Eur Arch Psychiatr Neurol Sci (1987) 236:214–222

European Archives of Psychiatry and Neurological Sciences

Test-Retest Reliability of a Standardized Psychiatric Interview (DIS/CIDI)*

Gert Semler[1], Hans-Ulrich Wittchen[1], Klaus Joschke[1], Michael Zaudig[2], Tobias von Geiso[2], Stephan Kaiser[1], Michael von Cranach[2], and Hildegard Pfister[1]

[1]Max-Planck-Institute for Psychiatry, Unit for Evaluation Research, Kraepelinstrasse 10, D-8000 München,
[2]Bezirkskrankenhaus, D-8950 Kaufbeuren, Federal Republic of Germany

Abstract. The reliability of DSM-III diagnoses using an expanded version of the Diagnostic Interview Schedule (DIS), called the Composite International Diagnostic Interview (CIDI), was evaluated by examining 60 psychiatric inpatients on a test-retest basis. Acceptable agreement coefficients of (kappa) 0.5 or above were found for all but two disorders: dysthymic disorder and generalized anxiety disorder. The subclassification of DSM-III affective disorders also revealed some discrepancies between the test and the retest interviews. When compared with results from earlier versions of the DIS, diagnostic reliability was found to have improved for the DSM-III anxiety disorders in particular. These improvements can possibly be attributed to some changes in the wording of the respective items of this section. Several reasons for lowered test-retest reliability are discussed.

Key words: Interview – Diagnosis – Affective disorders – Anxiety disorders – Schizophrenia

Introduction

Recently published, in part very promising data about the reliability and validity of the NIMH Diagnostic Interview Schedule (DIS)(Robins et al. 1981; Burnam et al. 1983; Anthony et al. 1985; Helzer et al. 1985; Wittchen et al. 1985) have stimulated considerable interest in the use of standardized diagnostic instruments in general and have encouraged further refinement and expansion of the DIS itself, such as the development of the expanded DIS version III, also called the Composite International Diagnostic Interview (CIDI). Because of the differing results and methodologies of DIS studies, however, clear judgments about the strengths and weaknesses of this particular standardized diagnostic interview are still difficult to make. Apart from a number of general problems with the reliability and validity studies done with the DIS, recent, more detailed discussions of DIS-related issues and controversies (Klerman 1985; Robins 1985; Burke 1986) have suggested that the variance in the results of these studies may reflect issues of design. Such issues would include, for example, the time interval between clinical examinations in test-retest studies, the temporal sequence and ordering of interviews, the characteristics of subjects under study (e.g., of inpatients, as opposed to subjects from the general population), and the differing professional backgrounds of the interviewers and extent of their respective clinical experience (Klerman 1985). Examined in this light, it becomes clear that only two of the studies report results on test-retest reliability in a strict sense, both being conducted with the earlier version II of the DIS. Thus, neither takes into account the many changes introduced in version III or its very recent expansion, the CIDI. In all other studies, reliability issues have either been confounded with issues of validity (e.g., the performance of lay interviewers is evaluated by comparison with that of clinical interviewers, Robins et al. 1981, 1982) or have been distorted (e.g., long intervals are permitted between test and retest interviews, which introduces the possibility of changes in the psychopathological state of the patients examined, Robins et al. 1981).

The aim of this paper was to analyze the test-retest reliability of the DIS approach both by using its latest version and by taking into account most of the critical methodological issues that have been raised concerning earlier studies. More specifically, we will address the following questions: (1) what is the test-retest reliability of DSM-III current and lifetime diagnoses when using the CIDI? (2) If there are discrepancies between the test and retest interviews on the diagnostic and symptom level, what do they mean? What is their source and significance? Finally, we compared our test-retest results with the results obtained by Robins et al. (1981) and Burnam et al. (1983).

Methods

The study was conducted at the Max-Planck-Institute for Psychiatry in Munich and at the County Hospital in Kaufbeuren, using the German translation (Semler 1983) of a preliminary version of the CIDI. The primary goal of the study was to determine the diagnostic test-retest reliability of the German translation of the DIS/CIDI under well-controlled experimental conditions. Thus, our results would potentially reflect the upper limits of the reliability obtainable with the DIS/CIDI approach, along with the value of the modifications introduced in the instrument since the earlier studies that used DIS versions II and III.

