt wurden 111 Fehler bei Nicht-Übereinstimmung von Diagnosen anhand von Protokollbögen, Supervisionsaufzeichnungen und Videos überprüft. Diese hohe Zahl kommt dadurch zustande, dass pro Interview oft mehrere Diagnosen nicht miteinander übereinstimmten oder mehrere Fehler gefunden wurden.



Als häufigste Quelle (42,3 %) für Nicht- Übereinstimmungen zeigte sich die unterschiedliche Einschätzung der Relevanz des vom Patienten geschilderten Problems (Beobachtungsvarianz). Zweithäufigste Fehlerquelle waren die unterschiedlichen Antworten, die die Patienten beim ersten und dann beim zweiten Interview auf die gleichen Fragen gaben (Informationsvarianz, 36 %). Einen kleineren Anteil an den Fehlern hatte die fehlerhafte Anwendung des F-DIPS (11,7%), dass F-DIPS oder DSM-IV unzureichend sind oder eine Präzisierungsfrage nötig war, mit deren Hilfe einer der beiden Interviewer schließlich die Diagnose gestellt hatte (insgesamt 9 %).

Zur Informationsvarianz führte, wenn Patienten

- unklare und widersprüchliche Antworten gaben, die sich auch nach mehrmaligem Nachfragen nicht auflösen ließen
- bei einem Interview sehr verschlossen waren, beim anderen offener
- aus unklaren Gründen andere Informationen gaben.

Zur Beobachtungsvarianz kam es öfters, wenn

- Patienten bagatellisierten oder übertrieben entweder durch ausgleichendes Ratingverhalten (z.B. bei Bagatellisierung höher raten) oder durch korrektes Übernehmen der Patientenantwort
- Patienten bei jeder Frage zunächst mit „ja“ antworten
- extrem viele Symptome berichtet werden
- das Auftreten des Patienten nicht mit den berichteten Symptomen übereinstimmt
- ein Interviewer die klinische Bedeutung der angegebenen Symptome über- oder unterschätzt.

Fehler bei der Anwendung des F-DIPS bezogen sich auf

- falsches Lesen innerhalb des Interviews (n=2)
- falsches Lesen der Kriterien (n=2)
- Symptome, die innerhalb einer Störung auftreten und auch bei anderen Störungen erfragt werden, werden mehrfach als Störung kodiert (z.B. Generalisierte Angststörung und Major Depression oder: Vermeiden von öffentlichen Plätzen wegen einer Zwangsstörung = Agoraphobie) (n=5)
- unvollständiges Befragen einzelner Störungen, z.B. weil das Interview bereits sehr lange dauerte (n=2)
- Übertragungsfehler (n=2)

Fehlerquellen, die an Mängeln des F-DIPS oder des DSM-IV lagen, ergaben sich durch:

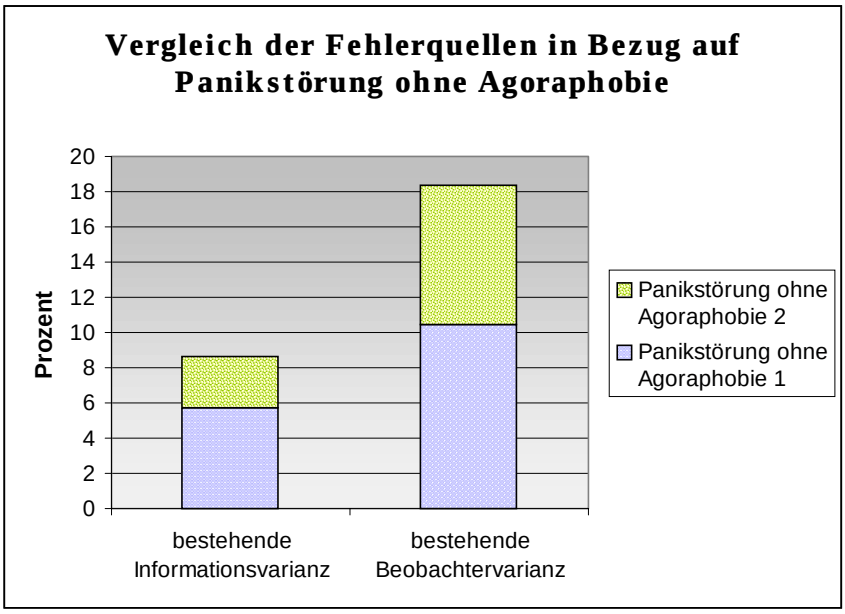
- zu komplex und allgemein formulierte Fragen, die teilweise aus sehr langen Sätzen mit Aufzählungen und mit und/oder- Verknüpfungen

bestehen (z.B.: „Gab es im letzten Jahr Zeiten, in denen Sie ganz plötzlich einen Ansturm von intensiver Angst, Furcht oder intensivem Unbehagen oder das Gefühl eines drohenden Unglücks spürten? Manche sagen dazu auch Angstanfälle oder Panikattacken.“)

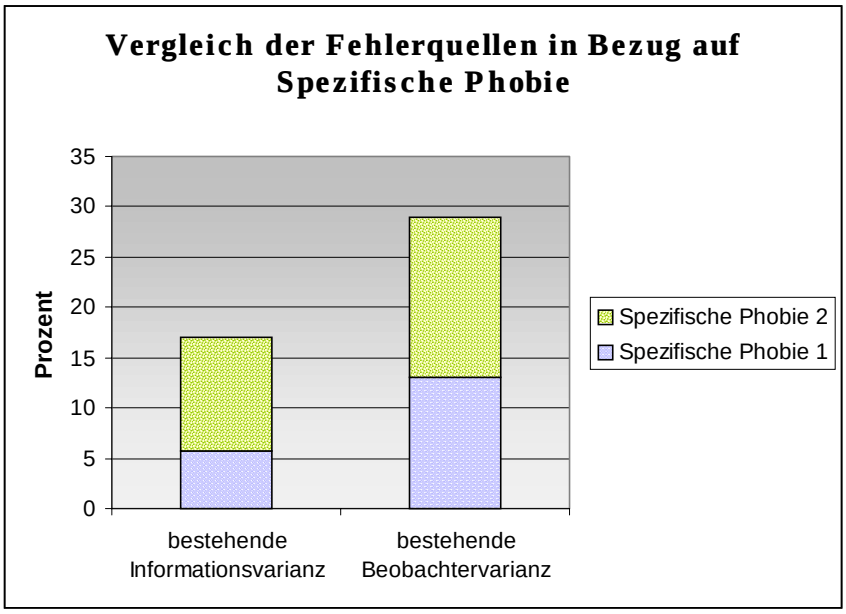
- Sprungregeln, mit deren Hilfe Fragen zu einigen Störungsbildern reduziert werden können und die zum nächsten Störungsbild weiterleiten. Diese sind teilweise von der ersten Antwort des Patienten abhängig (z.B. Panikstörung), teilweise werden sämtlichen möglichen Symptome erfragt, bevor ein Sprung möglich ist (Sozialphobie, Generalisierte Angststörung, Zwangsstörung). Antwortet ein Patient zunächst auf die Frage nach z.B. sozialen Ängsten verneinend, gibt dann jedoch eine problematische zwischenmenschliche Situation an, die er vermeidet, wird in diesem Fall weiter gefragt. Dies kann zu mehr positiven Diagnosen bei diesem Störungsbild führen und bringt Untersucher in die Schwierigkeit, inwiefern sie die unterschiedlichen Informationen werten
- bei der Sozialphobie wird gefragt, inwiefern die Patienten ängstlich oder nervös in sozialen Situationen werden oder diese vermeiden. Die Frage nach der *Ängstlichkeit oder Nervosität* ist dabei so weit formuliert, dass viele positive Antworten erfolgen, mehr als auf die Eingangsfrage für die Störung. Auf der anderen Seite lautet das Kriterium des DSM-IV: „Eine *dauerhafte und übertriebene Angst* vor einer oder mehreren sozialen oder Leistungssituationen...“. Der Interviewer muss sich also entscheiden, wie er die positiven Antworten auf die situativen Fragen bewertet und ob dies mit dem „schärfer formulierten“ DSM-Kriterium übereinstimmt
- die Reihenfolge der Befragung der Störungen. An der Stelle, an der die Generalisierte Angststörung (GAS) bewertet werden soll, sind die notwendigen Informationen, nämlich, ob eine Depression besteht, noch nicht vorhanden. Deshalb fällt die Bewertung der GAS schwer.
- das Fehlen einer Eingangsfrage nach dem Hauptproblem. Dadurch sind Interviewer und Befragte gleichermaßen bei den ersten Fragen verunsichert und halten sich bei der Abklärung zu lange auf.

Im folgenden wurden die Hauptfehlerquellen Beobachtervarianz und Informationsvarianz im einzelnen in ihrem Zusammenhang mit den Störungsbildern untersucht. Hierfür wurden diejenigen Interviews, bei denen der Fehler in der Beobachtervarianz lag, herausgesucht und mit denen verglichen, bei denen der Fehler in der Informationsvarianz lag. Wenn es komorbide Diagnosen gab, wurden diese nicht herausgefiltert, so dass die in den folgenden Grafiken abgebildeten Effekte eher größer sein dürften.

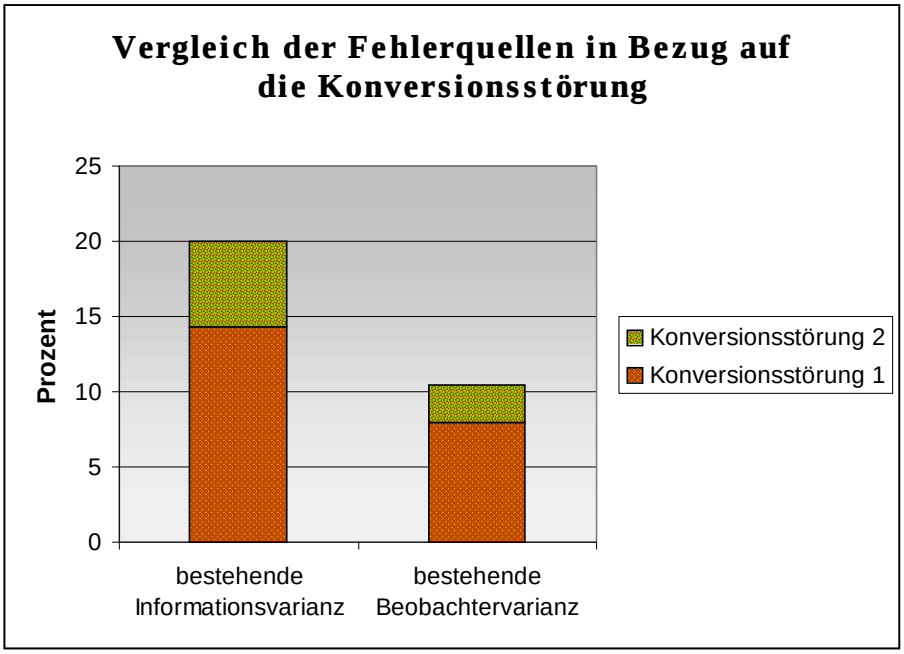
- Bei der Panikstörung ohne Agoraphobie kommen häufiger Fehler vor, die mit der Einschätzung der Relevanz des Problems (Beobachtervarianz) zu tun haben, vor, seltener solche, die mit unterschiedlicher Auskunft auf Seiten des Patienten zu tun haben.

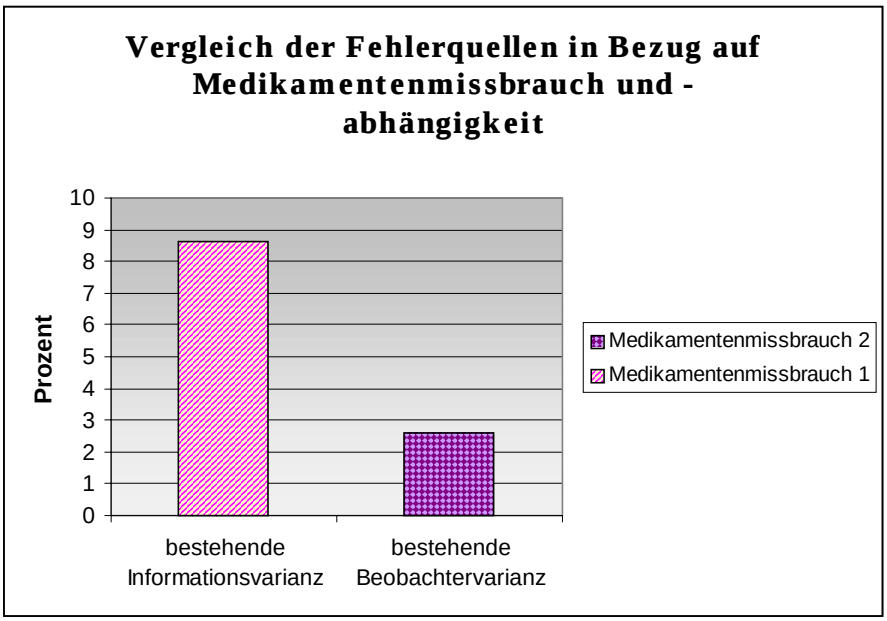


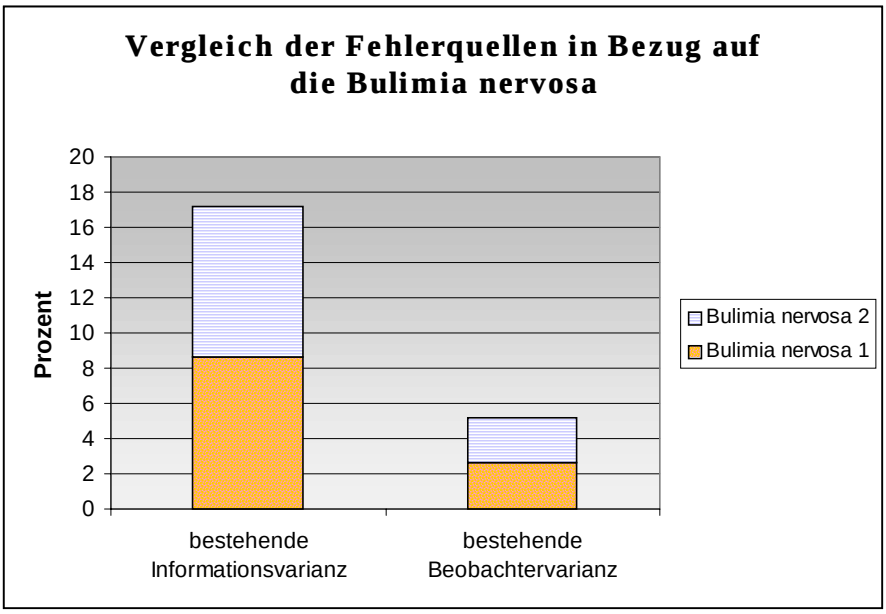
- Der gleiche Effekt kann auch bzgl. der Spezifischen Phobie festgestellt werden.



- Hingegen gibt es einen umgekehrten Effekt, nämlich dass die Patienten häufiger eine unterschiedliche Information auf die gleiche Frage geben, bei der Konversionsstörung, bei Medikamentenmissbrauch und –abhängigkeit sowie bei der Bulimia nervosa (siehe Grafiken).




