

Reliabilität des Diagnostischen Interviews bei Psychischen Störungen (DIPS für DSM-IV-TR) unter klinischen Routinebedingungen

Andrea Suppiger^a Tina In-Albon^a Chantal Herren^a Klaus Bader^b Silvia Schneider^a
Jürgen Margraf^c

^aKlinische Kinder- und Jugendpsychologie, Universität Basel,

^bVerhaltenstherapie-Ambulanz der Universitären Psychiatrischen Kliniken Basel,

^cKlinische Psychologie und Psychotherapie, Universität Basel, Schweiz

Schlüsselwörter

Diagnostik · DIPS für DSM-IV-TR · Strukturiertes Interview · Reliabilität · Fehleranalyse bei Nichtübereinstimmung

Zusammenfassung

Hintergrund: Ziel dieser Studie ist die Überprüfung der Interrater- und Retest-Reliabilität des an die DSM-IV-TR-Kriterien angepassten und erweiterten Diagnostischen Interviews bei Psychischen Störungen [DIPS für DSM-IV-TR; Schneider und Margraf, 2006]. **Methoden:** 10 geschulte Interviewerinnen führten 237 Interviews in klinischen ambulanten und stationären Einrichtungen durch. **Ergebnisse:** Die Ergebnisse sprechen für eine gute bis sehr gute Interrater-Reliabilität der Oberklassen Angststörungen, affektive Störungen, somatoforme Störungen, Schlafstörungen, Essstörungen, Substanz- und Alkoholmissbrauch/-abhängigkeit sowie aller überprüfbarer Diagnosekategorien und für den Ausschluss psychischer Störungen. Ein vergleichbares Bild ergab sich für die Retest-Reliabilitäten: Lediglich die Störungsoberklasse Schlafstörungen sowie die Diagnoseklassen generalisierte Angststörung und dysthyme Störung zeigten eine nicht zufriedenstellende Reliabilität. Die Reliabilitätswerte waren für die verschiedenen klinischen Settings (ambulant vs. stationär) vergleichbar. **Schlussfolgerung:** Die Ergebnisse zeigen, dass auch das vollständig überarbeitete und erweiterte DIPS für DSM-IV-TR ein reliables Interview zur Diagnostik psychischer Störungen darstellt und komorbide Störungsbilder, wie sie insbesondere im klinischen Alltag häufig anzutreffen sind, zuverlässig abbilden kann.

Key Words

Diagnostics · Structured Interview for DSM-IV-TR · Reliability · Analysis of non-consensus diagnosis

Summary

Reliability of the Structured Diagnostic Interview for Mental Disorders (DIPS for DSM-IV-TR) in Clinical Routine

Background: The aim of this study is to examine the interrater- and test-retest-reliability of the completely revised and expanded 'Diagnostisches Interview bei Psychischen Störungen für DSM-IV-TR' [DIPS for DSM-IV-TR; Schneider and Margraf, 2006]. **Method:** 10 certified interviewers conducted 237 interviews in in- and outpatient clinical facilities. **Results:** The results indicate a good interrater-reliability for the major diagnostic categories of anxiety disorders, mood disorders, somatoform disorders, eating disorders, sleeping disorders, alcohol and substance disorders, and for all specific psychiatric disorders, and the exclusion of psychiatric disorders. The results of the test-retest-reliability are similar: Only regarding the major diagnostic category of sleeping disorders and the specific psychiatric disorders of generalized anxiety disorder and dysthymic disorder the reliability was unsatisfactory. The reliability data are also similar in different clinical settings (in- vs. outpatient). **Conclusion:** The results indicate that the completely revised and expanded DIPS for DSM-IV-TR is a reliable interview for the assessment of mental disorders. Comorbid diagnoses, as often observed in daily clinical practice, can also be diagnosed reliably.

Hintergrund

Strukturierte und standardisierte Interviews haben sich in den letzten Jahren vorwiegend in der klinisch-psychologischen und in der psychiatrischen Forschung etabliert. In der Praxis verläuft die Etablierung eher zögerlich. Daher besteht ein Bedarf an diagnostischen Interviews, deren Gütekriterien sorgfältig abgeklärt werden müssen. Je nach Ausmaß der Standardisierung wird zwischen standardisierten und strukturierten Interviews unterschieden, wobei strukturierte Interviews im Gegensatz zu standardisierten Interviews eine flexiblere Struktur aufweisen und dem Anwender Beurteilungsfreiraum lassen. Im Folgenden werden unter diagnostischen Interviews immer strukturierte und standardisierte Interviews verstanden. Im deutschen Sprachraum existieren zur Erfassung psychischer Störungen im Erwachsenenalter derzeit die folgenden diagnostischen Breitbandinterviews: Composite International Diagnostic Interview [CIDI; Wittchen und Semler, 1990], Diagnostisches Expertensystem [DIA-X/M-CIDI; Wittchen und Pfister, 1997], Diagnostisches Interview bei Psychischen Störungen [DIPS für DSM-IV-TR; Schneider und Margraf, 2006], Strukturiertes Klinisches Interview für DSM-IV [SKID-I; Wittchen et al., 1997]. Tabelle 1 gibt einen Überblick über die Interrater- und Retest-Reliabilitäten dieser Interviews nach DSM-IV unter Angabe der untersuchten Studienpopulation. Insgesamt

werden für diese Interviews gute bis sehr gute Interrater- und Retest-Reliabilitäten berichtet. Ausnahmen bilden die generalisierte Angststörung und die dysthyme Störung, deren Kappa-Werte lediglich zwischen 0,13–0,67 bei allen erwähnten Interviews liegen und somit als unbefriedigend bis nur genügend gut interpretiert werden müssen. Betrachtet man die untersuchten Studienpopulationen, so fällt auf, dass die Überprüfung der Gütekriterien an Forschungspatienten beziehungsweise an Patienten einer psychosomatischen Klinik [z.B. Vorgängerversion des DIPS für DSM-IV-TR; Margraf et al., 1994], an einer Stichprobe aus der Allgemeinbevölkerung [z.B. M-CIDI; Wittchen et al., 1998] oder an einer vorselektierten Stichprobe [z.B. SKID-I; Zanarini et al., 2000] erfolgte. Für den Einsatz strukturierter Interviews im klinisch-psychologischen bzw. psychiatrischen Alltag ist jedoch zu prüfen, ob strukturierte und standardisierte Interviews auch unter klinischen ambulanten und stationären Routinebedingungen reliabel durchgeführt werden können. Deshalb war das Ziel der vorliegenden Studie, die Reliabilität und Praktikabilität der mit dem DIPS für DSM-IV-TR erstellten Diagnosen auch unter klinischen Alltagsbedingungen zu untersuchen. Neben Patienten aus psychosomatischen Kliniken und Teilnehmern aus Forschungsprojekten wurden gezielt Patienten ambulanter und stationärer psychiatrischer Einrichtungen einbezogen, um die Reliabilität des Interviews an Patienten mit unterschiedlich schweren