*This research was supported by the German Research Foundation (DFG). The study was done in close collaboration with members of the Task Force on Instrument Development, established within the framework of the joint WHO/ADAMHA Project on Diagnosis and Classification. Lee Robins, PhD, John Helzer, MD, John Wing, MD, Norman Sartorius, MD, and Jack Burke, MD, provided us with helpful comments. Parts of the reliability analysis were done under contract with the NIMH, Division of Biometry and Epidemiology.

Offprint requests to: G. Semler

The CIDI was developed by the Task Force on Instrument Development of the joint WHO/ADAMHA Project on Diagnosis and Classification. Using a single instrument it allows diagnostic assessment according to four different systems: (a) the Feighner Criteria, (b) the Research Diagnostic Criteria, (c) the Diagnostic and Statistical Manual of Mental Disorders version 3, and (d) ICD compatible classes derived from John Wing's Present State Examination (PSE) (Feighner et al. 1972; Wing et al. 1974; Spitzer et al. 1978; American Psychiatric Association 1980).

In its form and character, the CIDI is basically a version of the DIS (Robins et al. 1981), to which items adapted from the PSE (Wing et al. 1974) have been added to allow the derivation of many of the CATEGO classes. The interview is highly structured, permitting both physicians and nonphysicians to serve as interviewers following an intensive training program of at least 5 to 7 days. A sample page from the DIS/CIDI interview is given in Table 1.

Because the CIDI was originally developed for use in epidemiological studies to provide estimates of prevalence and

Table 1. Sample page from the DIS/CIDI:

"Operationalisation" of panic disorder-related questions in the CIDI (shortened and modified for this paper). In addition to codings for the presence of symptoms (NO = 1, YES = 5), the DIS/CIDI allows codings about how recent the symptom was (REC). There are six possible codes for time frames from "current" (= 1) to "lifetime" (= 6).

A. Could you tell me about one spell or attack like that?	RECORD: _____
B. Did your spell(s) ever seem to come on for no particular reason – without anything having happened that seemed to explain them?	NO REASON ...(ASK REC) .... 5 EXPLAINED BY SITUATION ....(GO TO C) 1
REC: When was the last time you had a spell that came on for no reason?	REC: 1 2 3 4 5 6
C. Have you ever had such a bad spell that you *had* to do something about it – like telephoning someone or leaving the room or house?	NO .....(GO TO C) .... 1 YES .....(ASK REC) ... 5
REC: When was the last time you had a spell that bad?	REC: 1 2 3 4 5 6

63. During one of your worst spells of suddenly feeling frightened or anxious or uneasy, did you ever notice that you had any of the following problems? During this spell: (READ EACH SYMPTOM AND CODE "YES" OR "NO" FOR EACH. REPEAT THE PHRASE "DURING THIS SPELL" FOR EACH AND CODE IN COLUMN I).

Column I	I NO	I YES	II RECENCY					
A. Were you *short of breath* – having trouble catching your breath?	1	5	1	2	3	4	5	6
B. Did *your heart pound?*	1	5	1	2	3	4	5	6
C. Were you *dizzy* or light-headed?	1	5	1	2	3	4	5	6
D. Did your *fingers or feet tingle?*	1	5	1	2	3	4	5	6
E. Did you have *tightness or pain in your chest?*	1	5	1	2	3	4	5	6
F. Did you feel like you were *choking or smothering?*	1	5	1	2	3	4	5	6
G. Did you feel *faint?*	1	5	1	2	3	4	5	6
H. Did you *sweat?*	1	5	1	2	3	4	5	6
I. Did you *tremble* or shake?	1	5	1	2	3	4	5	6
J. Did you feel *hot or cold flashes?*	1	5	1	2	3	4	5	6
K. Did *things* around you *seem unreal?*	1	5	1	2	3	4	5	6
L. Were you *afraid* either that *you might* die or that you might *act* in a *crazy* way?	1	5	1	2	3	4	5	6

Column II
IF ANY 5 COL. I A–L *AND* Q .62 REC = 1–5, ASK FOR EACH 5 IN A–L and CODE IN COL. II:
When was the last time you (had/were SX) during an attack or spell of feeling frightened or anxious?