Vergleich der Fehlerquellen in Bezug auf Medikamentenmissbrauch und -abhängigkeit
Prozent
10
9
8
7
6
5
4
3
2
1
0
bestehende Informationsvarianz
bestehende Beobachtervarianz
Medikamentenmissbrauch 2
Medikamentenmissbrauch 1


Vergleich der Fehlerquellen in Bezug auf die Bulimia nervosa
Prozent
20
18
16
14
12
10
8
6
4
2
0
bestehende Informationsvarianz
bestehende Beobachtervarianz
Bulimia nervosa 2
Bulimia nervosa 1

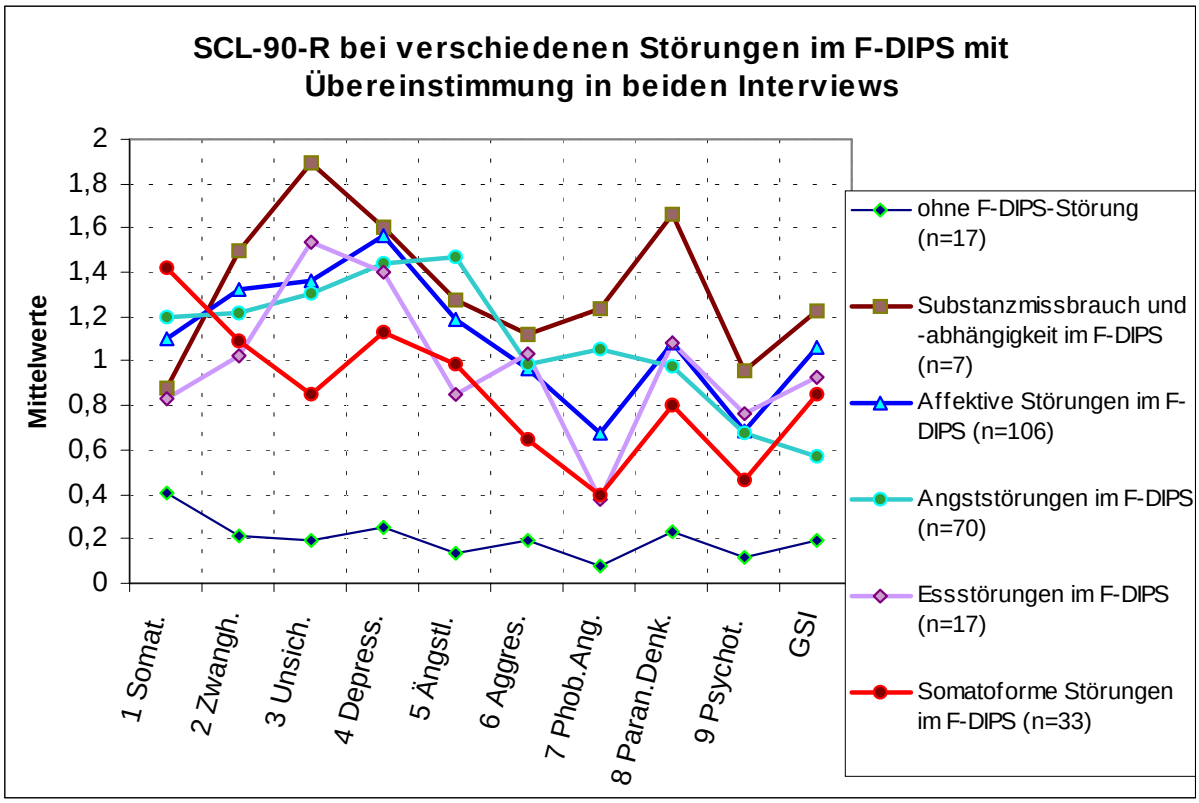
Bei allen anderen Störungen sind die Beobachtervarianz und die Informationsvarianz in etwa gleich verteilt, so dass sich dort kein Zusammenhang zwischen Störungsbild und Fehlerart zeigt.

Die Frage 6, ob es besonders häufige Fehlerquellen für eine Nichtübereinstimmung der Diagnosen gibt, lässt sich mit „ja“ beantworten. Es handelt sich um die Beobachtungs- und die Informationsvarianz.

Außerdem fällt auf, dass die Fehlerquellen sowohl bei der Panikstörung ohne Agoraphobie als auch bei der Spezifischen Phobie eher in der Beobachtervarianz zu suchen sind, die Interviewer also die Relevanz der geschilderten Symptome verschieden einschätzten. Bei der Konversionsstörung, bei Medikamentenmissbrauch oder –abhängigkeit und bei der Bulimia nervosa hingegen liegen die Fehlerquellen bei Nichtübereinstimmung eher in der Informationsvarianz, so dass die Patienten unterschiedliche Aussagen in beiden Interviews machten.

8.3.7 Validität des F-DIPS (Beantwortung der 7. Frage: Inwiefern diagnostiziert das F-DIPS die Störung, die der Patient tatsächlich hat?)

Vergleicht man zunächst die verschiedenen Störungsoberklassen bzgl. ihrer Ausprägung in den SCL-90-R-Skalen, ergeben sich folgende Verteilungen. Dabei wurden Komorbiditäten nicht herausgefiltert, so dass es sein kann, dass jemand einmal z.B. in der Stichprobe „Angststörung“ und einmal in der Stichprobe „Affektive Störung“ mit seinen SCL-Werten auftaucht. Die Patienten, die in der folgenden Grafik „ohne F-DIPS-Störung“ sind, sind solche, die in beiden Interviews übereinstimmend keine Störung erhalten haben.



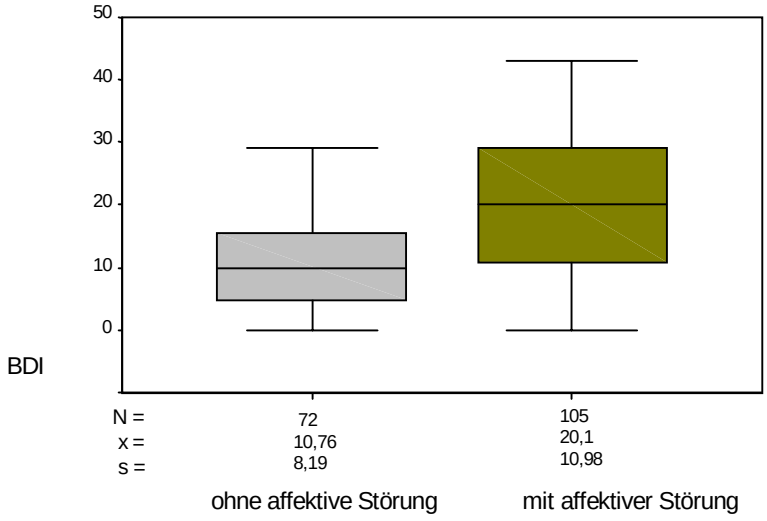
Spitzenwerte zeigen in der Skala „Somatisierung“ die Patienten, die auch im F-DIPS die eine somatoforme Störung erhalten haben. Bei den Skalen „Zwanghaftigkeit“, „Unsicherheit im Sozialkontakt“, „Phobische Angst“,

„Paranoides Denken“ und „Psychotizismus“ erreichen die Patienten, die im F-DIPS eine Substanzgebrauchsstörung erhielten, die höchsten Werte. Die höchsten Werte auf der Skala „Ängstlichkeit“ zeigt die F-DIPS Patientengruppe mit einer Angststörung. „Depressivität“ und „Aggressivität“ liegen bei mehreren Störungen ähnlich hoch. Die psychisch am belastetsten Patienten (GSI) sind die mit Substanzgebrauchsstörungen, die am wenigsten belasteten Patienten die mit Angststörungen im F-DIPS.

Im folgenden werden die Störungsoberklassen Affektive Störungen, Angststörungen (darunter Sozialphobie, Spezifische Phobie, Agoraphobie), Somatoforme Störungen (darunter Hypochondrie), Substanzmissbrauch und –abhängigkeit und Essstörungen mit den entsprechenden Fragebogenmittelwerten und mit den Entlassungsdiagnosen verglichen. Für die Darstellung der Fragebogenunterschiede wurden Boxplots ausgewählt.

1. Störungsoberklasse Affektive Störungen

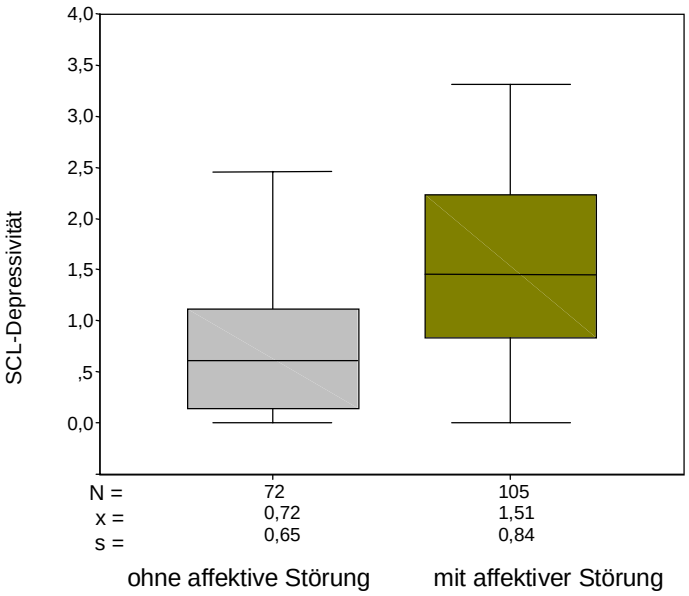
BDI-Werte bei Patienten ohne und mit F-DIPS Diagnose "Affektive Störung"



Es bestehen höchst signifikante Unterschiede im t-Test für unabhängige Stichproben ($p < 0{,}001$) zwischen Patienten mit affektiver Störung und

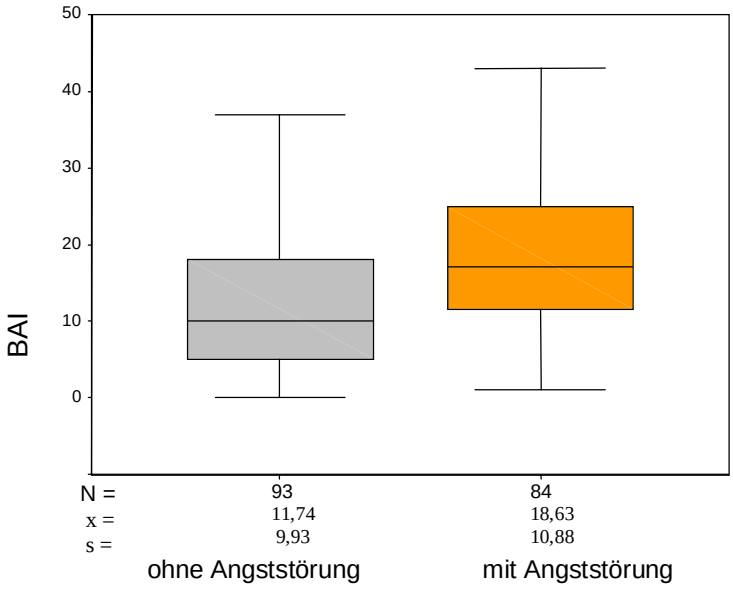
Patienten ohne affektive Störung im F-DIPS bzgl. ihrer BDI-Werte und bzgl. ihrer Werte in der SCL-90-R-Skala „Depressivität“.

SCL-90-R- Depressivität bei Patienten ohne und mit F-DIPS-Diagnose „Affektive Störung“



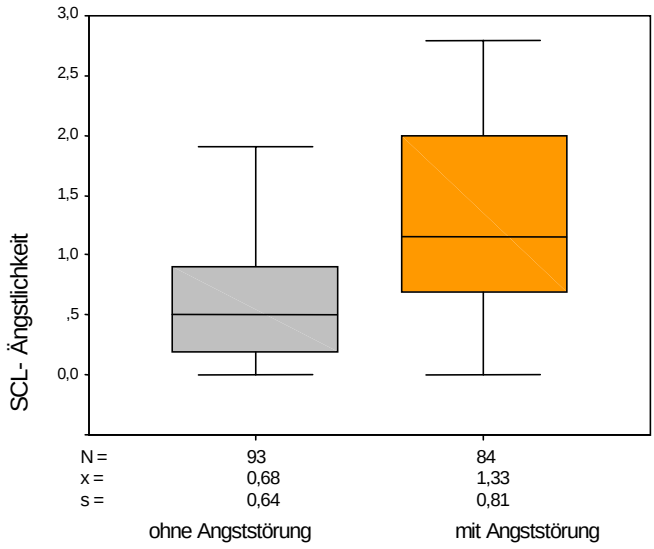
2. Störungsoberklasse Angststörungen

BAI-Werte bei Patienten ohne und mit F-DIPS-Diagnose "Angststörung"

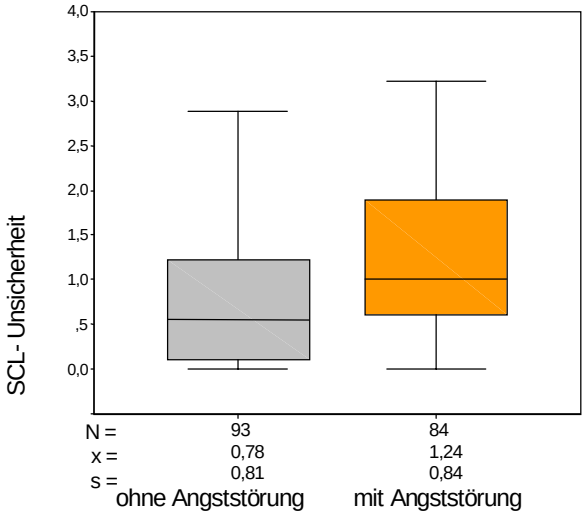


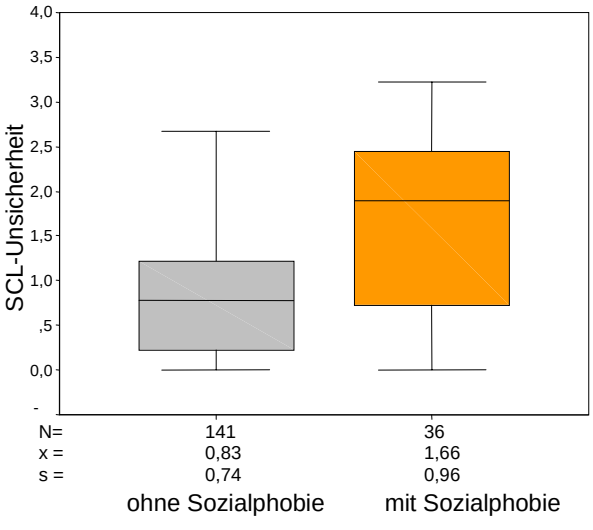
Es besteht ein höchst signifikanter Unterschied (p< 0,001) im BAI zwischen Patienten ohne und Patienten mit Angststörung im F-DIPS sowie auf der SCL-90-R-Skala „Ängstlichkeit“.

SCL-90-R-Ängstlichkeit- Werte bei Patienten ohne und mit F-DIPS-Diagnose „Angststörung“



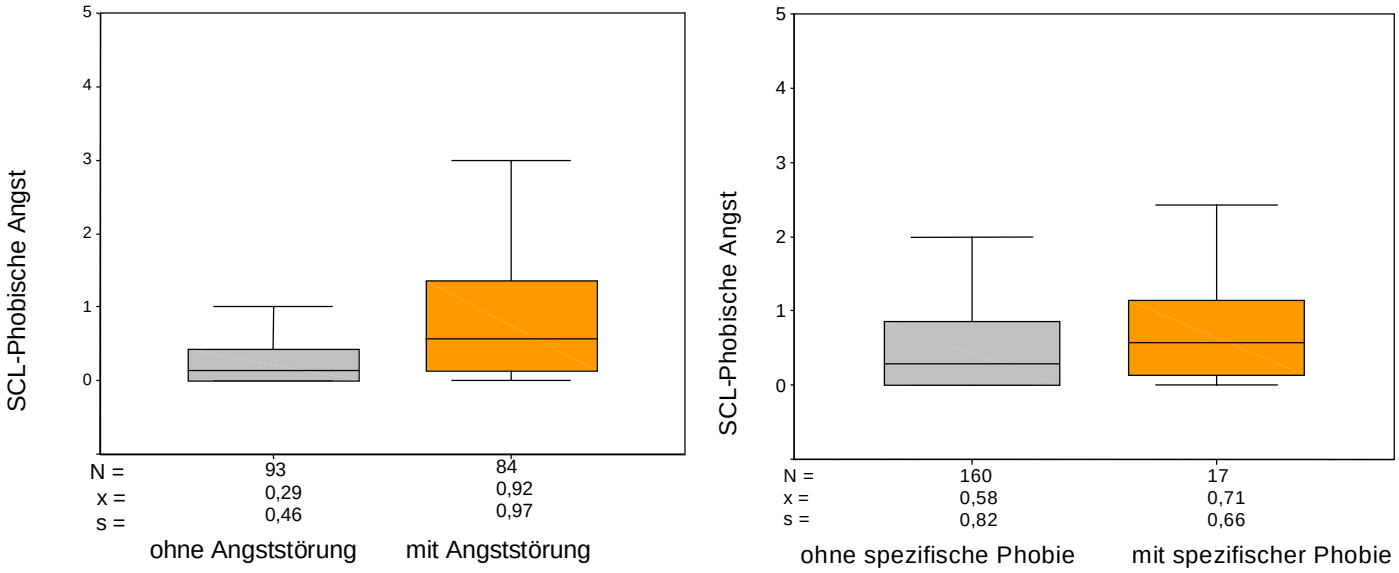
SCL-90-R-Unsicherheit im Sozialkontakt bei Patienten ohne und mit F-DIPS-Diagnose "Angststörung" / F-DIPS-Diagnose „Sozialphobie“





Die Unterschiede zwischen Patienten ohne Angststörung und mit Angststörung im F-DIPS bzgl. ihrer Ausprägung in der SCL-90-Skala „Unsicherheit im Sozialkontakt“ sind höchst signifikant (p<0,001). Ebenfalls höchst signifikant im t-Test und etwas eindrücklicher im Boxplot, unterscheiden sich Patienten mit und ohne Sozialphobie in der Skala „Unsicherheit im Sozialkontakt“.