Tab. 1. Interrater- und Retest-Reliabilität (Kappa-Koeffizienten) von deutschsprachigen strukturierten und standardisierten Interviews nach DSM-IV. Angaben in Klammern beziehen sich auf einzelne Diagnosen, nicht in Klammern gesetzte Werte auf Störungsoberklassen (z.B. Angststörungen); – entsprechende Kategorie wurde nicht untersucht oder die für die Berechnung erforderlichen Basisraten sind nicht gegeben

Interview	Stichprobe	Angststörungen	Affektive Störungen	Somatoforme Störungen	Essstörungen	Substanzen	Schlafstörungen	Keine Störung
<i>Interrater-Reliabilität</i>								
SKID-I ¹	n = 27 (vorselektiert)	(0,57–0,88)	(0,76–0,80)	–	0,77	1,0	–	–
DIPS für DSM-IV-TR ²	n = 237 (Forsch.setting, ambul., stat.)	0,78 (0,68–0,88)	0,81 (0,73)	–	0,89	0,92	0,72	0,88
DIA-X/M-CIDI ³	–	–	–	–	–	–	–	–
<i>Retest-Reliabilität</i>								
SKID-I ¹	n = 52 (vorselektiert)	(0,44–0,78)	(0,35–0,61)	–	0,64	(0,76–0,77)	–	–
DIPS DSM-IV-TR ²	n = 50 (Forsch.setting, ambul., stat.)	0,76 (0,13–0,83)	0,80 (0,50–0,79)	0,62	0,77	0,94 (0,83)	0,35	–
DIA-X/M-CIDI ³	n = 142 (Allg.bevölk., ambul., stat.)	(0,57–0,72)	0,77 (0,45–0,69)	(0,49–0,67)	(0,66–0,83)	(0,65–0,73)	–	–
DIA-X/M-CIDI ⁴	n = 60 (Allg.bevölk.)	0,81 (0,45–1,00)	0,68 (0,64–0,70)	0,62	0,56 (0,55–0,85)	0,64–0,78	–	–

¹Zanarini et al. [2000], für die deutschsprachige Fassung des SKID für DSM-IV liegen keine gesonderten Gütekriterien vor.

²Suppiger et al. [vorliegende Arbeit].

³Wittchen und Pfister [1997].

⁴Wittchen et al. [1998].

psychischen Störungen zu testen. Zudem wurden im Gegensatz zu anderen Reliabilitätsstudien [z.B. ADIS-IV-L; Brown et al., 2001] lediglich ungenügende Deutschkenntnisse als Ausschlusskriterium festgelegt.

Methodik

DIPS für DSM-IV-TR

Das DIPS für DSM-IV-TR [Schneider und Margraf, 2006] ist eine überarbeitete und erweiterte Version des DIPS [Margraf et al., 1994]. Beide DIPS-Versionen orientieren sich am amerikanischen «Anxiety Disorders Interview Schedule» [ADIS; Di Nardo und Barlow, 1988; resp. ADIS-IV-L; Di Nardo et al., 1994]. Die Neuauflagen des Diagnostischen und Statistischen Manuals psychischer Störungen [DSM-IV-TR; American Psychiatric Association, 2000] und der Internationalen Klassifikation von Krankheiten [ICD-10; World Health Organisation, 1991] sowie umfangreiche Erfahrungen mit dem DIPS in Forschung und Praxis machten eine vollständige Überarbeitung des DIPS notwendig. Als grundlegende Änderung ist die Anpassung aller diagnostischen Fragen an die Kriterien des DSM-IV-TR [American Psychiatric Association, 2000; Sass et al., 2003] zu nennen. Aufgrund der Häufigkeit in der klinischen Praxis wurde das Interview um folgende für den psychotherapeutischen Bereich relevanten psychischen Störungen ergänzt: akute Belastungsstörung, undifferenzierte somatoforme Störung, Forschungsdiagnose der Binge-Eating-Störung, die Schlafstörungen primäre Hypersomnie und primäre Insomnie, Missbrauch bzw. Abhängigkeit von Substanzen, Medikamenten und Alkohol. In der klinischen Praxis häufig und auf Symptomebene mit Achse-I-Störungen am wenigsten überlappend wurde die Borderline-Persönlichkeitsstörung auf Achse II neu aufgenommen. Somit erfasst das DIPS für DSM-IV-TR sämtliche Diagnosekriterien für folgende Störungskategorien: Angststörungen, affektive und somatoforme Störungen, Ess- und Schlafstörungen, Alkohol- und Substanzmissbrauch/-abhängigkeit, Borderline-Persönlichkeitsstörung sowie Screenings für körperliche Krankheiten, nichtorganische Psychosen, Tabakkonsum, Medikamentengebrauch und sexuelle Funktionsstörungen. Ebenfalls umfasst das DIPS einen klinisch-demographischen Teil, und es werden eine Anamnese psychischer Störungen, eine Familienanamnese psychischer Störungen sowie die Achsen IV und V des DSM-IV-TR erfasst. Anhand einer Tabelle können die DSM-IV-TR-Diagnosen formal in ICD-10-Diagnosen überführt werden. Inhaltlich muss dabei beachtet werden, dass die Überführung in ICD-10-Diagnosen aufgrund teilweise unterschiedlicher Diagnosekriterien nur näherungsweise erfolgen kann. Analog zur Vorgängerversion enthält das DIPS für DSM-IV-TR therapierelevante Fragen, um das Interview insbesondere auch für den psychotherapeutischen Gebrauch nutzbar zu machen. In Anlehnung an das ADIS-IV-L wurde die bisher fünfstufige klinische Einschätzung der Symptome (Beeinträchtigung und Belastung) auf eine neunstufige Punkte-Skala (0–8) erweitert [vgl. Wittchen et al., 1999]. Analog zur Vorgängerstudie [Margraf et al., 1991] konnte auch für das DIPS für DSM-IV-TR eine gute Validität der Oberklassen Angststörungen, affektive Störungen, somatoforme Störungen, Essstörungen, Substanz- und Alkoholmissbrauch/-abhängigkeit sowie einzelner, überprüfbarer Störungskategorien und für den Ausschluss psychischer Störungen nachgewiesen werden [In-Albon et al., 2008].