64. How old were you the first time you had one of these sudden spells of feeling frightened or anxious	ENTER AGE & GO TO Q. 65.
65. Have you ever had three spells like this close together – say within a three-week period?	NO .....(SKIP TO Q. 66) .... 1 YES .....(ASK REC) 5
REC: Have you had three spells in three weeks since (MO/YR)?	NO 6

Table 2. DSM-III diagnoses covered by the CIDI (brackets indicate possible corresponding ICD-9 codes)

Schizophrenic disorders	Anxiety disorders
– Schizophrenia (295.X)	– Panic disorder (300.0)
– Schizophreniform disorder (295.X)	– Generalized anxiety disorder (300.0)
	– Simple phobia (300.2)
Affective disorders	– Social phobia (300.2)
– Major depressive disorder	– Agoraphobia (300.2)
– Single episode (296.1, 300.4, 309.1)	– Obsessive-compulsive disorder (300.3)
– Recurrent (296.1)	
– Bipolar disorder (296.2/3/4/5)	Somatization disorder (300.1/5/7/8)
– Atypical bipolar disorder (296.6)	Psychosexual dysfunction (302.7)
– Dysthymic disorder (300.4)	Anorexia nervosa (307.1)
– Manic episode (296.0)	Alcohol use disorder (303)
	Drug use disorder (304.X)
	Organic brain syndrome (290–294, 310.X)

incidence rates of mental disorders in the general population, it generates interview data concerning the subject's most recent experience of symptoms or episodes. These data become a base for algorithms that permit, by use of a computer program, computation of "lifetime" as well as more current, cross-sectional (6-month, 4-week, 2-week) diagnoses. The instrument also asks age of onset for almost all symptoms, thus providing information on the development and chronological sequence of different disorders.

The original English version of the CIDI was translated into German in three stages. First, the instrument was twice translated independently – once by a clinical psychologist with 2 years experience with the DIS and other diagnostic interviews (G Semler) and again, by a psychiatrist in clinical training (H-U Rupp). Second, the two resulting versions were compared, discussed by the research team (four psychiatrists and five psychologists), and reduced to a single version. Third, this latter version was circulated to a number of German-speaking specialists in the field of psychopathology, who were asked for proposals concerning the contents and wording of items. To ensure the translation's maximum fidelity to the original version, some final refinements were made after a meeting with the authors of the instrument (Lee Robins, John Wing, John Helzer).

Although the German translation covers all diagnostic areas included in the original version, not all of them were considered in the test-retest study (Table 2). For various reasons (duration of interview, acute psychiatric inpatients, issues of compliance), we elected to omit antisocial personality disorder, tobacco use disorder, ego-dystonic homosexuality, pathological gambling, posttraumatic stress disorder, and transsexualism.

Subjects

The sample consisted of 30 male (mean age: 35 years) and 30 female (mean age: 37 years) inpatients. Fifteen were recruited from the Inpatient Department of the Max-Planck-Institute for Psychiatry, and 45 from the County Hospital in Kaufbeuren. They all met the following inclusion criteria: (a) at least 16 years old, (b) had no organic brain syndrome, and (c) willing to participate. Of the original 70 patients who were considered for inclusion in the study, 4 refused the test interview, and 6 patients refused the retest interview. Table 3 gives the distribution of the principal ICD-9 diagnoses for all 60 patients according to the clinicians judgment at the time of the patients' discharge.

The selection of patients was not random. Patients were preselected by the ward staff according to the foregoing criteria, but we then paid special attention to covering a broad spectrum of the different mental disorders assessable with the CIDI. The sample included newly admitted patients as well as long stay hospital patients. With the exception of 7 patients hospitalized for more than 1 year, most subjects were interviewed within 2 months after admission. After having agreed to participate, each was informed about the reasons for being interviewed twice with the same instrument and was instructed to regard both interviews as independent – that is, to answer all questions comprehensively on both occasions and not to consider the second interview a continuation of the first.

Design. All patients were interviewed twice by different interviewers. In order to avoid variance due to possible changes in the patients psychopathological state, the time interval between the examinations was kept to a minimum – 1 to 4 days with a mean of 1.7 days. We also assumed that the degree of concordance in test-retest studies would be influenced, in part, by distribution and assignment of interviewers and subjects. Thus, we strove for balance: each of four (male) interviewers conducted 30 interviews – 15 with males and 15 with females, and each interviewer was test-retest partner for each of his three colleagues across 10 subjects – for 5 as the test interviewer, for the remaining 5 as the retest interviewer.