SCL-90-R- Phobische Angst bei Patienten ohne und mit F-DIPS-Diagnose „Angststörung“ / „Spezifische Phobie“

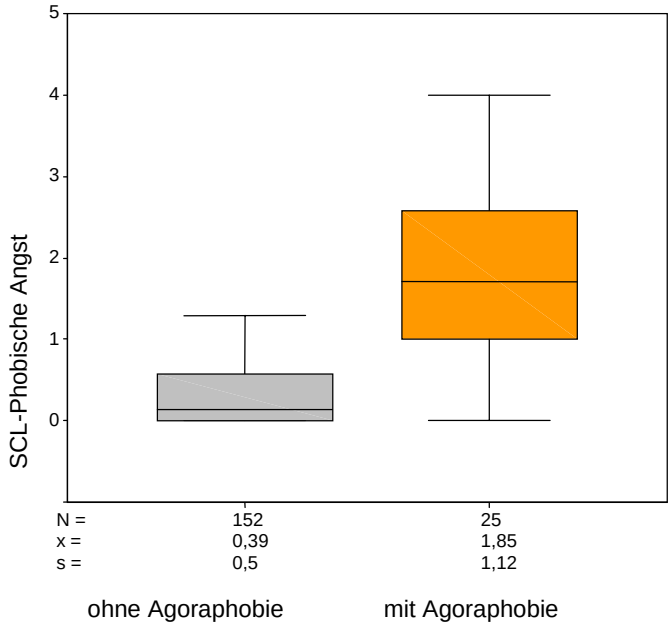


Die Mittelwertunterschiede im t-Test sind höchst signifikant (p< 0,001) unterschiedlich für Patienten mit Angststörung im F-DIPS zu denen ohne Angststörung im F-DIPS bzgl. der SCL-90-R-Skala „Phobische Angst“. Da unter der SCL-Skala „Phobische Angst“ eher agoraphobische Ängste beschrieben werden, dient die Skala nicht dazu, eine spezifische Phobie zu validieren. Dies zeigt sich dann in nicht signifikanten Mittelwert-Unterschieden im t-Test zwischen Patienten mit und ohne spezifischer Phobie im F-DIPS.

Wird zum Vergleich die Agoraphobie in ihrer Ausprägung auf der SCL-Skala „Phobische Angst“ untersucht, zeigt sich ein höchst signifikanter

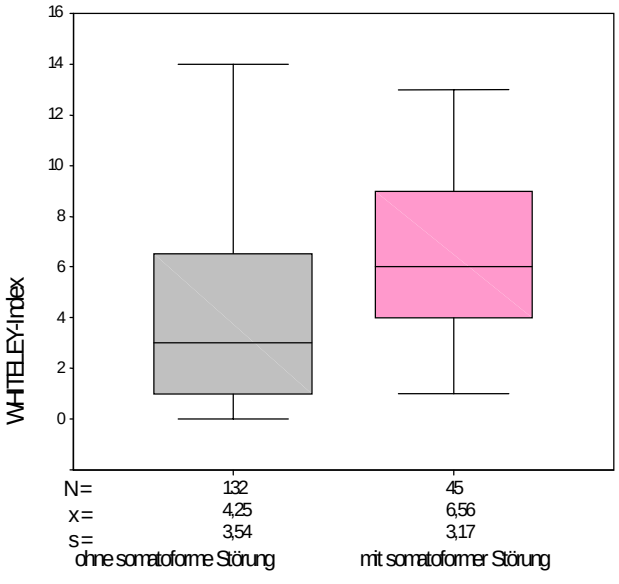
Unterschied (p< 0,001) zwischen Patienten mit und solchen ohne Agoraphobie im F-DIPS.

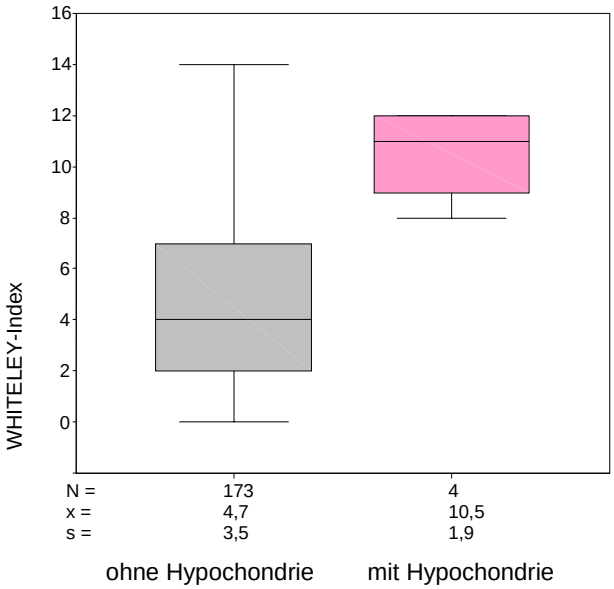
SCL-90-R- Phobische Angst bei Patienten ohne und mit F-DIPS-Diagnose „Agoraphobie“



3. Störungsoberklasse Somatoforme Störungen

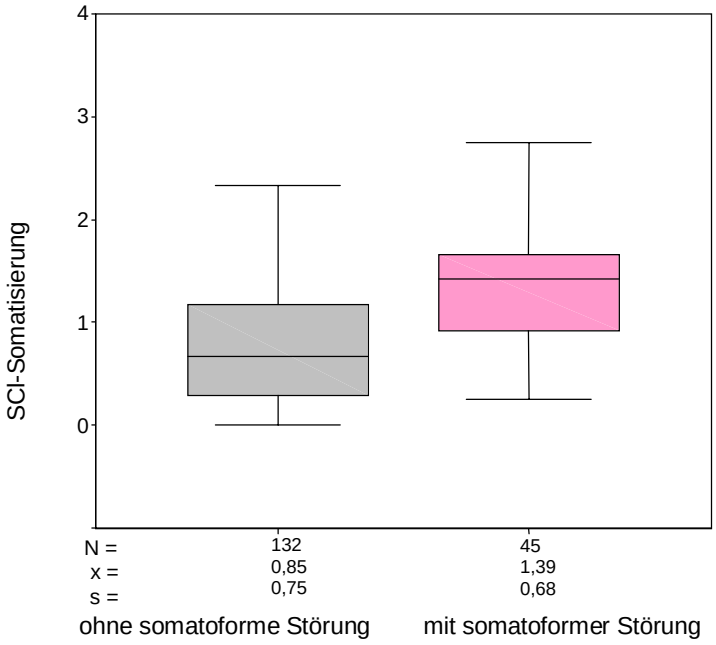
WHITELEY-Index bei Patienten ohne und mit somatoformer Störung im F-DIPS / Hypochondrie im F-DIPS





Patienten mit einer somatoformen Störung bzw. mit einer Hypochondrie im F-DIPS unterscheiden sich höchst signifikant (p< 0,001) von Patienten ohne somatoforme Störung bzw. ohne Hypochondrie bzgl. ihrer Whiteley-Indices.

SCL-90-R-Somatisierung bei Patienten ohne und
mit somatoformer Störung im F-DIPS



Es gibt höchst signifikante Unterschiede (p<0,001) in der Ausprägung auf der SCL-90-R-Skala „Somatisierung“ zwischen Patienten mit und ohne somatoformer Störung im F-DIPS.

In der folgenden Tabelle sind die Übereinstimmungen zwischen F-DIPS-Oberklassen und der Entlassungsdiagnose, zwischen F-DIPS-Oberklasse und der Patienteneinschätzung ihres Hauptproblems sowie zwischen der Entlassungsdiagnose und der Patienteneinschätzung berechnet worden, um zu sehen, ob die F-DIPS-Diagnosen valide sind.
Da die Behandler das Ergebnis aus dem F-DIPS mitgeteilt bekamen, hätte diese Information in die Entlassungsdiagnose mit einfließen können.

Die Ergebnisse wurden für den Fall berechnet, dass es eine Übereinstimmung bei den beiden F-DIPS-Diagnosen gegeben hatte.

Tabelle: Validität der F-DIPS-Diagnosen in Bezug auf Entlassungsdiagnose und Patienteneinschätzung

Störungsoberklasse	Vergleich zwischen	Häufigkeit +¹/+² -/+	+/- -/-	κ
Angststörungen	F-DIPS¹ vs. Entlassungsdiagnose²	21 0	26 69	.49
	F-DIPS¹ vs. Patienteneinschätzung²	35 3	28 77	.54
	Entlassungsdiagnose¹ vs. Patienteneinschätzung²	28 10	6 99	.70
Affektive Störungen	F-DIPS¹ vs. Entlassungsdiagnose²	70 36	11 48	.43
	F-DIPS¹ vs. Patienteneinschätzung²	44 1	52 53	.37
	Entlassungsdiagnose¹ vs. Patienteneinschätzung²	31 43	14 62	.24
Somatoforme Störungen	F-DIPS¹ vs. Entlassungsdiagnose²	24 24	9 112	.47
	F-DIPS¹ vs. Patienteneinschätzung²	25 22	3 101	.57
	Entlassungsdiagnose¹ vs. Patienteneinschätzung²	30 17	15 89	.50
Substanzabhängigkeit und -missbrauch	F-DIPS¹ vs. Entlassungsdiagnose²	4 1	3 176	.66
	F-DIPS¹ vs. Patienteneinschätzung²	1 1	4 159	.27
	Entlassungsdiagnose¹ vs. Patienteneinschätzung²	0 2	4 159	-.02
Essstörungen	F-DIPS¹ vs. Entlassungsdiagnose²	17 4	0 166	.88
	F-DIPS¹ vs. Patienteneinschätzung²	11 6	5 146	.63
	Entlassungsdiagnose¹ vs. Patienteneinschätzung²	14 3	6 145	.73

Die Angststörungen wurden im F-DIPS deutlich häufiger vergeben als in der Entlassungsdiagnose, so dass es nur zu einer mäßigen Übereinstimmung kam. Die Patienteneinschätzung bzgl. des Hauptproblems stimmte etwas besser mit der F-DIPS-Diagnose überein, nur beim Ver-

gleich „Entlassungsdiagnose vs. Patienteneinschätzung“ fand sich jedoch eine gute Übereinstimmung.

Hinsichtlich der affektiven Störungen gab es nur geringe Übereinstimmungen aus den verschiedenen diagnostischen Quellen, wobei die zwischen F-DIPS und Entlassungdiagnose noch als mäßig eingeschätzt werden kann, die zwischen den anderen diagnostischen Quellen als ungenügend.

Auch bei somatoformen Störungen liegt nur eine mäßige Übereinstimmung zwischen F-DIPS, Entlassungsdiagnose und Patienteneinschätzung vor.

4. Störungoberklasse Substanzabhängigkeit /-missbrauch

Bei der geringen Basisrate der Substanzabhängigkeiten in der Stichprobe kann die Übereinstimmung zwischen F-DIPS und Entlassungsdiagnose als gut bis ausgezeichnet bezeichnet werden. Die Patienteneinschätzung stimmt hier ungenügend mit F-DIPS oder noch weniger mit der Entlassungsdiagnose überein.

5. Störungsoberklasse Essstörungen

Die Essstörungen werden ausgezeichnet übereinstimmend von F-DIPS und Entlassungsdiagnose erfasst und stimmt auch gut zwischen Entlassungsdiagnose und Patienteneinschätzung überein. Die Übereinstimmung zwischen F-DIPS und Patienteneinschätzung ist als mittel gut zu bezeichnen.

Die Frage 7 bzgl. der Validität des F-DIPS lässt sich folgendermaßen beantworten:
Die Validität ist insgesamt gut, wobei Angststörungen, affektive Störungen und somatoforme Störungen zwar sehr gut mit Fragebogeninstrumenten (Selbstauskunft der Patienten zu ähnlichen Symptom-Fragen wie im F-DIPS) übereinstimmen, jedoch weniger gut mit der Einschätzung der Behandler (jeweils mäßige Übereinstimmung, κ = .43 bis .49) in Einklang zu bringen sind.
Substanzabhängigkeit und Essstörungen wurden nicht mit Fragebogeninstrumenten validiert. Die Übereinstimmung zwischen F-DIPS und Behandlern ist im Unterschied zu den vorher genannten Störungsoberklassen jedoch gut bis ausgezeichnet (κ = .66 und .88).

8.3.8 Vergleich der Gütekriterien des F-DIPS mit anderen Instrumenten (Beantwortung der 8. Frage: Wie zuverlässig und valide misst das F-DIPS psychische Störungen im Vergleich zu anderen diagnostischen Interviews?)

Um das F-DIPS mit anderen diagnostischen Interviews vergleichen zu können, wurde mittels der Literaturdatenbanken „Medline“ und „PsycLit“ unter den Stichworten „diagnostic interview“, „psychiatric interview“, „reliability“, „validity“, „DSM-IV“, „ICD-10“, „SCID“, „SKID“, „ADIS“, „SCAN“, „CIDI“, „DIS“, „IDCL“, verschiedenen Autorennamen und den Verknüpfungen nach aktuellen Studien gesucht und die Reliabilitäts- und Validitätsdaten anhand der gefundenen Originalartikel miteinander verglichen.

Dabei fiel auf, dass es zu Instrumenten, die nach DSM-IV klassifizieren, noch nicht viele Untersuchungen zur Retestreliabilität gibt, sondern oft

nur im Zusammenhang mit der Verwendung des Instruments in Studien, eine Überprüfung der Reliabilität für einen Teilstörungsbereich vorgenommen wurde. Aus diesem Grund wurde das F-DIPS zum Vergleich in eine Reihe mit auch älteren diagnostischen Interviews gestellt. Die folgende Tabelle zeigt die Reliabilitätskoeffizienten des F-DIPS im Vergleich zu den anderen diagnostischen Interviews.

Tabelle: Retest-Reliabilität im Vergleich zwischen F-DIPS und anderen diagnostischen Interviews

Diagnose	CIDI (DSM-III-R)	M-CIDI (DSM-IV)	ADIS (DSM-III-R)	SKID (DSM-III-R)	IDCL (DSM-IV)	DIPS lifetime (DSM-III-R)	DIPS derzeit (DSM-III-R)	F-DIPS (DSM-IV)
Major Depression	.66	.68	.57	.69	.73	.69	.71 Y	.72
Bipolare Störung	.47	.64		.70	.85	-.01		.83
Dysthyme Störung	.47	.70	.57	.80	.50	.51	.84 Y	.51 Y
Panikstörung	.84	1	.69	.27	.88	.70	.81	.69
Agoraphobie	.65	.84	.85	.36	.52	.68	.84	.73
Sozialphobie		.80				.68	.72	.53
Phobische Störungen	.57				.67	.68	.97	.60
Generalisierte Angststörung		.45				.46	.66 Y	.61 Y
Somatoforme Störungen		.62				.67 Y	.75 Y	.66
Alkoholabhängigkeit	.79	.78		.75	.80			.84 Y
Drogen-/ Medikamentenabhängigkeit	.73	.64		.75	.77			.77 Y
Essstörung		.56				.87	.80	.89

Alle in der Tabelle angegebenen Werte sind Kappa-Werte mit Ausnahme der durch Y gekennzeichneten Yule's Y-Werte, die dann angegeben wurden, wenn die Basisraten zu gering waren, um exakte Kappa-Koeffizienten berechnen zu können und somit die Y-Werte eine genauere Schätzung der Reliabilität angeben.