Ablauf der Untersuchung

Zur Bestimmung der Interrater-Reliabilität wurden 237 diagnostische Interviews in klinisch ambulanten und stationären Einrichtungen auf iPod aufgenommen und danach von einer zweiten unabhängigen Interviewerin gegenkodiert. Zur Bestimmung der Retest-Reliabilität wurden bei insgesamt 50 nacheinander angemeldeten Patienten (82% aus dem stationären Setting) zwei Interviews von voneinander unabhängigen Interviewerinnen im Abstand von 4,7 (SD = 4,2, Range 0–18) Tagen durchgeführt. 3,9%

der angesprochenen Patienten lehnten eine Re-Testung aus Zeitgründen oder psychischer Belastung ab. Die Interviewerinnen besprachen alle Interviews in regelmäßig stattfindenden Fallkonferenzen und bildeten bei abweichenden Diagnosen Konsensdiagnosen. Gründe für mangelnde Übereinstimmung wurden in von den Autoren entwickelten Checklisten separat für jede Störungskategorie erfasst. Die Interviews wurden möglichst vor Beginn einer Behandlung durchgeführt, um die Varianz des psychopathologischen Zustands der Patienten durch mögliche therapeutische Interventionen gering zu halten. Zur Erhebung von Lebenszeitdiagnosen wurden die Patienten nach derzeitigen (zum Zeitpunkt des Interviews) und nach früheren psychischen Problemen befragt. Alle Patienten wurden über die Fragestellung der Untersuchung aufgeklärt und gaben ihr schriftliches Einverständnis.

Interviewerinnen

Die Interviews wurden von 10 diplomierten Psychologinnen durchgeführt. Zwei Interviewerinnen hatten das Training zur Durchführung des ADIS am Center for Anxiety and Related Disorders in Albany respektive der Boston University erhalten. Die neu zu trainierenden Interviewerinnen erhielten ein standardisiertes Training in Anlehnung an das ADIS-IV-L Training [Brown et al., 2001] und wurden nach Abschluss einer erfolgreichen Zertifizierung in die Studie aufgenommen [siehe Handbuch des DIPS für DSM-IV-TR; Schneider und Margraf, 2006]. Für die Durchführung der Gegenkodierungen zur Bestimmung der Interrater-Reliabilität wurden zusätzlich zwei weitere Interviewer (50% weiblich) geschult.

Patienten

Im Zeitraum Januar 2005 bis Oktober 2005 wurden 204 Patienten (66% Frauen, 34% Männer) für die Reliabilitätsstudie in verschiedenen klinischen Einrichtungen rekrutiert. 67 Patienten (33%) stammten aus ambulanten, 98 Patienten (48%) aus stationären psychiatrischen oder psychosomatischen Einrichtungen und 39 Probanden (19%) aus einem Forschungsprojekt der Universität Basel. Die Rekrutierung erfolgte über die verschiedenen Institutionen, wobei zu Beginn der Studie vereinbart wurde, dass alle Patienten im benannten Zeitraum auf die Studienteilnahme angesprochen werden. Die Erfassung eines standardisierten Rekrutierungsprotokolls war aus organisatorischen Gründen leider nicht möglich. In den ambulanten Einrichtungen sowie im Forschungsprojekt konnten alle Patienten für die Studie rekrutiert werden, wobei ein unvollständiges Interview bei den Berechnungen ausgeschlossen wurde. Die Rekrutierungsrate in den drei stationären Einrichtungen betrug post hoc ermittelt 50%. Hauptgrund für eine Nicht-Teilnahme waren zeitorganisatorische Gründe. Da alle Interviews vor Beginn der Therapie stattfinden mussten, um den Einfluss möglicher Therapieeffekte auf die Berichterstattung zu kontrollieren, war es nicht immer möglich, in der ersten Woche des Klinikaufenthalts die Interviews durchzuführen. Insgesamt gehen wir daher nicht von einem systematischen Selektionseffekt aus.

Das Durchschnittsalter betrug 38 Jahre (SD = 12,89, Range 17–82). 38% der Patienten waren zum Zeitpunkt des Interviews verheiratet bzw. lebten in einer festen Partnerschaft, 47% waren ledig, 12% waren geschieden oder lebten getrennt und 3% waren verwitwet. Die durchschnittliche Anzahl besuchter Schuljahre lag bei 11 Jahren (SD = 2,5).

Maße zur Reliabilitätsbestimmung

Zur Bestimmung der Reliabilität werden drei Maße jeweils auf der Ebene von Oberklassen psychischer Störungen (z.B. affektive Störungen, Angststörungen) und auf der Ebene der einzelnen Diagnosekategorien (z.B. dysthyme Störung, Major Depression) angegeben: die prozentuale Übereinstimmung, der Kappa-Koeffizient [Cohen, 1960] und der Yules Y-Koeffizient [Yule, 1912]. Da Komorbidität im klinischen Alltag ein häufiges Phänomen ist, waren wir insbesondere daran interessiert, inwiefern mit Hilfe des DIPS auch mehrere Diagnosen gleichzeitig reliabel gestellt werden können. Dazu berechneten wir die totale prozentuale Übereinstimmung zwischen zwei Interviewerinnen. Eine Übereinstimmung lag nur dann vor, wenn beide Interviewerinnen in allen aktuellen und früheren

Diagnosen übereinstimmten, wobei zwischen Haupt- und Zusatzdiagnose nicht unterschieden wurde.

In Anlehnung an Fleiss [1981] werden Kappa-Werte wie folgt interpretiert: Sehr gute Übereinstimmung ($\kappa \geq 0,75$), gute Übereinstimmung ($0,60 \leq \kappa < 0,74$), genügende Übereinstimmung ($0,40 \leq \kappa < 0,59$), mangelhafte Übereinstimmung ($\kappa < 0,40$). Für Yules Y-Koeffizient gelten nach Landis und Koch [1977] erst Werte $> 0,81$ als ausgezeichnete und Werte zwischen 0,65 und 0,80 als ausreichende Übereinstimmung. Die Basisrate errechneten wir, indem wir aus den für die Kappa-Berechnung zugrunde liegenden Vierfeldertafeln die Werte aus den Feldern (ja/ja) und (nein/ja + ja/nein) dividiert durch 2) addierten.

Fehleranalyse

In den Fallbesprechungen wurden Gründe für mangelnde Übereinstimmung zwischen zwei Interviewerinnen auf vorgegebenen Checklisten (können bei der Erstautorin dieser Studie angefordert werden) für jede Störungskategorie festgehalten. Folgende acht potentielle Fehlerquellen wurden bei Nichtübereinstimmung unterschieden, wobei auch mehrere Fehlerquellen auftreten konnten.