The setting in which the interviews took place was standardized as far as possible. Test and retest interviews were con-

Table 3. Distribution of the clinicians' ICD-9 principal diagnoses at the time of the patients' discharge ($n = 60$)

ICD-9	Diagnosis	*n*
291./292./293.	Organic psychoses	5
295.1/3/5/6	Schizophrenic psychoses	14
295.7	Schizoaffective psychosis	4
296.1	Manic-depressive psychosis, depressive type	3
296.2/3/4	Manic-depressive psychosis, circular type	8
297.X	Paranoid states	2
298.X	Other, nonorganic psychoses	2
300.0/2	Anxiety states; phobic states	2
300.4	Neurotic depression	4
301.X	Personality disorders	1
302.X	Sexual deviations	2
303	Alcohol dependence	4
304.X	Drug dependence	3
307.1	Anorexia nervosa	2
309.X	Adjustment reactions	3
317	Mild mental retardation	1
Total		60

ducted in the same room and at the same hour of the day, with only the interviewer and the subject present.

Characteristics of the interviewers. The interviews were administered by two psychiatrists in residency training and two psychologists. One of the psychiatrists (M Zaudig) was highly experienced in clinical psychiatry, the other (T von Geiso) was a first-year resident. Both psychologists (K Joschke, S Kaiser) had less than 1 years practical experience in the field of psychiatry. All interviewers had participated in a 9-day CIDI training program and had conducted five "live interviews" under supervision. In addition, a 2-day training session was held by Drs Robins, Helzer, and Wing. The interviewers knew neither the clinical ICD diagnoses of the patients nor the results of the companion interviews and the diagnostic algorithms used to analyze the CIDI.

Statistical analyses. To translate the interview data from the expanded DIS/CIDI into DSM-III diagnoses, a slightly modified computer program, initially developed for the DIS (Boyd et al. 1985) was applied. The additional diagnostic codes optionally available in this computer program e.g., with regard to exclusion rules and clinical severity, were not used for this study, with the exception of those relating to organic brain syndrome.

To measure concordance between raters, we used three different coefficients: overall percentage agreement, kappa coefficients (Cohen 1960), and Y coefficients (Yule 1912). Overall percentage agreement and kappa coefficients represent at present – despite some controversies (Zubin 1967; Spitznagel and Helzer 1985) – the most commonly used measures to describe the extent of agreement between raters. Our use of them therefore guarantees a high degree of comparability with other studies. The Y coefficient, recently rediscovered by Spitznagel and Helzer (1985), is similar to kappa, in that it is a chance-corrected measure of association. Unlike the kappa coefficient, however, it is to a large extent independent of base rates and thus offers a new perspective on how to handle the base rate problem in reliability studies. Where A, B, C and D are the 2×2 classification frequencies, Y is defined as follows:

$$Y = \frac{\sqrt{AD} - \sqrt{BC}}{\sqrt{AD} + \sqrt{BC}}$$

It has to be noted, however, that when a single cell of the frequency table becomes 0, interpretation of Y becomes a problem, because in this case, Y reaches the endpoints of its range (+1 or −1). This would indicate perfect association or no association at all, even though percentage agreement is actually neither 0% nor 100%. Following the strategy recommended by Spitznagel and Helzer (1985), we handled the problem of zero cell frequencies in our contingency tables by using the pseudo-Bayes estimation procedure, described in detail by Bishop et al. (1975).

A test of significance of kappa was performed using formulas initially suggested by Bartko and Carpenter (1976). We restricted ourselves, however, to a one-tailed test that considers only positive deviations from 0. Kappa values of 0.40 and above were defined as acceptable, kappa values of 0.70 and above as excellent (Burke 1986). The same standards were applied to Yule's Y values.

Results

Diagnostic test-retest reliability

Table 4 summarizes the test-retest reliability of DIS/CIDI lifetime diagnoses with regard to DSM-III. Concordance rates (overall percentage agreement, kappa coefficients, Yule's Y coefficients) are presented for all DIS/DSM-III diagnoses found to be present in at least one patient. As indicated earlier, we did not use the DIS/CIDI diagnostic exclusion criteria for DSM-III that are optional in the computer program, nor did we include missing values in the calculations. Since some patients denied specific questions on the CIDI (e.g., questions asking for sexual experiences) or were too tired to complete the full interview, the diagnostic algorithms could not be applied to all interviews. Therefore, the frequencies in the 2×2 cross tabulation for the specific diagnoses do not sum up to $n = 60$.