Während es bei den verglichenen Instrumenten allgemein eine homogen (bei allen Interviews) ausgezeichnete Übereinstimmung bei den Substanzabhängigkeiten gibt, besteht eine homogen gute Übereinstimmung bei der Major Depression.
Heterogener sind die verschiedenen Instrumente bzgl. der Dysthymen Störung und den Angststörungen: Hier gibt es Schwankungen in den Kappa-Werten von .50 bis .80 bei der Dysthymen Störung, von .27 bis .88 bei der Panikstörung und von .36 bis .85 bei der Agoraphobie oder von .57 bis .97 bei den phobischen Störungen.

Um genau zu sehen, welchen Stellenwert das F-DIPS unter den diagnostischen Interviews einnimmt, wurden in der folgenden Tabelle die Reliabilitäts-Mittelwerte für einzelne Störungen, bei denen es genügend vergleichbare Werte gab, berechnet.

	DE-PRESSION	DYSTH.	PANIK-STÖR.	AGORA-PHOB.	SOZIAL-PHOBIE	PHOB. STÖR.	ALKOHOL-ABHÄNG.	ESS-STÖRUNG
Mittel-werte	.68	.61	.73	.68	.68	.70	.79	.78
F-DIPS	↑	↓	≅	↑	↓	↓	≅	↑

↑ bedeutet: F-DIPS besser, ↓: F-DIPS schlechter, ≅: F-DIPS in etwa gleich.

Vergleicht man die Übereinstimmungen bzgl. der Störungsoberklassen von dem Vorgänger-Interview DIPS mit denen des F-DIPS, wurden mit Hilfe des F-DIPS Angststörungen und die Kategorie „keine Störung“ deutlich weniger übereinstimmend diagnostiziert (siehe folgende Tabelle), die anderen Störungen in etwa gleich gut, bzw. etwas besser.

Tabelle: Retest-Reliabilität in Bezug auf die Störungsoberklassen – Vergleich DIPS – F-DIPS

Störungsoberklasse	DIPS (nur Primärdiagnosen) κ	DIPS Y	F-DIPS κ	F-DIPS Y
Angststörungen	.71	.75	.64	.65
Affektive Störungen	.56 (Depression)	.71 (Depression)	.71	.72
Somatoforme Störungen	.28	.75	.66	.72
Substanzmissbrauch und –abhängigkeit			.70	.85
Essstörungen	.80	.92	.89	.94
keine Achse I - Störung	.83	.92	.65	.76

Die Überprüfung der Validität des F-DIPS wurde ähnlich der des DIPS anhand einer Fragebogenbatterie durchgeführt. Hier erbrachte das F-DIPS dem DIPS ähnliche, und zwar gute (+) Ergebnisse für Angststörungen, Affektive Störungen und Somatoforme Störungen.

Tabelle: Validität der verschiedenen diagnostischen Interviews

Diagnose	CIDI-Auto (ICD-10)	CIDI (DSM-IV)	DIS (ICD-10)	CIS-R (DSM-III-R)	SCID (DSM-IV)	DIPS (DSM-III-R)	F-DIPS (DSM-IV)
Major Depression	-	~	-	-	~	+	+ (~)
Angststörungen	-	-		-	-	+	+ (~)
Somatoforme Störungen	-					+	+ (~)
Substanz-abhängigkeit	-	+			+		+
Essstörung	-	~				+	+

In der Tabelle sind diejenigen Felder frei gelassen worden, für die keine Hinweise aus Studien gefunden wurden. Die anderen Zeichen (+, ~ mittlere bis mäßige, - ungenügende) geben Tendenzen aus den gesichteten Studien an.

Wird das F-DIPS mit der Entlassungsdiagnose verglichen, kann die Validität für die Major Depression, die Angststörungen und die Somatoformen Störungen nur als mäßig angesehen werden, was durch ein „~“ in Klammern angegeben wird. Insgesamt stimmt das F-DIPS-Ergebnis mit der Entlassungsdiagnose von κ = .43 (Affektive Störungen) bis κ = .88 (Essstörungen) überein.

Vergleichbar sind diese Werte mit denen der Studie von Rosenman et al. (1997), in der ebenfalls Interview-Diagnosen mit klinischen Diagnosen verglichen wurden.

Es ergibt sich ein deutlich höherer Gesamt- κ –Koeffizient zwischen F-DIPS und Psychiater/Psychologe über die verschiedenen Störungsbereiche (κ =.59) als bei der Rosenman-Untersuchung zwischen dem CIDI in der computerisierten Version und einem Psychiater (κ =.23) und auch ein höherer als in der Untersuchung von Peters & Andrews (1995), die zwischen der computerisierten Version des CIDI und dem Kliniker einen Gesamt-Kappa-Koeffizienten von .40 ermittelten.

Auch in der Untersuchung von Kranzler (1996) wurde die mit dem SCID (Structured Clinical Interview for DSM-IV) erhaltene Diagnose mit einer durch einen Kliniker gestellten Diagnose verglichen. Die Validität entsprach in etwa der des F-DIPS (gut bei Substanzabhängigkeiten, mittel für Major Depression und (schlechter als beim F-DIPS) eine ungenügende für Angststörungen).

Diese Ergebnisse zur Validität sind nicht direkt mit den folgenden vergleichbar, da hier als Validierungskriterium ein zweites diagnostisches Interview verwendet wurde.

In Bezug auf die Depression gab es zwischen DIS und SCAN (Eaton et al., 2000) lediglich eine Übereinstimmung von κ =.20. Bei der Kreuzvalidierung von dem voll strukturierten Interview CIS-R (Clinician Interview Schedule-Revised) mit dem halbstrukturierten SCAN (Brugha et al., 1999) zeigte sich eine ungenügende Konkordanz für die ICD-10-Kategorie „Neurotische Störungen“ und für die Depressive Störung (κ < .25). Da das SCAN als Validierungsinstrument oft eingesetzt wird und es allgemein als valide gilt (Hesselbrock et al., 1999), sprechen diese Ergeb-

nisse gegen die DIS und das CIS-R. Verlässliche Daten zu der Güte des SCAN existieren jedoch nicht.

Frage 8 bzgl. des Vergleichs des F-DIPS mit anderen Instrumenten in Bezug auf die Testgüte lässt sich folgendermaßen beantworten:
Das F-DIPS ist bisher eines der wenigen diagnostischen Interviews, das nach DSM-IV klassifiziert und umfassend hinsichtlich seiner Retest-Reliabilität und der Validität untersucht wurde.
Die Reliabilität ist mit anderen Instrumenten vergleichbar mit Stärken gegenüber anderen Instrumenten in den Essstörungen, der Agoraphobie und der Major Depression. Schwächen in der Reliabilität bestehen im Vergleich zu anderen Instrumenten bei der Dysthymen Störung, der Sozialphobie und bei Phobischen Störungen.
Hinsichtlich der Validität ist das F-DIPS den standardisierten Instrumenten überlegen und kann, sonst vergleichbar mit dem SCID, hinsichtlich der Angststörungen validere Diagnosen als das SCID erstellen.

9 Diskussion und Ausblick

9.1 Diskussion der Ergebnisse

Die Stichprobe der Untersuchung war unselektiert und bestand aus 191 Patienten mit einem größeren Anteil von Frauen (78 % Frauen vs. 22 % Männer) und vielen Nicht-Erwerbstätigen (53 %). 73 Patienten kamen aus einer psychosomatischen Rehabilitationsklinik, 80 waren stationäre oder teilstationäre Patienten aus einer psychosomatischen Akutklinik und 36 waren Patienten aus zwei Ambulanzen von Akutkliniken (psychiatrisch und psychosomatisch). Außerdem nahmen zwei Probandinnen aus dem Projekt „Prädiktoren psychischer Gesundheit junger Frauen“ teil. Bei 17 der Patienten wurde übereinstimmend keine Störung mit dem F-DIPS diagnostiziert. Insgesamt wurden bei jedem Patienten durchschnittlich 1,7 Achse-I Störungen diagnostiziert, der Stichprobenmittelwert für das Beck-Depressionsinventar lag bei 16,5. Zudem wurde bei 69 % der Patienten ein Hinweis auf eine oder meist mehrere Persönlichkeitsstörungen durch den SKID-II-Fragebogen gefunden. Damit kann die Untersuchungsstichprobe insgesamt als schwer gestört bewertet werden.
Dies dürfte auch die Ergebnisse der Untersuchung maßgeblich beeinflusst haben.
Indem zwei Interviews beim selben Patienten durch zwei verschiedene Interviewer durchgeführt wurden, konnte die Retest-Reliabilität bestimmt werden. Diese lag bei 16 von 22 im F-DIPS erfragten Störungen im ausgezeichneten oder guten Bereich ($\kappa > .65$). Bei 3 Störungen fand sich eine mäßige bis mittlere (κ von .40 bis .65) und bei 3 Störungen eine ungenügende ($\kappa < .40$) Übereinstimmung. Auf die Störungsoberklassen bezogen gab es immer eine gute bis ausgezeichnete Übereinstimmung zwischen den beiden Interviewern mit Ausnahme der Angststörungen, wo die Übereinstimmung nur eine mittlere war. Das F-DIPS zeigte Stärken besonders bei der Zwangsstörung ($\kappa = .67$, Y= .90), der rezidivierenden Major Depression ($\kappa = .80$), der Bipolaren Störung ($\kappa = .83$), der Hypochondrie ($\kappa = .75$), bei Alkohol- ($\kappa = .40$, Y = .84) und Drogenabhängigkeit ($\kappa = .67$, Y = 1.0),

bei Anorexia (κ = .78, Y = .91) und Bulimia nervosa (κ = .90, Y = .95). Bei den Essstörungen erreicht das F-DIPS im Vergleich zu M-CIDI und DIPS deutlich höhere Übereinstimmungen. Im Vergleich zu anderen diagnostischen Interviews zeigt es außerdem eine gute Übereinstimmung bei der Agoraphobie (IDCL, SKID und DIPS für DSM-III-R) und bei Somatoformen Störungen (M-CIDI).

Warum das F-DIPS gerade in den Bereichen Essstörung und Somatoforme Störung besser abschneidet als das standardisierte Interview M-CIDI (Wittchen et al., 1998), das sonst immer sehr hohe Reliabilitätswerte aufweist, kann die detaillierte Fehleranalyse ein wenig erhellen.

Hier fiel auf, dass es bei der Bulimia nervosa und auch bei der Konversionsstörung häufiger zu einer Informationsvarianz in der Form, dass die Patientinnen unterschiedliche Aussagen in den beiden Interviews machen, kommt. Die Beobachtungsvarianz (unterschiedliche Einschätzung der Symptome) spielt hierbei nur eine sehr geringe Rolle für die Nichtübereinstimmung. Ein standardisiertes Interview aber minimiert die Beobachtungsvarianz, indem ein Interviewer vollkommen in seinem Handlungsspielraum festgelegt ist und ist darauf angewiesen (wie natürlich jedes andere Interview auch), dass die Patienten eine verlässliche Information geben. Da die Beobachtungsvarianz bei Bulimia nervosa, Konversionsstörung (und auch Medikamentenmissbrauch) nicht mehr wesentlich reduziert werden kann, ist ein standardisiertes Interview einem strukturierten gegenüber nicht mehr im Vorteil. Als ein weiteres Argument für diese Hypothese spricht, dass sowohl bei der Panikstörung ohne Agoraphobie (F-DIPS: κ = .46; Y = .69) als auch bei der spezifischen Phobie die Informationsvarianz gering ist, also Patienten in etwa die gleichen Antworten geben. Die Beobachtungsvarianz ist aber sehr hoch. Bei beiden Störungen zeigt sich das M-CIDI gegenüber dem F-DIPS deutlich überlegen.

Weitere F-DIPS- Störungen mit geringerer Retest-Reliabilität sind die Sozialphobie (κ = .53), Somatoforme Schmerzstörung (κ = .59), Generalisierte Angststörung (κ = .36; Y= .64), Dysthyme Störung (κ = .31; Y= .51) und Gemischte Angst-Depressions-Störung, die jedoch irrelevant ist aufgrund der niedrigen Basisrate von 0,5. Im Vergleich zu anderen diagnostischen

Interviews (M-CIDI, SKID für DSM-III-R, DIPS für DSM-III-R) fallen besonders die schlechteren Kappa-Werte bei der Dysthymen Störung und der Sozialphobie auf.

Betrachtet man die Fehleranalyse bei Nichtübereinstimmung der Diagnosen, sind die häufigsten Fehlerquellen die Beobachtungsvarianz (der Interviewer schätzt die Symptome anders ein: 42 %) und die Informationsvarianz (Inkonsistenzen auf Seiten des Patienten: 36 %). Interviewerfehler (12 %) oder Unklarheiten im F-DIPS bzw. DSM-IV (9 %) spielen eine untergeordnete Rolle. Es kann jedoch davon ausgegangen werden, dass mögliche Unklarheiten im F-DIPS oder in Kriterien des DSM-IV einen indirekten Einfluss auf die Höhe der Beobachtungs- bzw. der Informationsvarianz haben. So muss der Interviewer, wenn Fragen eine geringe Trennschärfe haben und es deshalb häufig zu z.B. positiven Antworten kommt, einschätzen, wie ernst er eine solche Antwort nimmt. Dadurch wird die Beobachtungsvarianz erhöht. Dies könnte z.B. eine Rolle bei der Panikstörung ohne Agoraphobie spielen.
Die Fehlerquellen bei Nichtübereinstimmung der F-DIPS-Diagnosen sind in etwa identisch mit denen bei der Durchführung der Vorgängerversion DIPS (vgl. Schneider et al., 1992). Die Informationsvarianz klärte dort 33 % aller Fehler auf, die Symptomgewichtung 25 % und die Interviewervarianz (unterschiedliche Durchführung des Interviews) 20 %. Die beiden letzteren Fehlerquellen entsprechen in etwa der Beobachtungsvarianz und sind damit mit insgesamt 45 % fast identisch.

Bei der Sozialphobie wie auch bei der Generalisierten Angststörung (GAS) könnte ein Grund für die geringe Übereinstimmung in der Störungseingangsfrage liegen, ob die Patienten soziale Ängste bzw. viele Sorgen haben, die jedoch in der Beantwortung des DSM-Kriteriums keine klare Bedeutung erhält. Sie kann höchstens zur höheren Sicherheit bei dem Interviewer führen, wenn sie kongruent mit den nachfolgenden Fragen beantwortet wurde. In diesen wird bei der Sozialphobie danach gefragt, inwiefern die Patienten ängstlich oder nervös in sozialen Situationen reagie-

ren oder diese vermeiden. Die Frage nach der *Ängstlichkeit oder Nervosität* ist dabei so weit formuliert, dass viele positive Antworten erfolgen, mehr als auf die Eingangsfrage für die Störung. Auf der anderen Seite lautet das Kriterium des DSM-IV: „Eine *dauerhafte und übertriebene Angst* vor einer oder mehreren sozialen oder Leistungssituationen...". Der Interviewer muss sich also entscheiden, wie er die positiven Antworten auf die situativen Fragen bewertet und ob dies mit dem „schärfer formulierten" DSM-Kriterium übereinstimmt. Diese Frage mag so weit formuliert sein, um möglichst viele subklinische Auffälligkeiten in der epidemiologischen Studie erfassen zu können, für die Klassifikation einer Störung nach DSM-IV ist sie jedoch ungeeignet.