1. Interviewerfehler: Sprungregel falsch angewendet, starke Abweichung von standardisierten Fragen, ungenügende Folgefragen zu Diagnoseklärung (z.B. ob Angst vor Fliegen aufgrund von Panik oder aufgrund von spezifischer Phobie auftritt).
2. Protokollierungsfehler: z.B. Flüchtighkeitsfehler bei der Protokollierung.
3. Unterschiedliche Symptomgewichtung: unterschiedliche Bewertung des Schweregrades der Symptome oder der Beeinträchtigung.
4. Auswertungsfehler: DSM-IV-TR-Kriterien falsch angewendet, differentialdiagnostische Aspekte falsch beachtet.
5. Nichtübereinstimmung aufgrund der Qualität der Aufnahme: Aufnahme fehlt, Antwort akustisch nicht nachvollziehbar (z.B. Patient nickt).
6. Unterschiedliche differentialdiagnostische Gewichtung: z.B. Schlafprobleme, die als Symptom der Major Depression oder als primäre Insomnie gewertet werden.
7. DSM-IV-TR oder DIPS unpräzise: unzureichender Informationsgewinn durch unpräzise Spezifikation eines Begriffs oder zu allgemein formulierte Diagnosekriterien.
8. Andere: z.B. Nichtübereinstimmung bei einer spezifischen Störung, die nicht direkt mit dem DIPS erfasst werden kann.

Bei Nichtübereinstimmung von Erst- und Retest-Interview kommen folgende zwei Fehlerquellen hinzu:

1. Informationsvarianz von Seiten des Patienten: Patient antwortet auf die gleiche Frage unterschiedlich.
2. Subjektvarianz: Zur Erfassung der Varianz wurde den Patienten direkt im Anschluss an das Retest-Interview ein Fragebogen mit fünf Items bezüglich Therapieeinfluss, Veränderung der Beschwerden oder der Interpretation der Beschwerden, unterschiedliche Tagesverfassung und unterschiedliche therapeutische Beziehung ausgeteilt.

Ergebnisse

96% der Interviews wurden vollständig durchgeführt. Im Durchschnitt vergaben die Interviewerinnen 2,3 Lebenszeitdiagnosen ($SD = 1,80$, Range 0–9). Die Interviews dauerten im Durchschnitt 106 min ($SD = 49,24$, Range 15–241).

Interrater-Reliabilität

In den Tabellen 2–5 sind die Ergebnisse zur Interrater-Reliabilität für Lebenszeitdiagnosen auf der Ebene von Oberklassen psychischer Störungen sowie auf der Ebene der einzelnen Diagnosekategorien aufgeführt. Bei den Oberklassen psychi-

Tab. 2. Interrater-Reliabilität des DIPS (Oberklasse, Lebenszeitdiagnosen)

DSM-IV-TR Oberklassen	Häufigkeiten		%	Kappa	Yules Y
	-/-	-/+			
	+/-	+/+			
Angststörungen	91	9	89	0,78	0,97
	17	120			
Affektive Störungen	88	8	91	0,81	0,98
	14	127			
Somatoforme Störungen	205	6	92	(0,57)	0,95
	12	14			
Essstörungen	201	3	98	0,89	1,00
	3	30			
Schlafstörungen	204	8	95	0,72	0,98
	5	20			
Störungen im Zusammenhang mit Substanzen	186	2	98	0,92	1,00
	4	45			
Keine Störung	32	3	97	0,88	1,00
	4	198			

Kappa-Koeffizienten, die aufgrund einer Basisrate der jeweiligen Oberklasse von $< 10\%$ ($< 23,8$ Fälle) eine Unterschätzung darstellen, sind in Klammern aufgeführt.

Tab. 3. Interrater-Reliabilität des DIPS (Störungsebene, Lebenszeitdiagnosen); Sektion Angststörungen

DSM-IV-TR Störungsebenen	Häufigkeiten		%	Kappa	Yules Y
	-/-	-/+			
	+/-	+/+			
Panikstörung mit/ohne Agoraphobie	195	3	97	0,88	1,00
	5	34			
Agoraphobie ohne Panikstörung	223	5	97	(0,65)	0,99
	2	7			
Soziale Phobie	162	6	92	0,80	0,98
	13	56			
Spezifische Phobie	203	4	94	0,68	0,98
	11	19			
Generalisierte Angststörung	217	4	97	(0,73)	0,99
	4	12			
Zwangsstörung	211	1	97	(0,83)	1,00
	6	19			
Posttraumatische Belastungsstörung	208	3	96	0,80	0,99
	6	20			

Kappa-Koeffizienten, die aufgrund einer Basisrate der jeweiligen Oberklasse von $< 10\%$ ($< 23,8$ Fälle) eine Unterschätzung darstellen, sind in Klammern aufgeführt.

scher Störungen ergibt sich eine insgesamt gute bis sehr gute Interrater-Reliabilität (Tab. 2). Die prozentuale Übereinstimmung beträgt bei allen Oberklassen mindestens 89%. Die Kappa-Werte liegen zwischen 0,72 und 0,92, Yules Y-Werte zwischen 0,95 und 1,00.

Bei den einzelnen Diagnoseklassen ergaben sich für die Interrater-Reliabilität von Lebenszeitdiagnosen gute bis sehr gute Werte (Tab. 3–5). Alle prozentualen Übereinstimmungen

Tab. 4. Interrater-Reliabilität des DIPS (Störungsebene, Lebenszeitdiagnosen); Sektionen affektive Störungen, somatoforme Störungen, Borderline-Persönlichkeitsstörung

DSM-IV-TR Störungsebenen	Häufigkeiten		%	Kappa	Yules Y
	-/-	-/+			
	+/-	+/+			
Major Depression	105	12	87	0,73	0,96
	20	100			
Bipolare I und II Störung	228	11	98	(0,71)	1,00
	3	5			
Dysthyme Störung	213	3	96	(0,71)	0,99
	7	14			
Hypochondrie	230	2	98	(0,43)	0,97
	3	2			
Somatisierungsstörung	226	3	96	(0,15)	0,83
	7	1			
Undifferenzierte somatoforme Störung	228	3	98	(0,71)	1,00
	1	5			
Schmerzstörung	228	3	97	(0,35)	0,95
	4	2			
Borderline-Persönlichkeitsstörung	221	2	97	(0,65)	0,99
	6	8			

Kappa-Koeffizienten, die aufgrund einer Basisrate der jeweiligen Oberklasse von <10% (<23,8 Fälle) eine Unterschätzung darstellen, sind in Klammern aufgeführt.