Overall percentage agreement varied between 72% and 98%. Generalized anxiety disorder and phobias showed relatively low percentage agreement, whereas somatization disor-

Table 4. Concordance rates for CIDI DSM-III lifetime diagnoses (without DSM-III exclusion rules)

DSM-III diagnosis	2. Interview		% Agreement	k	Y
	−	+			
1. Interview −	A	B			
+	C	D			
Organic brain syndrome	[b]		95	0.70**	–
Manic episode	49 / 1	3 / 3	93	0.56**	0.75
Major depressive disorder	32 / 5	4 / 16	84	0.66**	0.67
Dysthymic disorder	51 / 3	1 / 2	93	0.47*	0.71
Alcohol use disorder	35 / 4	1 / 15	91	0.79**	0.84
Drug use disorder	43 / 2	2 / 7	93	0.73**	0.79
Schizophrenic disorder	43 / 3	3 / 6	89	0.60**	0.69
Schizophreniform disorder	50 / 2	1 / 2	95	0.54*[a]	0.75
Obsessive-compulsive disorder	47 / 1	2 / 4	94	0.70**	0.81
Phobic disorder	34 / 8	2 / 11	81	0.57**	0.66
Panic disorder	51 / 2	0 / 6	97	0.84**	0.86[c]
Generalized anxiety disorder	27 / 7	9 / 14	72	0.41**	0.42
Somatization disorder	57 / 0	1 / 1	98	0.66*[a]	0.72[c]
No disorder	36 / 8	1 / 6	82	0.48**	0.68

* $P < 0.05$
** $P < 0.01$
[a] Base rate < 10%
[b] 3×3 table with 7 cases out of 56, 3 discrepancies
[c] Pseudo-Bayes estimation

der and panic disorder showed the highest values. With respect to k coefficients, the lowest kappa value was again found for generalized anxiety disorder ($k = 0.41$), whereas the highest kappa value was found for panic disorder ($k = 0.84$). In Table 4, k values for schizophreniform disorder and somatization disorder are market by an "a", because the number of positives for these diagnoses was below the generally accepted 10% level for the calculation of kappa coefficients. For all other diagnoses, k values were highly significant, except for dysthymic disorder, where k was significant at the 5% level only. Because of the low base rates for some disorders, the calculation of Yule's Y resulted in slightly higher values overall, as compared with kappa. Yule's Y ranged from 0.42 for generalized anxiety disorder to 0.86 for panic disorder. Because the calculation of Yule's Y coefficients requires dichotomized variables, the trichotomized diagnostic information in the program for organic brain syndrome (absent versus definitive mild or severe versus uncertain) precluded calculation of concordance rates for this disorder. Concordance rates for all other diagnoses were calculated in terms of absent versus present.

Subclassification of anxiety and depressive disorders

Although the base rates for most disorders were very small, the subgroups of some disorders were also analyzed in more detail. With respect to phobias (Table 5), relatively high concordance rates were found for agoraphobia, whereas remarkably lower kappa and Yule's Y values were obtained for social phobia and for simple phobia, for which the kappa statistic did not even reach the 5% level of significance.

Relatively poor agreement was also found for the subgroup of affective disorders (Table 6), varying between $k = 0.47$ for bipolar disorder and no more than chance agreement for atypical bipolar disorder. Comparable problems were observed for substance use disorders. Despite relatively high overall agreement, several subcategories, such as amphetamine use disorder and opioid use disorder, showed very low k and Y values (range of k values: -0.02–1.00).

Test-retest reliability with respect to different time frames

Unlike other diagnostic interviews, the CIDI optionally allows the assignment of diagnoses to various "time frames". Thus, we could determine whether symptoms or syndromes that might have occurred years before the examination would be assessed with lower reliability than the more recent symptoms and syndromes. The instrument generates data as to whether a person has suffered from a disorder "within the last 2 weeks", "within the last month", "within the last 6 months", "within the last year" or "more than 1 year ago". However, the task of assigning DSM-III diagnoses to the various resulting time frames is not a simple one. For example, the DSM-III diagnosis of depressive episode requires the presence of at least four of eight specific symptoms, each having been experienced nearly every day for at least 2 weeks. One cannot merely ask about the latest occurrence of "depressed mood". The CIDI thus defines a depressive episode specifically as "a period of 2 weeks or more when you had some of these problems and also felt depressed...", which does not perfectly match the DSM-III criteria. A similar procedure is required in order to assess the most recent occurrence of a manic episode and panic disorder.