Bei der GAS liegen die Bedingungen ähnlich. Zudem wird verlangt, dass die Antworten auf die verschiedenen befragten Sorgenbereiche der GAS nur abweichend von „0" bewertet werden, wenn die Sorgen nicht im Zusammenhang mit einer gleichzeitig bestehenden Achse I- Störung zu sehen sind. Dies kann zu dem Zeitpunkt, zu dem die GAS erfragt wird, jedoch noch gar nicht beurteilt werden.
Bei der Zwangsstörung wird ein ebenso zweistufiges Vorgehen laut Interviewmanual notwendig wie bei der GAS oder der Sozialphobie. Dennoch liegt die Retest-Reliabilität im ausgezeichneten Bereich. Dies kann daran liegen, dass zwar die Eingangsfrage für die Zwangstörung relativ offen formuliert ist, dann aber die Ergänzungsfragen zu der Art der Zwänge sehr präzise formuliert sind, so dass dabei sofort erkannt werden kann, dass die befragten Verhaltensweisen außerhalb der Norm liegen, was bei der Sozialphobie oder der GAS nicht der Fall ist.

Bei der Dysthymen Störung sind die Patienten oft überfordert, die vielen Fragen nach zeitlichen Kriterien zu beantworten (zwei Jahre / die meiste Zeit des Tages depressiv / mehr als die Hälfte der Tage in zwei Jahren / mindestens zwei Monate, in denen die Stimmung normal war). Dies könnte eine Ursache für Inkonsistenzen zwischen 1. und 2. Interview sein. Der andere Grund ist wahrscheinlich in der schwierigen Abgrenzung der Dysthymen Störung zum Normalen oder zur Major Depression zu suchen,

wie es auch in der Validitätsstudie von Margraf et al. (1991) beschrieben wird und was sich auch in anderen Studien (Schneider et al., 1992; Di-Nardo et al., 1983; Hiller et al., 1997) in nicht sehr hohen Reliabilitätswerten äußert (durchschnittlicher $\kappa = .50$).

Bei der Somatoformen Schmerzstörung ist ohne Zusatzinformationen aus der Krankenakte oft schwer zu entscheiden, ob das Schmerzsymptom eine organische Grundlage hat oder nicht. Die Patienten sind oft überzeugt, besonders wenn die Behandlung beginnt, dass es eine organische Grundlage für ihre Schmerzen geben muss, die nur noch nicht entdeckt wurde. Zudem bestehen meistens kleinere organische Auffälligkeiten, die das Ausmaß der Beschwerden jedoch nicht erklären, so dass vom Interviewer eine schwierige Bewertung gefordert ist. Beim zweiten Interview kann es zur Subjektvarianz kommen, da sich die Patienten bereits mit psychischen Ursachen ihrer Beschwerden befasst haben und nun eine andere Auskunft geben.

Die Retest-Reliabilität variiert nicht nur störungsspezifisch, sondern auch in Abhängigkeit von der Komorbidität mit einer Cluster B- Persönlichkeitsstörung (antisozial, Borderline, histrionisch, narzisstisch). Diese Personengruppe ist schwerer gestört mit mehr komorbiden Störungen, sie ist beim ersten Interview deutlich depressiver als beim zweiten, und die Interviews dauern länger als bei Non-Cluster-B-Patienten. Hiermit wurde eine Teilstichprobe gefunden, die durch ihre Wechselhaftigkeit und ihr unsicheres Selbstbild schwieriger zu diagnostizieren ist und nur in 50 % übereinstimmend von beiden Interviewern eingeschätzt wurde. Diese Teilstichprobe machte ein Drittel der Gesamtstichprobe aus, so dass hierdurch die Reliabilitätsergebnisse zu einem erheblichen Teil beeinflusst sind.
Bei Patienten mit einer höheren Anzahl an komorbiden Achse-I-Störungen kam es genauso zu schlechteren Übereinstimmungen, wobei sich diese Patientengruppe mit der Cluster-B- Persönlichkeitsstörungsgruppe stark überschneidet.

Den gleichen Effekt übt die Dauer des Interviews aus. Bei länger andauernden Interviews wurden schlechtere Übereinstimmungen der Diagnosen erreicht. Die beiden letzten Punkte (Anzahl der komorbiden Störungen und Dauer des Interviews) erfassen den gleichen Sachverhalt, da, beim Vorliegen vieler Störungen, deutlich mehr Fragen gestellt werden müssen. Die Auswirkung auf die Übereinstimmung kann nun in zweifacher Weise erfolgen, zum einen dadurch, dass die Konzentration oder Motivation der Patienten oder die der Interviewer im Laufe des langen Interviews nachlässt und zum anderen dadurch, dass bei schwerer gestörten Patienten Informationen schwerer zu erhalten sind.

Einen weiteren Einfluss auf die Retest-Reliabilität hat die Erfahrung der Anwendung des F-DIPS insofern, als die Interviewer, die bereits durch das Projekt „Psychische Gesundheit junger Frauen“ Erfahrung gesammelt hatten, deutlich schlechtere Übereinstimmungen erzielten als die Interviewer, die direkt an einer Patientenstichprobe ihre ersten F-DIPS Erfahrungen machen konnten. Dies hängt vermutlich zum einen damit zusammen, dass die vermeintlich erfahrenen Projekt-Interviewer mit weniger gestörten Probandinnen Erfahrung gesammelt hatten und dadurch kleinere Auffälligkeiten überbewertet haben. Zum anderen lag bei ihnen die Schulung bereits länger zurück, so dass sich bereits individuellere Fragegewohnheiten entwickelt haben könnten, die längere Zeit nicht mehr korrigiert worden waren. Dies wirkte sich besonders bei zwei erfahrenen F-DIPS-Interviewern untereinander aus, die sowohl bei Angststörungen (κ = .33) als auch bei somatoformen Störungen (κ = .39) und affektiven Störungen (κ = .55) zu extrem geringeren Übereinstimmungen als alle anderen Interviewern kamen.

Der Einfluss der klinisch-psychiatrischen Erfahrung auf die Reliabilität konnte aufgrund der geringen Fallzahlen für die Paarung „zwei klinisch erfahrene Interviewer“ nicht bewertet werden. Bei Angst- und Affektiven Störungen sind die Übereinstimmungen jedoch ausgezeichnet, bei den klinisch unerfahrenen Interviewern nur gut. Ob sich die klinische Erfahrung so auf alle Störungsbereiche auswirkt und auch bei größeren Fall-

zahlen konstant bleibt, bleibt leider offen. Aufgrund anderer Studien (z.B. Hiller, 1997; Schneider, 1992) mit klinisch erfahreneren Interviewern und dennoch nicht entscheidend besseren Übereinstimmungen und aufgrund der Studie von Ventura et al. (1998), die keine Unterschiede nach einem SCID- Training der Interraterreliabilität zwischen klinisch erfahrenen und neu angelernten Interviewern fanden, kann diese Hypothese jedoch als unwahrscheinlich angesehen werden.

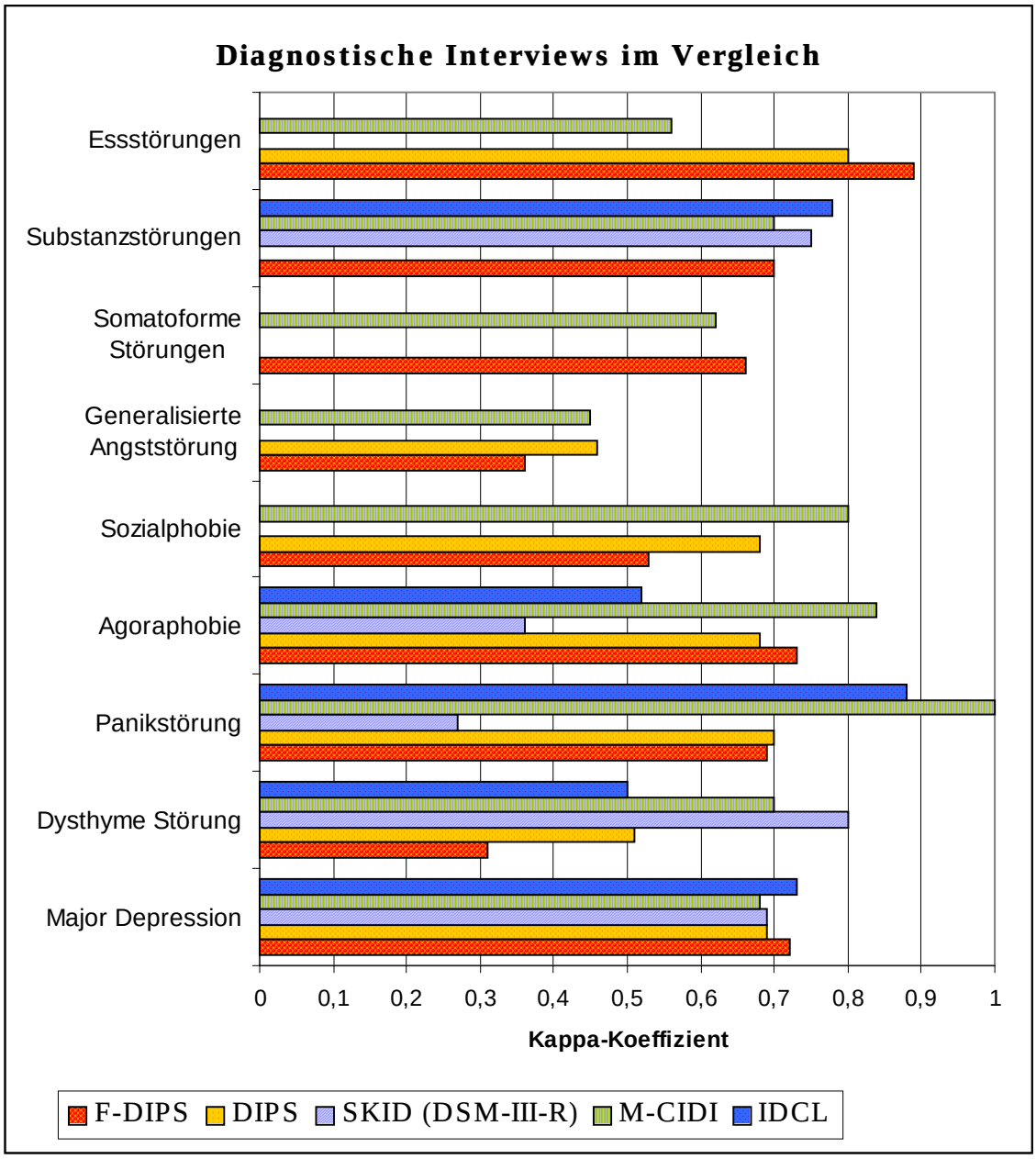
Die geringste Übereinstimmung zwischen klinisch erfahrenen Interviewern und klinisch unerfahrenen Interviewern macht sich erneut bei den Angststörungen mit $\kappa = .50$, aber auch bei den somatoformen Störungen mit $\kappa = .56$ fest. Bei der genaueren Fehleranalyse zeigt sich auch hier die Relevanz der Beobachtungsvarianz (s.o.). Da klinisch unerfahrene Interviewer weniger Störungen gesehen haben, bewerten sie kleinere Auffälligkeiten bei Angststörungen höher als klinisch Erfahrene. Bei somatoformen Störungen neigen sie dazu, falsch negativ zu entscheiden, wenn Patienten betonen, dass ihre Beschwerden wahrscheinlich organisch bedingt seien. Die Frage „Warum sind Sie dann hier in der psychosomatischen Klinik in Behandlung?“ ist nicht als Frage im F-DIPS vorgesehen und wird dann meist nicht gestellt.

Ein fast signifikanter Befund ergab sich aus dem zeitlichen Abstand der beiden Interviews voneinander. Hierfür könnte zum einen der Erinnerungseffekt verantwortlich sein, so dass mehr Patienten, die nach einer Woche bereits ihr zweites Interview hatten, noch die Fragen und ihre Angaben aus dem ersten Interview erinnern konnten. Zum anderen könnte der Patient bei sehr viel späteren Befragungszeiträumen nach drei oder vier Wochen eine andere Einstellung zu seinen Symptomen bekommen haben, so dass er ihnen nicht mehr die Wertigkeit gibt (Subjektvarianz) wie im ersten Interview.
Die mit dem BDI zu beiden Interviewzeitpunkten gemessene Depressivität, das globale Funktionsniveau des Patienten und das Ausmaß an Vertrauen zum Interviewer trennten die Patienten mit und ohne übereinstimmende

Diagnose nicht signifikant voneinander. Allerdings veränderte sich die Depressivität im BDI hoch signifikant von einem Messzeitpunkt zum nächsten in beiden Teilstichproben. Geht man davon aus, dass das aktuelle Maß an Depressivität einen Einfluss auf die Bewertung der Realität hat, könnte dieses Ergebnis bedeuten, dass generell auch die Auskunft bzgl. der Symptome oder die Bewertungen der Symptome im F-DIPS verändert sein können.

Auch das Vertrauen zum Interviewer änderte sich zwischen erstem und zweitem Interview positiv. Dieses Ergebnis ist unabhängig davon, ob die Patienten zur Gruppe mit übereinstimmenden Diagnosen oder zu der ohne übereinstimmende Diagnosen gehören und unabhängig von der Person des Interviewers, so dass davon auszugehen ist, dass sich die Patienten allgemein nach einem Interview an die Art der Vorgehensweise gewöhnt haben und sich deshalb im zweiten Interview sicherer fühlen.

Das Ausmaß an Sicherheit, mit der eine Diagnose vergeben wurde, spielt keine Rolle für die Übereinstimmung in den Diagnosen. Bei der Substanzabhängigkeit wurde von Interviewern mit einem sehr sicheren Gefühl sogar eine deutlich geringere Retest-Reliabilität erreicht. Da die Interviewer im Durchschnitt mit 82 %iger Sicherheit eine F-DIPS-Diagnose stellen und sich mehr als doppelt so häufig wie erfahrene Kliniker (in einer klinisch gestellten Diagnose) sehr sicher mit ihrer Diagnose sind, kann davon ausgegangen werden, dass sich ein Gefühl der Sicherheit allein durch die Verwendung des strukturierten Interviews aufbaut. Evtl. könnte auch die relative Unerfahrenheit der Interviewer eine Sicherheit dadurch entstehen lassen, dass sie nicht so viele Verläufe kennen und nicht so viele differenzialdiagnostische Erwägungen mit einbeziehen. Damit würde das Vertrauen in das F-DIPS und die damit gestellten Diagnosen eine Scheinsicherheit erzeugen.
Insgesamt zeigt sich das F-DIPS im Vergleich zu anderen strukturierten bzw. standardisierten Interviews bzw. Checklisten ähnlich reliabel.



Die in der Grafik „Diagnostische Interviews im Vergleich“ sehr geringen Kappa-Koeffizienten des F-DIPS bei der Dysthymen Störung und der Generalisierten Angststörung sind durch die kleinen Basisraten dieser Störungsbilder bedingt. Dennoch wurden zur Einheitlichkeit ausschließlich Kappa-Werte verwandt. Die Ergebnisse für das DIPS (1991) für die Somatoformen Störungen wurden aufgrund einer noch geringeren Basisrate nicht in die Grafik integriert, da dies wenig aussagekräftig gewesen wäre.