Tab. 5. Interrater-Reliabilität des DIPS (Störungsebene, Lebenszeitdiagnosen); Sektionen Essstörungen, Schlafstörungen, Störungen im Zusammenhang mit Substanzen

DSM-IV-TR Störungsebenen	Häufigkeiten		%	Kappa	Yules Y
	-/-	-/+			
	+/-	+/+			
Anorexia nervosa	220	2	99	(0,90)	1,00
	1	14			
Bulimia nervosa	223	2	99	(0,92)	-
	0	12			
Binge-Eating-Störung	227	2	99	(0,82)	1,00
	1	7			
Primäre Insomnie	209	6	95	(0,73)	0,98
	5	17			
Primäre Hypersomnie	231	3	98	(0,49)	0,99
	1	2			
Alkoholmissbrauch	225	3	97	(0,57)	0,98
	4	5			
Alkoholabhängigkeit	208	5	97	0,85	1,00
	2	22			
Medikamentenabhängigkeit	232	2	99	(0,75)	-
	0	3			
Substanzmissbrauch	230	2	98	(0,59)	0,99
	2	3			
Substanzabhängigkeit	222	2	97	(0,62)	0,99
	6	7			

Kappa-Koeffizienten, die aufgrund einer Basisrate der jeweiligen Oberklasse von <10% (<23,8 Fälle) eine Unterschätzung darstellen, sind in Klammern aufgeführt.

Tab. 6. Retest-Reliabilität des DIPS (Oberklasse, Lebenszeitdiagnosen)

DSM-IV-TR Oberklassen	Häufigkeiten		%	Kappa	Yules Y
	-/-	-/+			
	+/-	+/+			
Angststörungen	20	3	88	0,76	0,96
	3	24			
Affektive Störungen	12	2	92	0,80	0,98
	2	34			
Somatoforme Störungen	42	1	92	0,62	0,97
	3	4			
Essstörungen	41	3	94	0,77	-
	0	6			
Schlafstörungen	42	1	88	0,35	0,89
	5	2			
Störungen im Zusammenhang mit Substanzen	38	0	98	0,94	-
	1	11			
Keine Störung	47	1	98	(0,79)	-
	0	2			

Kappa-Koeffizienten, die aufgrund einer Basisrate der jeweiligen Oberklasse von <10% (<23,5 Fälle) eine Unterschätzung darstellen, sind in Klammern aufgeführt.

lagen bei den einzelnen Diagnoseklassen über 87%. Die Kappa-Werte lagen zwischen 0,68 und 0,88, die Y-Koeffizienten zwischen 0,83 und 1,00. Die globale prozentuale Übereinstimmung zeigte zudem, dass in 61% der Interviews zwei unabhängige Interviewerinnen in allen aktuellen und früheren Diagnosen übereinstimmten, wobei Komorbiditätsmuster von bis zu neun Diagnosen in diese Berechnungen eingeschlossen wurden.

Retest-Reliabilität

Die Ergebnisse zur Retest-Reliabilität bestätigen die Ergebnisse zur Interrater-Reliabilität. Bei den Oberklassen (Tab. 6) lagen die prozentualen Übereinstimmungen für Lebenszeitdiagnosen alle bei mindestens 88%. Die Kappa-Werte lagen bis auf die Oberklasse der Schlafstörungen ($\kappa = 0,35$) zwischen 0,62 und 0,94, die Y-Koeffizienten zwischen 0,89 und 0,98.

Auf der Ebene der einzelnen Diagnoseklassen ergab sich ebenfalls eine gute bis sehr gute Übereinstimmung. Die prozentuale Übereinstimmung lag bei mindestens 84%. Die Kappa-Werte lagen zwischen 0,73 und 0,83, Yules Y-Koeffizienten zwischen 0,91 und 0,99. Eine Ausnahme bilden die dysthyme Störung und die generalisierte Angststörung, die mit Kappa-Werten von 0,50 und 0,13 genügend respektive ungenügend ausfallen. Der Ausschluss von Lebenszeitdiagnosen erfolgte in 98% der Fälle übereinstimmend.

Weitere Analysen

Da bei Lebenszeitdiagnosen die zeitliche Reihung der Diagnosen nicht beachtet wird, wurden dieselben Berechnungen auch für die Hauptdiagnosen berechnet. Dabei liegt eine Übereinstimmung nur dann vor, wenn beide Diagnostiker die-

Tab. 7. Interrater-Reliabilität des DIPS (Oberklasse, Lebenszeitdiagnosen) im stationären Setting

DSM-IV-TR Oberklassen	Häufigkeiten		%	Kappa	Yules Y
	-/-	-/+			
	+/-	+/+			
Angststörungen	45	6	87	0,73	0,96
	11	69			
Affektive Störungen	24	5	89	0,69	0,96
	10	92			
Somatoforme Störungen	110	5	92	0,64	0,96
	5	11			
Essstörungen	112	1	99	0,94	1,00
	1	17			
Schlafstörungen	112	7	92	0,55	0,94
	4	8			
Störungen im Zusammenhang mit Substanzen	91	1	96	0,91	1,00
	4	35			
Keine Störung	5	1	95	0,76	0,99
	2	123			

Kappa-Koeffizienten, die aufgrund einer Basisrate der jeweiligen Oberklasse von <10% (<13,1 Fälle) eine Unterschätzung darstellen, sind in Klammern aufgeführt.

selbe Diagnose als primär eingestuft hatten. Für die ausschließlich als primär eingestuften Diagnosen zeigt sich ein ähnliches Bild wie für die Lebenszeitdiagnosen. Auf der Ebene der Oberklassen liegt die prozentuale Übereinstimmung bei mindestens 88%, Yules Y fällt mit Werten zwischen 0,94 und 1,00 ebenfalls sehr zufriedenstellend aus, die Kappa-Werte sind gut bis sehr gut (0,74–0,92). Auch bei den einzelnen Diagnoseklassen sind die Werte aus der Übereinstimmung von Lebenszeitdiagnosen mit der ausschließlichen Wertung der Hauptdiagnosen vergleichbar. Die prozentuale Übereinstimmung liegt in allen Diagnoseklassen bei mindestens 85% mit Kappa-Werten zwischen 0,67 und 0,85 und Y-Koeffizienten zwischen 0,93 und 1,00. Auf die ausschließliche Wertung der Hauptdiagnosen bei der Retest-Reliabilität wird aufgrund zu geringer Basisraten verzichtet.

Betrachtet man ausschließlich die im stationären Setting erhobenen Diagnosen, so zeigt sich sowohl bei den Oberklassen wie auch bei den einzelnen Diagnoseklassen ein vergleichbares Bild. Bei den Oberklassen (Tab. 7) liegt die prozentuale Übereinstimmung für Lebenszeitdiagnosen bei mindestens 87%, Yules Y fällt mit Werten zwischen 0,94 bis 1,00 sehr zufriedenstellend aus. Mit einer Ausnahme der Oberklasse der Schlafstörungen (Kappa = 0,55) sind die Kappa-Werte für die im stationären Setting erhobenen Diagnosen gut bis sehr gut (0,64–0,91). Bei den einzelnen Diagnoseklassen liegt die prozentuale Übereinstimmung der ausschließlich im stationären Setting erhobenen Lebenszeitdiagnosen bei mindestens 82%, die Kappa-Werte sind mit Ausnahme der Major Depression ($\kappa = 0,62$) höher als $\kappa = 0,72$, die Y-Werte liegen zwischen 0,90 und 1,00.