Table 5. Concordance rates for DSM-III lifetime diagnoses. Subclassification of phobic disorders

DSM-III diagnosis	2. Interview −	2. Interview +	% Agreement	k	Y
	1. Interview − A	B			
	1. Interview + C	D			
Agoraphobia	40	0	88	0.65**	0.75[b]
	7	9			
Simple phobia	44	3	82	0.19	0.34
	7	2			
Social phobia	46	2	89	0.44*	0.61
	4	3			
Agoraphobia with or without panic attacks	[a]		83	0.55**	–

* $P < 0.05$
** $P < 0.01$
[a] 3 × 3 table with 17 cases out of 59, 10 discrepanices
[b] Pseudo-Bayes estimation

Table 6. Concordance rates for CIDI DSM-III lifetime diagnoses. Subclassification of affective disorders

DSM-III diagnosis	2. Interview −	2. Interview +	% Agreement	k	Y
	1. Interview − A	B			
	1. Interview + C	D			
Bipolar disorder	[b]		91	0.47*	–
Major depressive disorder, Single episode	49	2	91	0.40	0.60
	3	2			
Major depressive disorder, Recurrent	40	5	79	0.33*	0.41
	7	5			
Atypical bipolar disorder	52	0	93	0.0[a]	0.0
	4	0			

* $P < 0.05$
[a] Base rate < 10%
[b] 4 × 4 table with 7 cases out of 56, 5 discrepanices

Results of the comparison between test and retest interviews with respect to 4-week, 6-month, and 12-month diagnoses are given in Table 7. The computations were based on 2 × 2 contingency tables with illness present or absent during the chosen time interval – that is, an illness having an earlier offset was considered absent. To avoid problems in interpreting the respective results, diagnoses with base rates less than or equal to 10% are indicated with an "a". Organic brain syndrome was omitted, because that disorder, by definition, must be diagnosed as present once the appropriate criteria are met. On the other hand, schizophrenic disorder was considered in the calculations even though DSM-III requires some signs of the illness to be continuously present. Recency of schizophrenia is thus defined here as the latest occurrence of an active phase of schizophrenia.

Table 7 illustrates how interviewer agreement changes along with various time criteria with regard to the latest occurrence of the disorder. Concordance rates displayed a slight tendency for percentage rates to increase from the lifetime to the 4-week time frame, whereas k and Y values tended in the

Table 7. Diagnostic concordance in different "time frames" (4 weeks, 6 months, 12 months) as compared with lifetime diagnoses

DSM-III diagnosis	Time frames											
	4 Weeks			6 Months			12 Months			Lifetime		
	(%)	k	Y	(%)	k	Y	(%)	k	Y	(%)	k	Y
Panic disorder	97	0.73**[a]	0.79[b]	98	0.88**	0.87[b]	97	0.78**	0.82[b]	97	0.84**	0.86[b]
Generalized anxiety disorder	79	0.33*	0.41	77	0.40**	0.46	74	0.36**	0.39	72	0.41**	0.42
Phobic disorder	87	0.59**	0.75	80	0.45**	0.67	78	0.41**	0.56	81	0.57**	0.66
Obsessive-compulsive disorder	96	0.65**[a]	0.72[b]	94	0.64**	0.79	94	0.64**	0.79	95	0.70**	0.81
Schizophrenic disorder	86	0.25	0.42	87	0.46**	0.59	87	0.46**	0.59	89	0.60**	0.69
Major depressive disorder	84	0.52**	0.64	81	0.53**	0.58	81	0.55**	0.58	84	0.66**	0.67
Manic episode	93	−0.04[a]	−0.001[b]	93	0.46*	0.70	95	0.64**	0.73[b]	93	0.56**	0.75
Alcohol abuse	96	0.0[a]	0.0	91	0.40	0.60	93	0.67**	0.79	89	0.74**	0.77
Alcohol dependence	98	0.0[a]	0.0	95	0.54*[a]	0.75	96	0.81**	0.84[b]	93	0.81**	0.83
Drug abuse	100	0.0[a]	0.0[a]	98	0.66*[a]	0.71	98	0.79**[a]	0.81[b]	93	0.67**	0.76
Drug dependence	98	0.66*[a]	0.71	94	0.37[a]	0.67	93	0.46*	0.66	94	0.79**	0.85

* $P<0.05$
** $P<0.01$
[a] Base rate < 10%
[b] Pseudo-Bayes estimation

opposite direction. A dramatic drop in k and Y values, however, was only obtained in schizophrenia. In the case of phobias, the 4-week classification displayed higher values than the lifetime classification.