Während es bei den verglichenen Instrumenten allgemein eine homogen hohe Übereinstimmung bei den *Substanzabhängigkeiten* und bei der *Major Depression* gibt und eine homogen niedrige Reliabilität bei der Generalisierten Angststörung, sind die verschiedenen Instrumente hinsichtlich ihrer Reliabilität bei der *Dysthymen Störung* und bei den anderen *Angststörungen* heterogener: Hier gibt es Schwankungen in den Kappa-Werten von .50 bis .80 bei der Dysthymen Störung, von .27 bis .88 bei der Panikstörung und von .36 bis .85 bei der Agoraphobie oder von .57 bis .97 bei den phobischen Störungen. Bei den Angststörungen sind besonders beim M-CIDI, also einem standardisierten Interview, höhere Übereinstimmungen zu finden, was wahrscheinlich mit dem o.g. höheren Anteil an Beobachtungsvarianz bei einigen Angststörungen zusammenhängt.
Zur Beantwortung der Frage, weshalb das F-DIPS als Nachfolgeversion des DIPS schlechtere Übereinstimmungen (κ = .64 vs. κ = .71) hinsichtlich der Angststörungen zeigt, kann eine genauere Analyse der Veränderungen zwischen DIPS und F-DIPS beitragen. Bei der Eingangsfrage zur Panikstörung (Gab es im letzten Jahr Zeiten, in denen Sie ganz plötzlich einen Ansturm von intensiver Angst, Furcht ... oder das Gefühl eines drohenden Unglücks spürten?) gibt es nur den Unterschied, dass im F-DIPS die Frage noch länger ist als im DIPS durch den Zusatz „oder intensivem Unbehagen“. Allerdings bedient sich das F-DIPS auch im weiteren Verlauf einer komplizierteren Sprache mit vielen Wiederholungen. Z.B. heißt es in der nächsten Frage anstatt wie im DIPS „In welchen Situationen hatten Sie diese Gefühle?“, „Wann hatten Sie zum letzten Mal einen derartigen Ansturm von intensiver Angst, Furcht oder intensivem Unbehagen?“ und dann weiterhin: „Erlebten Sie in den vergangenen 7 Tage einen plötzlichen Ansturm von intensiver Angst, Furcht oder intensivem Unbehagen oder das Gefühl eines drohenden Unglücks?“. Dieses mehrfach nach dem gleichen Sachverhalt Fragen, könnte zur Verwirrung von Patient und Interviewer beitragen und ist im DIPS durch prägnantere Fragen besser gelöst.

Der Hauptgrund für die schlechteren Übereinstimmungswerte bei Angststörungen des F-DIPS sind jedoch die deutlich geringeren Prävalenzraten für eine Angststörung in der heterogener gestörten F-DIPS-Stichprobe im

Vergleich zur Stichprobe des DIPS, die zu über der Hälfte aus Patienten mit Angststörungen und aus einem Viertel „Patienten“ ohne Störung bestand, also nicht so schwer gestört war. Außerdem liegt bei den für die Reliabilitätsüberprüfung des F-DIPS in der Hauptsache untersuchten psychosomatischen Patienten eine diffusere körperliche Symptomatik vor, die dann unterschiedlich als Angst- oder als somatoforme Symptomatik gewertet werden kann und somit schwer zu beurteilen ist.
Bzgl. der Diagnose-Oberklassen Affektive Störungen, Somatoforme Störungen und Essstörungen ist das F-DIPS geringfügig reliabler als sein Vorgänger DIPS.

Die Validität des F-DIPS wurde wie die des Vorgängermodells DIPS (Margraf et al., 1991) anhand von störungsspezifischen Fragebögen (BDI, BAI, Whiteley-Index, SCL-90-R), überprüft und darüber hinaus mit Hilfe der Entlassungsdiagnose der Behandler und anhand von Patienteneinschätzungen der Hauptprobleme ermittelt. Es zeigte sich wie auch beim DIPS eine insgesamt gute Validität, wobei *Angststörungen, affektive Störungen* und *somatoforme Störungen* zwar sehr gut mit Fragebogenergebnissen übereinstimmen, jedoch weniger gut mit der Einschätzung der Behandler in Einklang zu bringen sind. Bei *Substanzabhängigkeit und –missbrauch* und bei den *Essstörungen* stimmen die F-DIPS-Diagnosen besser (gut bis ausgezeichnet) mit den Entlassungsdiagnosen überein.
Als interessantes (Neben-) Ergebnis zeigte sich der Zusammenhang von Entlassungsdiagnose mit der Patienteneinschätzung des Hauptproblems bei den Angststörungen. Dieser erbrachte mit .70 den höchsten κ-Koeffizienten, sogar etwas höher als zwischen beiden F-DIPS-Diagnosen (κ = .64). Damit zeigt sich, dass die Patienten, wenn eine Angststörung in der Entlassungsdiagnose steht, diese auch als ihr Hauptproblem benennen, während eine Angststörung in der F-DIPS Diagnose nur mäßig mit dem zusammenhängt, was die Patienten als ihr Hauptproblem darstellen (κ = .54).

Insgesamt stimmt das F-DIPS-Ergebnis mit der Entlassungsdiagnose von $\kappa = .43$ (Affektive Störungen) bis $\kappa = .88$ (Essstörungen) überein.
Vergleichbar sind diese Werte mit denen der Studie von Rosenman et al. (1997), in der ebenfalls Interview-Diagnosen mit klinischen Diagnosen verglichen wurden. Es besteht jedoch ein deutlich höherer Gesamt- κ – Koeffizient zwischen F-DIPS und Psychiater/Psychologe über die verschiedenen Störungsbereiche (κ =.59) als bei der Rosenman-Untersuchung zwischen dem CIDI in der computerisierten Version und einem Psychiater (κ =.23) und auch ein höherer als in der Untersuchung von Peters & Andrews (1995), die zwischen der computerisierten Version des CIDI und dem Kliniker einen Gesamt-Kappa-Koeffizienten von .40 ermittelten. Die Validität des SCID (Structured Clinical Interview for DSM-IV), gemessen in einer Studie von Kranzler (1996), der die SCID-Diagnose ebenfalls mit einer klinischen Diagnose verglich, entsprach in etwa der des F-DIPS (gut bei Substanzabhängigkeiten, mittel für Major Depression und - schlechter als beim F-DIPS - ungenügend für Angststörungen).
Dies zeigt erneut die Überlegenheit eines strukturierten Interviews, insbesondere des F-DIPS, im Vergleich zu einem standardisierten Interview hinsichtlich seiner Validität.

9.2 Diskussion der Untersuchungsmethoden

Bei Planung der Untersuchung wurden zunächst andere Untersuchungsbedingungen angenommen, z.B. dass die Interviews ausschließlich durch klinisch oder F-DIPS-erfahrene Interviewer durchgeführt werden mit Probandinnen aus dem Dresdner Projekt „Prädiktoren der psychischen Gesundheit junger Frauen“. Erst als sich dies als nicht möglich erwies, wurden auch Psychologie-Studenten als Interviewer angeworben und Klinikpatienten für die Stichprobe ausgewählt.
Im nachhinein beurteilt, war es von Vorteil, dass sich die Stichprobe aus Klinik-Patienten zusammensetzte, besonders, da die Befürchtung, dass die Behandlung zu Stimmungsveränderungen und Veränderungen der Symptomatik führen würde, sich nur in einem das Ergebnis nicht beein-

flussenden Ausmaß bewahrheitet hat. So bestanden klinisch realistische Bedingungen sowohl für die Durchführung der Interviews als auch für das Ausmaß an Störungen mit einer Komorbidität von Persönlichkeitsstörungen. Durch die Vielfalt an unterschiedlichen Patienten (ambulant, Rehabilitationsklinik, Universitätsklinik) ließ sich auch zeigen, dass das F-DIPS überall angewandt werden kann, jedoch unterschiedlich zeitintensiv ist und bei schwer gestörten Patienten zur Belastung werden kann und daher unter Umständen nicht an einem einzigen Termin durchgeführt werden sollte.

Trotz der relativ breit variierenden Stichprobe gab es geringe Basisraten unter 10 % mit allen dazugehörigen statistischen Schwierigkeiten bei der Panikstörung ohne Agoraphobie, der Agoraphobie ohne Panikstörung, der Spezifischen Phobie, der Generalisierten Angststörung, bei der Zwangsstörung, der Posttraumatischen Belastungsstörung, der Dysthymen Störung, der Bipolaren Störungen, der Gemischten Angst-Depressionsstörung, der Hypochondrie, der Somatisierungsstörung, der Konversionsstörung, bei allen Substanzabhängigkeiten und beiden Essstörungen. In der Untersuchung wurden keine Gewichtungen von primärer und sekundärer Diagnose vorgenommen, da die Basisraten sonst noch geringer geworden wären und kein erheblicher Informationsgewinn zu erwarten gewesen wäre.

9.3 Ausblick

Ein strukturiertes Interview wie das F-DIPS hat Vorteile in der systematischen Erfassung aller für eine Psychotherapie relevanten Störungen. Das korrekte Vorgehen im Hinblick auf die Kriterien des DSM-IV wird mit Hilfe des Interviews erleichtert. Es wurden zwar viele Interviewleitfäden entwickelt, um die Klassifikation nach DSM-IV oder ICD-10 zu verbessern – ob die Instrumente dies wirklich tun, wurde bisher allerdings lediglich umfassend für das CIDI (Retest-Reliabilität und Validität) (Wittchen et al., 1998; Cooper et al., 1998; Rosenman et al., 1997; Peters & Andrews,

1995), die ICDL (Retest-Reliabilität) (Hiller et al., 1997) und nun für das F-DIPS überprüft.
Das hier untersuchte F-DIPS kann in den Bereichen Affektive Störungen, Somatoforme Störungen, Substanzmissbrauch und –abhängigkeit und Essstörungen als gut reliables und für alle Bereiche als valides Instrument angesehen werden. Es ist jedoch relativ zeitaufwendig (durchschnittliche Durchführungsdauer 97 Minuten ± 41 Minuten), was sich besonders bei den schwerer gestörten Patienten der Universitätsklinik zeigte, wo die Interviews durchschnittlich noch 10 Minuten länger als bei Rehaklinikpatienten dauerten. Dabei wurde deutlich, dass gerade bei den länger andauernden Interviews eine geringere Übereinstimmung in den Diagnosen zu finden ist.

Möglichkeiten, um das Interview etwas zu straffen, liegen in der Reihenfolge der Störungen, die abgefragt werden und in den Formulierungen, die knapper und eindeutiger werden könnten (s.o.). Während das DIPS noch als Interview mit Schwerpunkt auf der Diagnostik von Angststörungen entwickelt worden ist, ist das F-DIPS gleichmäßiger ausdifferenziert worden auf alle Störungen. Insofern spricht nichts mehr dafür, gerade die Angststörungen als ersten Bereich abzufragen. Es wäre sinnvoll, die Affektiven Störungen vorweg zu stellen, da diese die am häufigsten vorkommende psychische Störung darstellen und evtl. andere Störungen, wie z.B. die Generalisierte Angststörung, ausschließen können, die dann nicht mehr vollständig (außer Zeitkriterien) befragt werden müsste.
Sprachlich könnten einige Formulierungen überarbeitet werden, da sie für Patienten sonst oft während des Interviews vereinfacht oder übersetzt werden müssen. Beispielsweise bei der Agoraphobie wird gefragt: „Gab es in den vergangenen 7 Tagen Situationen, wie z.B. Kaufhäuser, Autofahren, Menschenmengen oder enge, geschlossene Räume, in denen Sie Angst hatten oder die Sie vermieden, weil Sie sich dort sehr ängstlich fühlen könnten?“. Wie viel leichter wäre die Frage, wenn sie lautete: „Hatten Sie in den letzten 7 Tagen Angst, alleine das Haus zu verlassen, mit dem Bus zu fahren, in Kaufhäuser zu gehen oder in einer Schlange anzustehen, oder haben Sie die Situationen aus Angst vermieden?“.

Ein weiterer Weg, um dem Interviewer etwas mehr Sicherheit zu geben, in welche Richtung die Beschwerden des Patienten gehen, besonders, wenn der Patient ausnahmslos Fragen bejaht, wäre eine Eingangsfrage nach dem Hauptproblem des Patienten wie in einem klinischen Interview. Diese Frage wird im F-DIPS erst am Schluss gestellt, wahrscheinlich, um den Interviewer nicht durch Hypothesenbildung zu leiten und dadurch evtl. einzelne Fragen zu vernachlässigen. Tatsächlich führt der Verzicht einer allgemeinen Eingangsfrage aber dazu, gerade bei den ersten Fragen zu viel Zeit zu verlieren durch zusätzliches Nachfragen, um klarere Antworten und eine Richtung zu bekommen. Zudem bleibt die Atmosphäre des Gesprächs gespannt, wenn der Patient zu Beginn des Interviews unter Druck steht, aber über sein Hauptproblem vielleicht noch lange nicht sprechen darf.

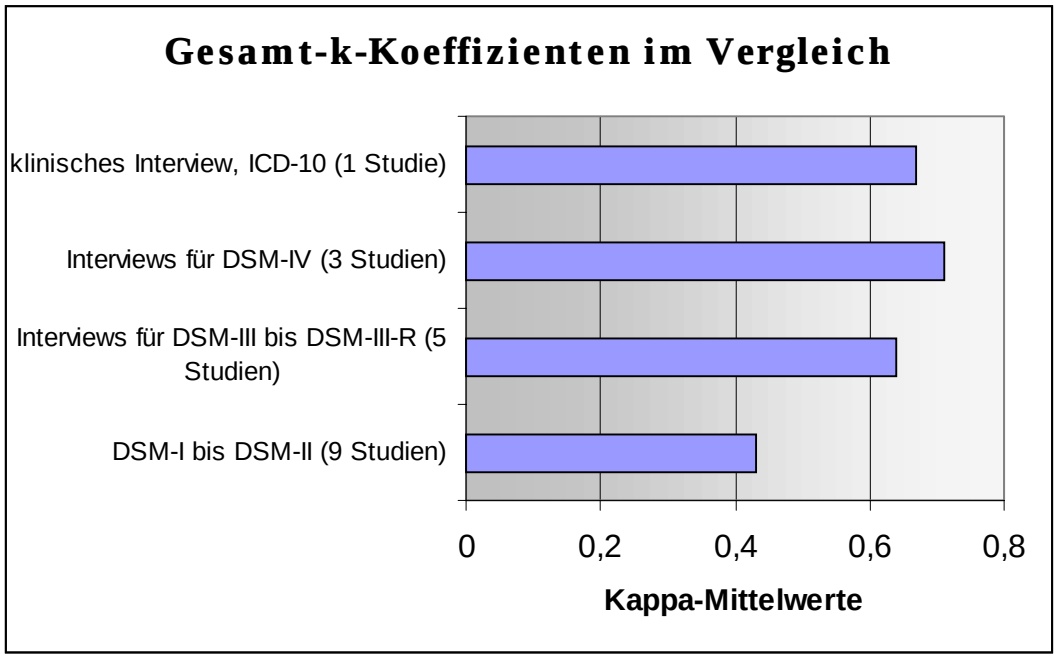
Methodisch interessant wäre es zur Klärung der Frage, wie sinnvoll und aussagekräftig Retest-Reliabilitätsstudien überhaupt sind, den Einfluss der die Reliabilität beeinflussenden Faktoren mit Hilfe eines Lisrel-Modelles zu berechnen. Darüber könnte mehr Klarheit entstehen, wie viel Varianz sich tatsächlich allein störungsspezifisch erklären lässt und wie viel von den anderen Faktoren wie dem Ausmaß der Komorbidität, der Cluster-B Persönlichkeitsstörungen, der Dauer des Interviews, der Interviewer-Erfahrung oder dem zeitlichen Abstand zwischen den Interviews abhängig ist. Leider hätte hierfür die Stichprobe noch etwas größer als 191 Patienten sein müssen, weshalb in diesem Rahmen darauf verzichtet wurde.

In den meisten Studien werden zwar Kappa-Koeffizienten angegeben, es fehlen jedoch Angaben von Basisraten als auch solche zu den genaueren Untersuchungsbedingungen oder sie werden bezüglich ihres Einflusses auf die Untersuchungsergebnisse nicht oder kaum beachtet.

In Anbetracht des fast signifikanten Befundes, dass die Übereinstimmung der diagnostischen Urteile vom zeitlichen Abstand der beiden Interviews voneinander abhängen könnte (mit besseren Übereinstimmungen bei kürzerem Abstand) könnte auch diskutiert werden, ob es sinnvoll ist, eine

zweimalige Testvorgabe innerhalb einer Woche durchzuführen, wie es in vielen Untersuchungen geschieht, oder ob zur Vergleichbarkeit alle Untersuchungen den Wochenabstand wählen sollten. Auf der anderen Seite macht es Untersuchungen fragwürdig, wenn Artefakte wie Erinnerungseffekte genutzt werden, um die Reliabilität eines Tests künstlich zu erhöhen, anstatt die evtl. niedrigere Reliabilität konstruktiv zu diskutieren.
Auch würde es einen wirklichen Vergleich von Instrumenten erleichtern, wenn vergleichbare Methoden z.B. für die Validierung angewendet werden würden. Hier bestehen jedoch noch große Unsicherheiten, welche Kriterien sich eignen, um die durch ein strukturiertes Interview gewonnenen Diagnosen zu messen. Dies wirft wiederum insgesamt die Frage auf, was als Störung anerkannt wird: die Störung, die behandelt wird, oder die behandlungsbedürftig wäre, oder die, die sich auch im Fragebogen abbildet (vgl. Margraf et al., 1991), oder die, die ein Psychiater diagnostiziert (vgl. Peters & Andrews, 1995; Rosenman et al, 1997; Kranzler et al., 1996), oder die, die mit einem vergleichbaren diagnostischen Interview (vgl. Cooper et al., 1998; Wittchen, 1994b; Eaton et al., 2000) erfasst wird. Diese Ambivalenz erklärt vermutlich, wieso es so wenige Studien zur Überprüfung der Validität von diagnostischen Interviews gibt.