Ergebnisse der Fehleranalyse

Die häufigste Fehlerquelle bei mangelnder Übereinstimmung bestand darin, dass die beiden Diagnostiker den Schweregrad von Symptomen unterschiedlich stark gewichteten (30%). Weitere Fehler kamen dadurch zustande, dass die Interviewerin die Diagnosekriterien respektive das DIPS falsch angewendet (23%), einen Fehler bei der Interviewführung gemacht (18%) oder eine unterschiedliche differentialdiagnostische Gewichtung vorgenommen hat (12%). Alle weiteren erhobenen Fehlerquellen spielten bei Nichtübereinstimmungen kaum eine Rolle (<9%). Bei den Retest-Interviews bestand die häufigste Fehlerquelle (42%) darin, dass der Patient beim ersten und zweiten Interview unterschiedliche Angaben machte.

Diskussion

Die Ergebnisse der vorliegenden Studie zeigen, dass die überarbeitete Version des DIPS für DSM-IV-TR auch unter schwierigen Routinebedingungen eine gute bis sehr gute Interrater-Reliabilität aufweist. Dies gilt sowohl für die DIPS-Oberklassen Angststörungen, affektive Störungen, somatoforme Störungen, Essstörungen, Schlafstörungen und Störungen im Zusammenhang mit Substanzen als auch für die einzelnen DIPS-Diagnosen Panikstörung mit/ohne Agoraphobie, Agoraphobie ohne Panikstörung, soziale Phobie, spezifische Phobie, generalisierte Angststörung, Zwangsstörung, posttraumatische Belastungsstörung, Major Depression, bipolare Störung I und II, dysthyme Störung, Hypochondrie, Somatisierungsstörung, undifferenzierte somatoforme Störung, Schmerzstörung, Anorexia und Bulimia nervosa, Binge-Eating-Disorder, primäre Insomnie und Hypersomnie, Alkoholmissbrauch und -abhängigkeit, Substanzmissbrauch und -abhängigkeit und Borderline-Persönlichkeitsstörung. Der Ausschluss von psychischen Störungen erwies sich ebenfalls als sehr reliabel. Lag keine Lebenszeitdiagnose vor, stimmte das diagnostische Urteil der Interviewerinnen in 97% der Fälle überein.

Die Ergebnisse zur Retest-Reliabilität untermauern die Ergebnisse zur Interrater-Reliabilität. Mit Ausnahme der Oberklasse der Schlafstörungen ($\kappa = 0,35$) sind alle Kappa- und Y-Koeffizienten als gut bis sehr gut zu betrachten. Da die Schlafstörungen im SKID und im DIA-X/M-CIDI nicht erfasst werden, ist ein Vergleich mit diesen Interviews nicht möglich. Der niedrige Kappa-Wert bei der Oberklasse der Schlafstörungen ist aufgrund der durchgeführten Fehleranalyse bei Nichtübereinstimmung vor allem auf die unpräzise Operationalisierung der DSM-IV-TR-Kriterien zur primären Insomnie und primären Hypersomnie zurückzuführen. 50% der Nichtübereinstimmungen waren bedingt durch unterschiedliche differentialdiagnostische Gewichtung (Kriterium D: «Das Störungsbild tritt nicht ausschließlich im Verlauf einer anderen psychischen Störung auf») und unterschiedliche Symptomgewichtung (vor allem beim Kriterium B: «Die Schlafstörung verursacht in klinisch bedeutsamer Weise Leiden oder Beeinträchtigung»).

Anzufügen ist, dass sich für Schlafstörungen die Internationale Klassifikation der Schlafstörungen [ICSD; Schramm und Riemann, 1995] durchgesetzt hat, da das DSM-IV-TR und die ICD-10 die Schlafstörungen zu wenig differenzieren [Ohayon und Roberts, 2001]. Auf der Ebene der einzelnen Diagnoseklassen ergab sich insgesamt ebenfalls eine gute bis sehr gute Retest-Reliabilität. Eine Ausnahme bilden die dysthyme Störung und die generalisierte Angststörung, die mit Kappa-Werten von 0,50 und 0,13 nur genügend respektive ungenügend ausfallen. Frühere Reliabilitätsstudien mit dem DIPS respektive dem englischsprachigen Anxiety Disorders Interview Schedule for DSM-IV [ADIS-IV; Brown et al., 1994] berichten auch von unbefriedigender Übereinstimmung bei der dysthymen Störung [Schneider et al., 1992; Brown et al., 2001] respektive generalisierten Angststörung [Schneider et al., 1992; Keller, 2000]. Genauso berichten Reliabilitätsstudien zum SKID-I [Zanarini et al., 2000] und zum DIA-X/M-CIDI [Wittchen et al., 1998] von nur genügend hohen Kappa-Werten bei generalisierter Angststörung. Bei der dysthymen Störung wird auch für das SKID-I [Zanarini et al., 2000] eine mangelhafte und für das DIA-X/M-CIDI [Wittchen und Pfister, 1997] eine genügende Übereinstimmung berichtet. Der niedrige Kappa-Wert bei der dysthymen Störung kann einerseits durch die schwierige Abgrenzung von der Major Depression erklärt werden: Werden die beiden Diagnoseklassen zusammengefasst auf der Oberklasse der affektiven Störungen betrachtet, fallen die Reliabilitätswerte gut aus. Andererseits ergab die Betrachtung der Fehleranalyse, dass 60% der Nichtübereinstimmungen durch Informationsvarianz von Seiten des Patienten zustande gekommen sind. Detaillierte Analysen haben zudem gezeigt, dass die Patienten in dieser Sektion vor allem Mühe hatten, reliable Angaben zur Dauer der affektiven Symptomatik zu machen. Die niedrige Reliabilität der generalisierten Angststörung geht einher mit dem Befund, dass dieses Störungsbild auch eine schlechte diskriminante Validität aufweist [In-Albon et al., 2008]. Durch die enge Verknüpfung strukturierter und standardisierter Interviews mit der klassifikatorischen Diagnostik kann die vorliegende Studie auch als Überprüfung der Reliabilität der Diagnosekriterien des DSM-IV-TR betrachtet werden. Die beschriebenen Schwachstellen des DIPS, welche sich mit anderen strukturierten und standardisierten Interviews decken, zeigen somit auch die nach wie vor bestehenden unbefriedigenden Operationalisierungen einzelner Diagnoseklassen im DSM-IV-TR.