Sources of variance

In order to identify possible reasons for discrepancies observed between the test and the retest interviews, we examined the reliability in more detail on diagnostic and symptom levels. Only some diagnostic sections (those with relatively low kappa coefficients and at least 10% positives) and only lifetime diagnoses were taken into account for this analysis.

Schizophrenic disorder. Table 4 shows that at least one of the interviewers found a schizophrenic disorder to be present in 12 out of 55 patients. In 6 of those 12 patients both interviewers agreed; in the other 6 they disagreed about the presence of the disorder. A closer inspection of the interview protocols revealed that the DSM-III "A" criterion for schizophrenia was scored concordantly by both interviewers in 4 of the 6 discrepant cases. Additional criteria, however, such as age of onset, prodromal symptoms, and residual symptoms, were assigned only in one of both interviews. This was confirmed by findings from test-retest comparisons on the item level, which indicated generally higher concordance rates for the core symptoms as set forth in the "A" criterion of schizophrenia than for the additional symptom questions that would establish the other criteria. No indication was found that the lower reliability was due to unreliable assessment of the 6-month criterion.

Major Depressive Disorder. This diagnosis was assigned to 25 of the 57 patients, 9 of them being diagnosed discrepantly. Most of these discrepancies were clearly due to difficulties in determining what should be regarded as the "worst period". Information on the "worst period", which is defined in the DIS/CIDI as the period with the largest number of symptoms, is elicited by a summary question at the end of the depression section. As indicated earlier, the interviewer determines by this information whether the subject has ever had a depressive episode by assessing the number of symptoms (DIS/CIDI symptom groups) present at the time of the "worst period". Comparisons of discrepancies on the diagnostic level showed that 6 of the 9 patients diagnosed discrepantly did report depressive episodes in both interviews, but picked out different "worst periods" indicated by different ages. For all 6 patients, a sufficient number of symptoms (more than three for a DSM-III diagnosis of major depression) were met in only one of the two interviews. This might indicate that the subject implicitly has different criteria for the "worst episode" than is assumed by the DIS/CIDI.

Manic episode. Surprisingly, discrepancies between test and retest interviews in the mania section could not be explained by discrepant "worst period" assessment, although manic episodes are assessed in the same way as depressive episodes. Only one out of four discrepancies could be related to a problem of determining the worst episode. Discordance with respect to the remaining three patients was due both to patients' explicit denial of symptoms or to patients' attribution of the respective manic symptoms to periods of alcohol abuse. It may be possible that patients remember manic episodes more easily than depressive episodes, simply because manic episodes occur less frequently.

Phobic disorders. Agreement about the lifetime presence of phobic disorders was found in the case of 11 patients; disagreement was found in the case of 10 patients. Of the 10 diagnostic disagreements, 8 were the result of the patients' having denied in the retest interview symptoms that they had already reported in the initial interview. Only 2 of the disagreements were the result of a denial occurring in the initial interview.

General remarks on methodology

The latter results indicated a more general problem in the application of a test-retest design to comprehensive and complicated standardized interviews such as the DIS/CIDI. Patients who find the instrument's probe and recency questions boring

Table 8. Diagnostic discordance: number of unconfirmed first and second interview diagnoses

DSM-III diagnosis	Number of diagnoses in the	
	First interview	Second interview
Schizophrenic disorder	3	3
Schizophreniform disorder	2	1
Major depressive disorder	5	4
Dysthymic disorder	3	1
Manic episode	1	3
Alcohol use disorder	4	1
Drug use disorder	2	2
Obsessive-compulsive disorder	1	2
Phobic disorder	8	2
Panic disorder	2	0
Generalized anxiety disorder	8	6
Somatization disorder	0	1
Total	39	26

in the initial interview may attempt to avoid these questions in the retest interview by simply denying symptoms already reported in the first interview. This was substantiated by test-retest comparisons on item level, which showed an overall significant decrease of reported symptoms per patient from the first interview ($n = 35$ symptoms) to the second ($n = 28$ symptoms), as well as a decrease in the average interview duration from 102 min to 87 min. In accordance with these results, the number of diagnoses assigned in the first interview ($n = 166$) was about 10% larger than the number assigned in the second interview ($n = 151$). However, the decrease in the number of resulting diagnoses in the second interview was not equally distributed in all diagnostic areas. Table 8 shows a balanced relation of the number of unconfirmed diagnoses resulting from either the first or the second interview for schizophrenia and depression, whereas other diagnoses, such as phobias and alcoholism, showed a considerable drop from the first to the second interview.