In den vergangenen Jahrzehnten wurde dem Vorwurf, psychiatrische Diagnosen seien wenig reliabel und hätten deshalb nur einen begrenzten Wert für Behandlungen, Prognosen, Ätiologieverständnis oder für phänomenologische Beschreibungen (Spitzer et al., 1975), durch eine immer stärkere Ausdifferenzierung von Klassifikationssystemen und die Entwicklung von diagnostischen Interviews begegnet. In der folgenden Grafik sind für mehrere o.g. Retest-Reliabilitätsstudien, die dasselbe Klassifikationssystem überprüften, Gesamt-Kappa-Werte über alle Störungen und alle Instrumente gemittelt worden. Wie die Grafik zeigt, gab es einen sehr hohen Qualitätssprung der diagnostischen Instrumente von DSM-II zu DSM-III Anfang der 80er Jahre.



Dieser Entwicklungssprung konnte durch die Entwicklung des DSM-IV nicht in der gleichen Stärke wiederholt, aber doch ausgebaut werden. Aufgrund dieser Erfolge ist es wahrscheinlich, dass auch in den nächsten Jahren die Forschung in Richtung der Fortentwicklung von syndromatologischer Diagnostik und diagnostischen Interviews geht. Auch in der klinischen Praxis setzen sich diagnostische Interviews vermutlich noch stärker durch, um Ansprüchen der Qualitätssicherung zu genügen. Jedoch zeigt die Studie von Rosenman et al. (1997), die als einzelne Studie in der Grafik aufgeführt ist, dass auch die psychiatrische Diagnostik ohne Verwendung eines Interviewleitfadens anscheinend in der Lage sein kann, übereinstimmende Diagnosen zu erbringen. Auch solche zu strukturierten diagnostischen Interviews alternative Vorgehensweisen zur Diagnosesicherung sollten von der zukünftigen Forschung zur Güte von Diagnostik nicht vernachlässigt werden.

10 Zusammenfassung

In der Diagnostik psychischer Störungen werden kategoriale Systeme mit syndromatologischer Klassifikation verwendet. Diese Entwicklung entstand vor allem aufgrund von Reliabilitätsproblemen bei der klinisch-psychiatrischen Klassifikation, aber auch durch die Pharmakotherapie, die eher syndromorientiert als nosologisch vorgeht (Möller, 1998). In der Forschung ist besonders das DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, APA, 1994) aufgrund seiner guten Operationalisierung der Störungskriterien weit verbreitet.

Um den sorgfältigen Gebrauch des Klassifikationssystems zu gewährleisten, wurden eine Reihe von diagnostischen Interviewleitfäden entwickelt (halb strukturiert, voll strukturiert oder standardisiert).

In der vorliegenden Untersuchung geht es um das (voll) strukturierte Interview F-DIPS (Diagnostisches Interview bei psychischen Störungen- Forschungsversion für DSM-IV von Margraf et al., 1996), das bisher vor allem innerhalb der Studie zu „Prädiktoren psychischer Gesundheit/Krankheit bei jungen Frauen“ eingesetzt wurde. Mit seiner Hilfe ist es möglich, eine Auswahl an Achse I- Störungen (Angststörungen, Affektive Störungen, Somatoforme Störungen, Substanzmissbrauch und –abhängigkeit sowie Essstörungen) zu diagnostizieren, eine soziale und medizinische Anamnese zu erheben und eine Einschätzung des allgemeinen Funktionsniveaus (GAF, Achse V des DSM-IV) vorzunehmen.

In dieser Untersuchung wurden 191 Patientinnen und Patienten mit dem F-DIPS von zwei unabhängigen Interviewern im Abstand von etwa zwei Wochen zu ihren Symptomen im letzten Jahr befragt. Hierüber wurde die Retest-Reliabilität bestimmt. Außerdem füllten die Patienten störungsspezifische Fragebögen aus. Mit deren Hilfe sowie mit der Entlassungsdiagnose und der Selbstauskunft der Patienten zu ihrem Hauptproblem wurde die Validierung des F-DIPS vorgenommen.

Die Interviewerinnen und Interviewer setzten sich bis auf zwei Ausnahmen aus Psychologie-Studenten zusammen, die ein spezielles Training in der Anwendung des F-DIPS erhalten hatten.

Insgesamt zeigte sich eine mit anderen Instrumenten vergleichbare Retest-Reliabilität mit Stärken in den Essstörungen (Kappa 0,89, Yules Y 0,94) und Substanzmissbrauch und –abhängigkeit (Kappa 0,70, Yules Y 0,85). Eine ausgezeichnete Retest-Reliabilität ergab sich auch bei der Zwangsstörung (κ = .67, Y= .90), der Major Depression rezidivierend (κ = .80), bei bipolaren Störungen (κ = .83), und der Hypochondrie (κ = .75).

Gute Übereinstimmungen mit einem Kappa-Koeffizienten über .65 zeigten sich bei den meisten Angststörungen mit Ausnahme der Panikstörung ohne Agoraphobie (κ = .46, Y= .69), der Sozialphobie (κ = .53, Y= .61) und der Generalisierten Angststörung (κ = .36, Y= .61) und bei den somatoformen Störungen mit Ausnahme der somatoformen Schmerzstörung (κ = .59, Y= .67). Auch die Dysthyme Störung weist eine geringe Reliabilität (κ = .31, Y= .51) auf, sogar noch geringer, als die anderer Retest-Reliabilitätsstudien.

Eine Erklärung für die schlechteren Reliabilitätswerte des F-DIPS bezüglich der Angststörungen im Vergleich mit dem Vorgängerinterview für DSM-III-R DIPS ergibt sich aus den komplizierteren Frageformulierungen des DIPS, aber vor allem mit den niedrigeren Prävalenzraten für Angststörungen in der heterogeneren, schwerer gestörten F-DIPS-Stichprobe mit diffusen körperlichen Symptomen liegen. Hinsichtlich aller, auch komorbider Störungen wurden 61 % der Patienten von beiden Interviewern völlig übereinstimmend bewertet. Insgesamt wurden in beiden Interviews 1,7 Diagnosen vergeben. Werden die Interviewer in der Güte ihrer Diagnosestellungen überprüft, zeigte sich, dass die vier erfahrenen „Dresdner-Projekt-Interviewer“ die schlechtesten Übereinstimmungen erzielten, was wahrscheinlich zum einen mit der bei ihnen länger zurückliegenden Schulung, aber vielleicht auch mit einer durch die Vielzahl der geführten Interviews erworbenen Scheinsicherheit zusammenhängt, die sich jedoch auf eine meist gesunde Stichprobe der Allgemeinbevölkerung bezog, so dass die Einschätzung klinischer Störungsbilder mit multiplen Auffälligkeiten Schwierigkeiten bereiteten. Das Gefühl von Sicherheit beim Stellen einer Diagnose hat keinen Einfluss auf die tatsächliche Übereinstimmung der Diagnosen.

Unter möglichen weiteren Faktoren zeigte sich ein Einfluss auf die Übereinstimmung in den Diagnosen, wenn ein Patient Hinweise auf eine Cluster B- Persönlichkeitsstörung (antisoziale, Borderline, histrionische, narzisstische Persönlichkeitsstörung) bot, wenn er eine Vielzahl von Diagnosen erhielt, die Interviews länger dauerten und wenn die Patienten selbst angaben, dass sie nur teilweise die gleichen Antworten in beiden Interviews gegeben hatten. Einen beinahe signifikanten Zusammenhang gab es zwischen dem zeitlichen Abstand der beiden Interviews zueinander ($p=0,06$) und der Übereinstimmung in den Diagnosen, wobei es eine höhere Übereinstimmung gab, wenn die beiden Interviews innerhalb einer Woche geführt wurden.
Als Fehlerquellen bei den nicht-übereinstimmenden Interviews zeigten sich die Beobachtungsvarianz (Interviewer bewerten die geschilderte Problematik unterschiedlich) mit 42 % als häufigste Ursache, die Informationsvarianz (Patienten geben auf die gleiche Frage abweichende Antworten) mit 36 % als zweithäufigste. In 12 % der Fälle spielte die fehlerhafte Anwendung (Übertragungsfehler, keine vollständige Befragung, Sprungregeln missachten) des F-DIPS eine Rolle und in ca. 9 % waren F-DIPS oder DSM-IV unzureichend, so dass sie direkt für die Nichtübereinstimmung verantwortlich waren.

Das andere Ziel der Untersuchung betraf die Validität. Das F-DIPS erwies sich in den Störungsbereichen, die mit Fragebögen validiert wurden (Angststörungen, Depression, somatoforme Störungen) als sehr valide. Die Störungsbereiche Substanzmissbrauch und –abhängigkeit wurden ausschließlich durch die Entlassungsdiagnose validiert. Auch diese Störungen konnten mit dem F-DIPS sehr valide erfasst werden. Bei den Essstörungen lag der Kappa-Wert bei .88, bei den Substanzabhängigkeiten bei .66.
Mit der Entlassungsdiagnose verglichen, lagen die Angststörungen, Affektiven Störungen und Somatoformen Störungen deutlich niedriger in der Übereinstimmung (Kappa-Werte von .43 bis .49), jedoch immer noch höher als bei einem standardisierten diagnostischen Interview.

Insgesamt liegt mit dem F-DIPS eines der wenigen auf beide Hauptgütekriterien untersuchten Instrumente vor, das in einer Stichprobe aus psychosomatischen Patienten Stärken bei der Diagnostik von Substanzmissbrauch, Essstörungen und Affektiven Störungen aufwies. Bezüglich seiner Validität gehört es zu den besten der untersuchten diagnostischen Interviews.

Es eignet sich grundsätzlich auch für den Einsatz bei schwer gestörten Patienten, wobei sich eine problematische Teilstichprobe – die Cluster-B-Persönlichkeitsstörungen – ermitteln ließ, bei der die Reliabilität der Diagnose deutlich litt.

11 Literatur

Alpert, J.E., Uebelacker, L.A., McLean, N.E., Nierenberg, A.A., Pava, J.A., Worthington, J.J. 3rd, Tedlow, J.R., Rosenbaum, J.F. & Fava, M. (1997). Social phobia, avoidant personality disorder and atypical depression: co-occurrence and clinical implications. Psychological Medicine, 27, 627-633.

Amelang, M. & Zielinski, W. (1997). Psychologische Diagnostik und Intervention. (2. Aufl.) Berlin: Springer.

Arbeitskreis OPD (Hrsg.) (1998). Operationalisierte Psychodynamische Diagnostik. (2. Aufl.) Bern: Huber.

Asendorpf, J. & Wallbott, H.G. (1979). Maße der Beobachterübereinstimmung: Ein systematischer Vergleich. Zeitschrift für Sozialpsychologie, 10, 243-252.

Baumann, U. & Stieglitz, R.-D. (1994). Psychodiagnostik psychischer Störungen: Allgemeine Grundlagen. In R.-D. Stieglitz & U. Baumann (Hrsg.), Psychodiagnostik psychischer Störungen (S. 3-20). Stuttgart: Enke.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J.E. & Erbaugh, J.K. (1962). Reliability of psychiatric diagnoses: A studiy of consistency of clinical judgments and ratings. American Journal of Psychiatry, 119, 351-357.

Berger, M. (Hrsg.). (1999). Psychiatrie und Psychotherapie. München: Urban & Schwarzenberg.

Bortz, J. & Döring, N. (1995). Forschungsmethoden und Evaluation. (2. Aufl.) Berlin: Springer.

Brennan, R.L. & Prediger, D.J. (1981). Coeffizient κ: Some uses, misuses and alternatives. Educational and Psychological Measurement, 41, 687-699.

Brown, T.A., DiNardo, P.A. & Barlow, D.H. (1994). Anxiety Disorders Interview Schedule for DSM-IV (ADIS-IV). Albany, New York: Graywind.

Brugha, T.S., Bebbington, P.E., Jenkins, R., Meltzer, H., Taub, N.A., Janas, M., Vernon, J. (1999). Cross validation of a general population survey diagnostic interview: a comparison of CIS-R with SCAN ICD-10 diagnostic categories. Psychological Medicine, 29, 1029-1042.

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

Cooper, J.E., Kendell, R.E., Gurland, B.J. (1969). Cross-national study of mental disorders: Some results from the first comparative investigation. American Journal of Psychiatry, 125, 21-28.

Cooper, L., Peters, L. & Andrews, G. (1998). Validity of the Composite International Diagnostic Interview (CIDI) psychosis module in a psychiatric setting. Journal of Psychiatric Research, 32, 361-368.

Davison, G.C. & Neale, J.M. (1996). Klinische Psychologie. (4. Aufl.) Weinheim: Psychologie Verlags Union.

Derogatis, L.R. & Cleary, P.A. (1977). Confirmation of the dimensional structure of the SCL-90: a study in construct validation. Journal of Clinical Psychology, 33, 981-989.

Dilling, H., Mombour, W., Schmidt, M.H. (Hrsg.) (1991). Internationale Klassifikation psychischer Störungen. Weltgesundheitsorganisation. ICD-10 Kapitel V (F). Bern: Huber.

DiNardo, P.A., O'Brian, G.T., Barlow, D.H., Waddell, M.T. & Blanchard, E.B. (1983). Reliability of DSM-III-R anxiety disorder categories using a new structured interview. Archives of General Psychiatry, 40, 1070-1074.

Easton, C., Meza, E., Mager, D., Ulug, B., Kilic, C., Gogus, A., Babor, T.F. (1997). Test-retest reliability of the alcohol and drug use disorder sections of the schedules for clinical assessment in neuropsychiatry (SCAN). Drug Alcohol Depend, 25, 187-194.

Eaton, W.W., Neufeld, K., Chen, L.S., Cai, G. (2000). A comparison of self-report and clinical diagnostic interviews for depression: Diagnostic Interview Schedule and Schedules for Clinical Assessment in Neuropsychiatry in the Baltimore Epidemiologic Catchment Area follow-up. Archives of General Psychiatry, 57, 217-222.

Erdman, H.P., Klein, M.H., Greist, J.H., Skare, S.S., Husted, J.J., Robins, L.N., Helzer, J.E., Goldring, E., Hamburger, M. & Miller, J.P. (1992). A comparison of two computer-administered versions of the NIMH Diagnostc Interview Schedule. Journal of Psychiatric Research, 26, 85-95.

Eysenck, H.J. (1986). A critique of contemporary classification and diagnosis. In T. Millon & G.L. Klerman (Eds.), Contemporary directions in psychopathology: Toward the DSM-IV (pp. 73-88). New York: Guilford.

Fiedler, P. (1994). Persönlichkeitsstörungen. Weinheim: Psychologie Verlags Union.

Fleiss, J.L. (1981). Statistical Methods für Rates and Proportions, ed2. New York: Wiley.

Frances, A., Widiger, T. & Fyer, M.R. (1990). The influence of classifikation methods on comorbidity. In J.D. Maser & C.R.Cloninger (Eds.), Comorbidity of mood and anxiety disorders (pp. 41-59). Washington: American Psychiatric Press.

Franke, G.H. (1995). SCL-90-R. Die Symptomcheckliste von Derogatis. Göttingen: Beltz Test.

Fydrich, T. (1997). Diagnostik und Intervention in der Klinischen Psychologie. In M. Amelang & W. Zielinski, Psychologische Diagnostik und Intervention (S. 444-458) (2. Aufl.). Berlin: Springer.