Als häufigste Fehlerquellen bei mangelhafter Übereinstimmung wurden unterschiedliche Handhabung der Diagnosekriterien (Symptomgewichtung, differentialdiagnostische Gewichtung) sowie fehlerhafte Handhabung des DIPS respektive der DSM-IV-TR-Diagnosekriterien (Auswertung und Durchführung des Interviews) verantwortlich gemacht. Bei den Retest-Interviews kam als häufigste Fehlerquelle hinzu, dass die Patienten in den beiden Interviews unterschiedliche Angaben machten. Eine Reliabilitätsprüfung der Fehlerquellenanalyse wurde in der vorliegenden Studie nicht vorgenommen.

Trotz der zunehmenden Operationalisierung des DSM bleibt dem Diagnostiker bei der Anwendung strukturierter Interviews Spielraum bei differentialdiagnostischen Überlegungen und bei der Einschätzung des Schweregrades von Symptomen, was die Bedeutung eines systematischen Trainings unterstreicht. Standardisierte Verfahren, wie z.B. das CIDI [Wittchen und Semler, 1990] und das diagnostische Expertensystem DIA-X/M-CIDI [Wittchen und Pfister, 1997] verzichten im Vergleich zum DIPS explizit auf das klinische Urteil des Diagnostikers. Obwohl das DIPS als strukturiertes Interview dem klinischen Urteil deutlich mehr Gewicht einräumt als z.B. ein Fragebogen oder ein voll standardisiertes Interview, konnten mit Hilfe des Interviews auch unter klinischen Routinebedingungen sehr reliable Diagnosen gestellt werden, die vergleichbar sind mit denen von standardisierten Verfahren (siehe Tab. 1).

Ein Anliegen der Studie war es, für die Überprüfung der Reliabilität des DIPS für DSM-IV-TR nicht in erster Linie Teilnehmer aus Forschungsprojekten oder aus der Allgemeinbevölkerung heranzuziehen, sondern vielmehr Patienten aus unterschiedlichen klinisch-psychologischen und psychiatrischen Institutionen zu rekrutieren und unter Alltagsbedingungen zu interviewen. Wittchen et al. [1998] merken in ihrer Studie zur Überprüfung der Retest-Reliabilität des M-CIDI z.B. an, dass sie nur wenig stark beeinträchtigte Patienten interviewt hatten. Für stärker beeinträchtigte Patienten würden sie andere Reliabilitätswerte erwarten. In Anbetracht dessen, dass die erhobene Stichprobe in unserer Studie groß war ($n = 237$ Interviews), zu etwa gleichen Teilen aus stationären und ambulanten Einrichtungen stammte und insgesamt ein hohes Maß an Komorbidität aufwies, können die Ergebnisse der psychometrischen Überprüfung als sehr gut beurteilt werden. Zudem wurde während der Datenerhebung darauf verzichtet, Patienten mit spezifischen Störungen gezielt zu rekrutieren, da dadurch eine hypothesenfreie Befragung der Patienten nicht mehr gewährleistet wäre. Beim Vergleich von Reliabilitätsdaten aus verschiedenen Studien ist zudem zu beachten, ob nur einzelne Störungssektionen (z.B. nur die Sektion Angststörungen oder die Sektion Alkohol-, Substanz- und Medikamentenmissbrauch /-abhängigkeit) oder ein Interview als Ganzes, welches die ganze Bandbreite an psychischen Störungen abdeckt, hinsichtlich seiner Gütekriterien untersucht wurde. Je mehr Störungsbereiche auf einmal überprüft werden und je höher die Komorbidität der Stichprobe ausfällt, desto schwieriger und komplexer wird die Beurteilung von differentialdiagnostischen Überlegungen (z.B. ob das Störungsbild nicht auf die direkte körperliche Wirkung einer Substanz zurückgeht oder die Symptomatik nicht besser durch eine andere psychische Störung erklärt werden kann). Im Vergleich zu anderen Reliabilitätsstudien [z.B. ADIS-IV-L; Brown et al., 2001] wurden bei der Überprüfung des DIPS für DSM-IV-TR nebst ungenügender Deutschkenntnisse bewusst keine weiteren Ausschlusskriterien, wie z.B. psychotische Symptomatik, Alkohol- oder Substanzmissbrauch/-ab-

hängigkeit oder Suizidalität festgelegt. Die insgesamt 32 verschiedenen Störungsbereiche wurden zudem in einer einzigen Datenerhebung überprüft. Die durchschnittliche Anzahl von 2,4 Diagnosen pro Patient spricht für eine relativ hohe Komorbidität, welche komplexe differentialdiagnostische Überlegungen erforderlich machte.

Abschließend soll auf Einschränkungen der vorliegenden Studie aufmerksam gemacht werden. So ist festzuhalten, dass bei einigen Störungskategorien, z.B. den somatoformen Störungen, die Stichprobe klein war ($n = 11$) und somit die Ergebnisse zur Validität einzelner Störungskategorien als Vorbefunde zu betrachten sind [In-Albon et al., 2008]. Zudem müssen die Ergebnisse zur Retest-Reliabilität aufgrund der relativ kleinen Stichprobe ($n = 50$) als vorläufig betrachtet werden. Bezüglich der Praktikabilität im klinischen Alltag wird eine durchschnittliche Interviewdauer von 106 min ($SD = 49,24$, Range 15–241) von psychologischen bzw. ärztlichen Psychotherapeuten und Psychiatern oft als kritisch und zudem belastend für die Patienten betrachtet. Die Daten aus einer Akzeptanzstudie zum DIPS zeigen jedoch eindrücklich, dass Patienten die strukturierte Diagnostik mit dem DIPS für DSM-IV-TR schätzen und sie sich durch die genaue Befragung gut verstanden fühlen [Suppiger et al., in press]. Eine gründliche Diagnostik schafft Vertrauen in die folgende Therapie und ist ihre essentielle Basis [Hoyer et al., 2006; Ehler, 2007]. Bedenkt man zudem, dass eine psychotherapeutische Behandlung in der Regel 25 Stunden und länger dauert, so scheint ein Aufwand von durchschnittlich zwei Therapiestunden für eine

sorgfältige Befunderhebung mehr als gerechtfertigt. In vielen anderen Bereichen der Medizin ist das Verhältnis von Diagnostik zu Therapie deutlich ungünstiger.

Die vorliegende Studie zeigt, dass das strukturierte Interview DIPS eine reliable Erhebung psychischer Störungen nicht nur bei ausgewählten Forschungsteilnehmern klinisch-psychologischer/psychiatrischer Studien erlaubt, sondern auch bei ambulanten und stationären Patienten mit zum Teil sehr komplexen Diagnosen. Gemeinsam mit den zufriedenstellenden Befunden zur Validität [In-Albon et al., 2008] und der hohen Akzeptanz des DIPS bei den befragten Patienten [Suppiger et al., in press] sprechen diese Ergebnisse für den routinemäßigen Einsatz strukturierter Interviews im klinischen Alltag. Da von Seiten der Behandelnden nicht selten Vorbehalte gegenüber strukturierten Interviews bestehen, da sie als zu zeitaufwendig oder belastend für den Patienten erlebt werden, widmen wir uns derzeit der Frage, wie eine höhere Akzeptanz strukturierter Interviews erzielt und eine häufigere Anwendung strukturierter Interviews erreicht werden kann.

Dank

Die Autoren danken allen Interviewerinnen für die Durchführung und Auswertung der Interviews sowie PD Dr. Ernst Hermann (Klinik Meisenberg), Prof. Dr. Franz Müller-Spahn (Universitäre Psychiatrische Klinik Basel) und Dr. Sascha Wendler (Klinik Nexus, Baden-Baden) für die Unterstützung des Forschungsprojektes. Das Projekt wurde von der «Freiwilligen Akademischen Gesellschaft» finanziell unterstützt (DPE2020).

Literatur

- American Psychiatric Association (Ed): Diagnostic and Statistical Manual of Mental Disorders, ed 4, revised (DSM-IV-TR). Washington, DC, American Psychiatric Association, 2000.
- Brown TA, Di Nardo PA, Barlow DH: Anxiety Disorders Interview Schedule for DSM-IV (ADIS-IV). New York, Graywind, 1994.
- Brown TA, Di Nardo PA, Lehman CL, Campbell LA: Reliability of DSM-IV anxiety and mood disorders: implications for the classification of emotional disorders. *J Abnorm Psychol* 2001;110(1):49–58.
- Cohen J: A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
- Di Nardo PA, Barlow DH: Anxiety Disorders Interview Schedule-Revised (ADIS-R). Albany, NY, Graywind, 1988.
- Di Nardo PA, Brown TA, Barlow DH: Anxiety Disorders Interviews Schedule for DSM-IV: Lifetime Version (ADIS-IV-L). Albany, NY, Graywind, 1994.
- Ehler U: Eine Psychotherapie ist immer nur so gut wie ihre Diagnostik. *Verhaltenstherapie* 2007;17:81–82.
- Fleiss JL: Statistical Methods for Rates and Proportions, ed 2. New York, Wiley, 1981.
- Hoyer J, Scholz D, Ruhl U, Wittchen HU: Patients' feedback after computer-assisted diagnostic interviews for mental disorders. *Psychother Res* 2006;16:357–363.
- In-Albon T, Suppiger A, Schlup, B, et al: Validität des Diagnostischen Interviews für Psychische Störungen. *Z Klin Psychol Psychother* 2008;37(1):33–42.
- Keller A: Die Klassifikation psychischer Störungen nach DSM-IV mit Hilfe des strukturierten diagnostischen Interviews (F-DIPS). Ruprecht-Karls-Universität Heidelberg, unveröffentlichte Dissertation, 2000.
- Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174.
- Margraf J, Schneider S, Spörkel H: Therapiebezogene Diagnostik: Validität des Diagnostischen Interviews bei Psychischen Störungen (DIPS). *Verhaltenstherapie* 1991;1(2):110–119.
- Margraf J, Schneider S, Ehlers A: Diagnostisches Interview bei Psychischen Störungen, ed 2. Heidelberg, Springer, 1994.
- Ohayon MM, Roberts RE: Comparability of sleep disorders diagnoses using DSM-IV and ICD classifications with adolescents. *Sleep* 2001;24(8):920–925.
- Sass H, Wittchen HU, Zaudig M, Houben I: Diagnostisches und Statistisches Manual psychischer Störungen, Textrevision (DSM-IV-TR). Göttingen, Hogrefe, 2003.
- Schneider S, Margraf J: Diagnostisches Interview bei Psychischen Störungen, ed 3 (DIPS für DSM-IV-TR). Heidelberg, Springer, 2006.
- Schneider S, Margraf J, Spörkel H, Franzen U: Therapiebezogene Diagnostik: Reliabilität des Diagnostischen Interviews bei Psychischen Störungen (DIPS). *Diagnostica* 1992;38(3):209–227.
- Schramm E, Riemann D: Internationale Klassifikation der Schlafstörungen. Weinheim, Psychologie Verlags Union, 1995.
- Suppiger A, In-Albon T, Hendriksen S, Hermann E, Margraf J, Schneider S: Acceptance of structured diagnostic interviews for mental disorders in clinical practice and research settings. *Behav Ther*, in press.
- Wittchen HU, Pfister H: DIA-X Interviews: Manual für Screening, Verfahren und Interview. Interviewheft Längsschnittuntersuchung (DIA-X lifetime); Ergänzungsheft (DIA-X), Interviewheft Querschnittuntersuchung (DIA-X 12 Monate); PC-Programm zur Durchführung des Interviews (Längs- und Querschnittsuntersuchung); Auswertungsprogramm. Frankfurt, Swets & Zeitlinger, 1997.
- Wittchen HU, Semler G: Composite International Diagnostic Interview (CIDI, Version 1.0). Weinheim, Beltz, 1990.
- Wittchen HU, Wunderlich U, Gruschwitz S, Zaudig M: SKID-I. Strukturiertes Klinisches Interview für DSM-IV Achse I: Psychische Störungen. Interviewheft und Beurteilungsheft. Eine deutschsprachige, erweiterte Bearbeitung der amerikanischen Originalversion des SCID-I. Göttingen, Hogrefe, 1997.
- Wittchen HU, Lachner G, Wunderlich U, Pfister H: Test-retest reliability of the computerized DSM-IV version of the Munich-Composite International Diagnostic Interview (M-CIDI). *Soc Psychiatry Psychiatr Epidemiol* 1998;33(11):568–578.
- Wittchen HU, Üstün TB, Kessler RC: Diagnosing mental disorders in the community. A difference that matters? *Psychol Med* 1999;29:1021–1027.
- World Health Organisation: Tenth Revision of the International Classification of Diseases, Chapter V (F): Mental and Behavioural Disorders (Including Disorders of Psychological Development). Clinical Descriptions and Diagnostic Guidelines. Genf, WHO, 1991.
- Yule GU: On the methods of association between two attributes. *J Roy Statist Soc* 1912;75:579–642.
- Zanarini MC, Skodol AE, Bender D, et al: The collaborative longitudinal personality disorder study: reliability of axis I and II diagnoses. *J Personal Disord* 2000;14(4):291–309.