Fydrich, T., Renneberg, B., Schmitz, B. & Wittchen, H.-U. (1997). Strukturiertes Klinisches Interview für DSM-IV. Achse II: Persönlichkeitsstörungen (SKID-II). Göttingen: Hogrefe.

Grove, W.M., Andreason, N.C., McDonald-Scott, P., Keller, M.B. & Shapiro, R.W. (1981). Reliability studies of psychiatric diagnosis. Archives of General Psychiatry, 38, 408-413.

Gülick-Bailer, M. van, Maurer, K. & Häfner, H. (Eds.). (1995). Schedules for Clinical Assesssment in Neuropsychiatry (SCAN). Bern: Huber.

Helmchen, H. (1975). Der Einfluß der statistischen Erfassung psychopathologischer Daten auf nosologische Konzepte in der Psychiatrie. In: K. Heinrich (Hrsg.), Zur Kritik der psychiatrischen Nosologie. Stuttgart: Schattauer.

Hesselbrock, M., Easton, C., Bucholz, K.K., Schuckit, M. & Hesselbrock, V. (1999). A validity study of the SSAGA- a comparison with the SCAN. Addiction, 94, 1361-1370.

Hiller, W., Zaudig, M. & Mombour, W. (1997). Internationale Diagnosen Checkliste (IDCL) für DSM-IV. Göttingen: Hogrefe.

Jacobson, N.S. & Revenstorf, D. (1988). Statistics for assessing the clinical significance of psychotherapy techniques: Issues, problems, and new developments. Behavioral Assessment, 10, 133-145.

Janca, A., Robins, L.N., Bucholz, K.K., Early, T.S. & Shayka, J.J. (1992). Comparison of Composite International Diagnostic Interview and clinical DSM-III-R criteria checklist diagnoses. Acta Psychiatrica Scandinavica, 85,, 440-443.

Janzarik, W. (1974). Themen und Tendenzen der deutschsprachigen Psychiatrie. Berlin: Springer.

Jaspers, K. (1973). Allgemeine Psychopathologie. (9. Aufl.) Berlin: Springer.

Kessler, R.C., McGonagle, K.A., Zhao, S., Nelson, C.B., Huges, M., Eshleman, S., Wittchen, H.-U. & Kendler, K.S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States : Results form the national comorbidity survey. Archives of General Psychiatry, 51, 8-19.

Kranzler, H.R., Kadden, R.M., Babor, T.F., Tennen, H. & Rounsaville, B.J. (1996). Validity of the SCID in substance abuse patients. Addiction, 91, 859-868.

Krauth, J. (1995). Testkonstruktion und Testtheorie. Weinheim: Psychologie Verlags Union.

Kreitman, N., Sainsbury, P., Morrisey, J. (1961). The reliability of psychiatric assessment: An analysis. Journal of Mental Science, 107, 887-891.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Margraf, J. (1996). Klassifikation psychischer Störungen. In J. Margraf (Hrsg.), Lehrbuch der Verhaltenstherapie. Band 1 (S. 83-101). Berlin: Springer.

Margraf, J. & Ehlers, A. (1998). Beck Angstinventar. Deutsche Version. Bern: Huber.

Margraf, J., Schneider, S. & Ehlers, A. (1991,1994). DIPS. Diagnostisches Interview bei psychischen Störungen. Berlin: Springer.

Margraf, J., Schneider, S., Soeder, U., Neumer, S. & Becker, E. (1996). F-DIPS. Diagnostisches Interview bei psychischen Störungen (Forschungsversion). Unveröffentlichtes Manuskript. TU Dresden.

Margraf, J., Schneider, S. & Spörkel, H. (1991). Therapiebezogene Diagnostik: Validität des Diagnostischen Interviews für psychische Störungen (DIPS). Verhaltenstherapie, 1, 110-119.

Margraf, J. & Schneider, S. (1996). Diagnostik psychischer Störungen mit strukturierten Interviews. In J. Margraf (Hrsg.), Lehrbuch der Verhaltenstherapie. Band 1. Berlin: Springer.

Martin, C.S., Pollock, N.K., Bukstein, O.G., Lynch, K.G. (2000). Interrater reliability of the SCID alcohol and substance use disorders section among adoloscents. Drug Alcohol Depend, 59, 173-176.

Möbius, P.J. (1892). Über die Einteilung der Krankheiten. Neurologische Betrachtungen. Centralbibliothek Nervenheilkunde Psychiatrie, 15, 289-301.

Möller, H.-J. (1998). Probleme der Klassifikation und Diagnostik. In H. Reinecker (Hrsg.), Lehrbuch der Klinischen Psychologie. Modelle psychischer Störungen (3-24). Göttingen: Hogrefe.

Mombour, W., Zaudig, M., Berger, P., Gutierrez, K., Berner, W., Berger, K., Cranach v., M., Giglhuber, O., Bose v., M. (1996). International Personality Disorder Examination (IPDE). Göttingen: Hogrefe.

Mundt, Ch. (1991). Endogenität von Psychosen – Anachronismus oder aktueller Wegweiser für die Pathogeneseforschung? Nervenarzt, 62, 3-15.

Nathan, P.E. (1994). DSM-IV: Empirical, accessible, not yet ideal. Journal of Clinical Psychology, 50, 103-110.

Peters, L. & Andrews, G. (1995). Procedural validity of the computerized version of the Composite International Diagnostic Interview (CIDI-Auto) in the anxiety disorders. Psychological Medicine, 25,1269-1280.

Richter, P. (1991). Zur Konstruktvalidität des Beck-Depressionsinventars bei der Erfassung depressiver Verläufe. Ein empirischer und methodologischer Beitrag. Regensburg: Roderer.

Richter, P., Werner, J. & Bastine, R. (1994). Psychometrische Eigenschaften des Beck-Depressionsinventars (BDI): Ein Überblick. Zeitschrift für Klinische Psychologie, 23, 3-19.

Rief, W., Greitemeyer, M. & Fichter, M. (1991). Die Symptom Check List SCL-90-R: Überprüfungen an 900 psychosomatischen Patienten. Diagnostica, 37, 58-65.

Rief, W., Hiller, W., Geissner, E. & Fichter, M.M. (1994). Hypochondrie: Erfassung und erste klinische Ergebnisse. Zeitschrift für Klinische Psychologie, 23, 34-42.

Robins, L.N., Helzer, J.E., Croughan, J. & Ratcliff, K.S. (1981). National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, an validity. Archives of General Psychiatry, 38, 381-389.

Rosenman, S.J., Korten, A.E., Levings, E.T. (1997). Computerised diagnosis in acute psychiatry: validity of CIDI-Auto against routine clinical diagnoses. Journal of Psychiatry Research, 31, 581-592.

Sandifer, M.G., Pettus, C, Quade, D. (1964). A study of psychiatric diagnosis. Journal of Nervous Mental Disorders, 139, 350-361.

Saß, H., Wittchen, H.-U. & Zaudig, M. (1996). Diagnostisches und Statistisches Manual Psychischer Störungen DSM-IV. Göttingen: Hogrefe.

Satlow, E. (1996). Die Interrater-Reliabilität im Diagnostischen Interview bei psychischen Störungen, Forschungsversion (F-DIPS) bei der Studie „Gesundheit in Dresden“. Unveröffentlichte Diplomarbeit. Berlin: Institut für Diagnostik und Intervention der FU Berlin.

Schmidt, H.O.& Fonda, C.P. (1956). The reliability of psychiatric diagnosis: A new look. Journal of Abnormal, Social Psychology, 52, 262-275.

Schneider, K. (1987). Klinische Psychopathologie. (13. Aufl.) Stuttgart: Thieme.

Schneider, S., Margraf, J., Spörkel, H., & Franzen, U. (1992). Therapiebezogene Diagnostik: Reliabilität des Diagnostischen Interviews bei psychischen Störungen (DIPS). Diagnostica, 38, 209-227.

Schulte, D. & Wittchen, H.-U. (1988). Wert und Nutzen klassifikatorischer Diagnostik für die Psychotherapie. Diagnostica, 34, 85-98.

Segal, S.P., Watson, M.A., Nelson, S. (1986). Consistency in the application of civil commitment standards in psychiatric emergency rooms. Jounal of Psychiatry Law, 14, 125-147.

Semler, G., Wittchen, H.U., Joschke, K., Zaudig, M., von Geiso, T., Kaiser, S., von Cranach, M. & Pfister, H. (1987). Test-retest reliability of a standardized psychiatric interview (DIS/CIDI). European Archives of Psychiatry and Neurological Sciences, 236, 214-222.

Shrout, P.E., Spitzer, R.C. & Fleiss, J.L. (1987). Quantification of agreement in psychiatric diagnosis revisited. Archives of General Psychiatry, 44, 172-177.

Spitzer, R.L., Endicott, J. & Robins, E. (1975). Clinical criteria for psychiatric diagnosis and DSM-III. American Journal of Psychiatry, 132, 1187-1192.

Spitzer, R.L. & Fleiss, J.L. (1974). A re-analysis of the reliability of psychiatric diagnoses. British Journal of Psychiatry, 125, 341-374.

Spitzer, R.L. & Wilson, P.T. (1975). Nosology and the official psychiatric nomenclature. In A. Freedman, A.H. Kaplan & B.J. Sadock (Eds.), Comprehensive textbook of psychiatry – II (Vol.1). Baltimore: Williams & Wilkins.

Spitznagel, E.L. & Helzer, J.E. (1985). A proposed solution to the base rate problem in the κ statistic. Archives of General Psychiatry, 42, 725-729.

Steller, M. (1994). Diagnostischer Prozeß. In R.-D. Stieglitz & U. Baumann (Hrsg.), Psychodiagnostik psychischer Störungen (S. 37-46). Stuttgart: Enke.

Stieglitz, R.-D. (2000). Diagnostik und Klassifikation psychischer Störungen. Göttingen: Hogrefe.

Stieglitz, R.-D. & Baumann, U. (Hrsg.). (1994). Psychodiagnostik psychischer Störungen. Stuttgart: Enke.

Stieglitz, R.-D., Fähndrich, E. & Möller, H.-J. (Hrsg.). (1998). Syndromale Diagnostik psychischer Störungen. Göttingen: Hogrefe.

Stieglitz, R.-D. & Freyberger, H.J. (1999a). Psychiatrische Untersuchung und Befunderhebung. In M. Berger (Hrsg.), Psychiatrie und Psychotherapie (4-30). München: Urban & Schwarzenberg.

Stieglitz, R.-D. & Freyberger, H.J. (1999b). Psychiatrische Diagnostik und Klassifikation. In M. Berger (Hrsg.), Psychiatrie und Psychotherapie (32-62). München: Urban & Schwarzenberg.

Thornton, Ch., Russell, J. & Hudson, J. (1998). Does the Composite International Diagnostic Interview Underdiagnose the Eating Disorders? International Journal of Eating Disorders, 23, 341-345.

Ventura, J., Liberman, R.P., Green, M.F., Shaner, A., Mintz, J. (1998). Training and quality assurance with the Structured Clinical Interview for DSM-IV (SCID-I/P). Psychiatry Research, 79, 163-173.

Wakefield, J.C. (1997a). Diagnosing DSM-IV – Part I: DSM-IV and the concept of disorder. Behavior Research Therapy, 35, 633-649.

Wakefield, J.C. (1997b). Diagnosing DSM-IV – Part II: Eysenck (1986) and the essentialist fallacy. Behavior Research Therapy, 35, 651-665.

Way, B.B., Allen, M.H., Mumpower, J.L, Stewart, T.R. & Banks, S.M. (1998). Interrater agreement among psychiatrist in psychiatric emergency assessments. American Journal of Psychiatry, 155, 1423-1428.

Wittchen, H.-U. (1994a). Klassifikation. In R.-D. Stieglitz & U. Baumann (Hrsg.), Psychodiagnostik psychischer Störungen (S. 49-63). Stuttgart: Enke.

Wittchen, H.-U. (1994b). Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): a critical review. Journal of Psychiatry Research, 28, 57-84.

Wittchen, H.-U., Lachner, G., Wunderlich, U., Pfister, H. (1998). Test-retest reliability of the computerized DSM-IV version of the Munich-Composite International Diagnostic Interview (M-CIDI). Social Psychiatry and Psychiatric Epidemiology, 33, 568-578.

Wittchen, H.-U., Müller, N., Pfister, H., Winter, S. & Schmidtkunz, B. (1999). Affektive, somatoforme und Angststörungen in Deutschland – Erste Ergebnisse des bundesweiten Zusatzsurveys „Psychische Störungen“. Gesundheitswesen, 61, 216-222.

Wittchen, H.-U. & Schulte D. (1988). Diagnostische Kriterien und operationalisierte Diagnosen. Grundlagen der Klassifikation psychischer Störungen. Diagnostica, 34, 3-27.

Wittchen, H.-U. & Semler, G. (1991). Composite International Diagnostic Interview. Weinheim: Beltz.

Wittchen, H.-U. & Vossen, A. (1996). Komorbiditätsstrukturen bei Angststörungen. In J. Margraf (Hrsg.), Lehrbuch der Verhaltenstherapie. Band 1 (S. 217-233). Berlin: Springer.

Wittchen, H.-U., Wunderlich, U., Gruschwitz, S. & Zaudig, M. (1997). Strukturiertes klinisches Interview für DSM-IV, Achse-I (SKID). Göttingen: Hogrefe.

Wittchen, H.-U. & Zerssen, D. von (Hrsg.). (1987). Verläufe behandelter und unbehandelter Depressionen und Angststörungen. – Eine klinisch-psychiatrische und epidemiologische Verlaufsuntersuchung. Berlin: Springer.

Wittchen, H.-U., Zaudig, M., Spengler, P., Mombour, W., Hiller, W., Essau, C.A., Rummler, R., Spitzer, R.L. & Williams, J. (1991). Wie zuverlässig ist operationalisierte Diagnostik? – Die Test-Retest-Reliabilität des Strukturierten Klinischen Interviews für DSM-III-R. Zeitschrift für Klinische Psychologie, 20, 136-153.

Wottawa, H. (1981). Psychologische Methodenlehre. (2. Aufl.) München: Juventa.

Yule, G.U. (1912). On the methods of measuring association between two attributes. Journal of the Royal Statistical Society, 75, 579-642.

Zimmerman, M. & Mattia, J.I. (1999). Psychiatric diagnosis in clinical pracice: is comorbidity being missed? Comprehensive Psychiatry, 40, 182-191.

Lebenslauf

Persönliche Angaben

- **Andrea Keller**
- geboren am: 11. August 1962 in Darmstadt

Ausbildung

1968 - 1971	Grundschule Rüsselsheim
1971 - 1981	Immanuel-Kant-Gymnasium Rüsselsheim **Abschluss: Abitur**
1981 - 1982	Johann-Wolfgang-Goethe Universität Frankfurt Studium der Kunstgeschichte und Ethnologie
1982 - 1985	Krankenpflegeausbildung **Abschluss: Krankenpflegeexamen**
1985 - 1991	Albert-Ludwigs-Universität Freiburg i. Br. **Studium der Psychologie, Abschluss: Diplom**

Berufliche Entwicklung

1991 - 1993	Neurologische Universitätsklinik Freiburg Forschungsprojekt „Prognose akuter Aphasien“
1992 - 1993	Psychiatrische Universitätsklinik Freiburg Schwerpunkt Krisenintervention, Depressionen
1994 - 1995	Psychosomatische Fachklinik Bad Pyrmont Behandlung von Essstörungen
1995 - 1997	Psychiatrische Universitätsklinik Heidelberg Behandlung von Persönlichkeitsstörungen, Depressionen
1997 - 1998	Neurologische Rehabilitationsklinik für Kinder und Jugendliche Kreischa bei Dresden: Diagnostik und kognitive Trainings
seit 6/1998	Klinik für Psychotherapie und Psychosomatik am Universitätsklinikum Dresden Leiterin der Tagesklinik Behandlung somatoformer Störungen, Angststörungen, Essstörungen und Persönlichkeitsstörungen

Spezielle Qualifikationen

1991 - 1996	Tübinger Akademie für Verhaltenstherapie **Abschluss: Verhaltenstherapie nach KV-Richtlinien**
1999	Freistaat Sachsen **Approbation: Psychologische Psychotherapeutin**