Discussion

Before discussing and judging the DIS/CIDI approach more generally, we should emphasize again that the primary aim of our study was to provide test-retest reliability data on the German translation of the DIS/CIDI derived in a well-controlled experimental study. To achieve this, special attempts were made to reduce as much as possible method-related sources of variance. Thus, the time interval between the examinations was kept very short, and the interview situation was standardized to a great extent, taking into account the time of the day the interview was conducted as well as room conditions. In addition, only four interviewers were used, and each conducted the same number of interviews. Unlike the test-retest study by Burnam et al. (1983), the study was conducted in an inpatient setting, where more severe disorders can be found together with a generally higher number of lifetime diagnoses. Moreover, special attempts were made to identify possible sources for low test-retest reliability of items, in order to make suggestions for improvement of the instrument (Wittchen, unpublished report to the NIMH). Because of these specific methodological characteristics, our results, as indicated earlier, probably reflect the upper limits of the reliability obtainable with the DIS/CIDI approach. Thus, they should not necessarily be regarded as representative of the overall reliability of the DIS/CIDI approach in epidemiological or other clinical settings.

Diagnostic test-retest reliability

In this study, the majority of DIS/DSM III diagnoses were reliably assessed in the test-retest format. Taking a kappa value of 0.50 as acceptable agreement, only two disorders had agreement coefficients below this criterion: generalized anxiety disorder and dysthymic disorder. Whereas a lower k value for dysthymic disorder might be influenced by the low base rate of this disorder in our sample – as indicated by the high Yule's Y value – a number of obvious diagnostic disagreements between the first and the second interviews were found for generalized anxiety disorder. Furthermore, although the overall classification of phobic disorders yielded acceptable levels of reliability ($k = 0.57$), there is evidence that simple phobias cannot be assessed reliably. Given that in all previous studies similar results were obtained, as indicated by kappa values of below. 0.5 (Robins et al. 1981; Burnam et al. 1983), this seems to be a robust finding. Similar problems of a reliable subclassification were identified for affective disorders. No agreement between the test and the retest interviews was found for the diagnoses of atypical bipolar disorder; the division of single and recurrent episodes of major depression was subject to some disagreement between the test and the retest interviews.

In comparison with the test-retest study by Burnam et al. (1983), who used the Spanish version of the DIS, slightly better results were found for most diagnostic sections. The same holds true for the data of Semler and Wittchen (1983) about the test-retest reliability of version II of the DIS. Our results are quite similar to the "lay interviewer/clinicians" comparison by Robins et al. (1982), with two exceptions: (1) the agreement coefficient for panic disorder was $k = 0.84$ in our study as compared with 0.40 in the study by Robins et al., representing a considerable improvement, and (2) there was a slight improvement in agreement for dysthymic disorder. Both improvements could possibly be attributed to a better formulation of the respective items to assess the diagnostic criteria for these diagnoses. It should also be emphasized that although a high number of psychotic inpatients (almost 40%) were included in our study, acceptable test-retest coefficients for schizophrenic and schizophreniform disorders were found. This indicates that the DIS/CIDI approach can be used with more severely disturbed psychiatric inpatients, when the instrument is used by trained clinicians.

Sources of test-retest variance

With regard to sources of low reliability, a few reasons became evident. One was the significant drop in the number of positively answered and coded symptom questions in the retest interviews. This drop occurred primarily in the section assessing anxiety disorders, especially in the section for phobic disorders. This result was primarily design-related and not directly caused by the instrument, in that the patients had to an-

swer the interview questions twice within a rather short time period of 1 to 4 days; however, the finding may indicate that some patients quickly learn during the interview how to avoid the strict and complicated – but sometimes boring – probe questions. It is surprising, nevertheless, that this source of variance was found only with regard to anxiety disorders and alcoholism.

A second source of variance relevant for schizophrenia, schizophreniform disorders, and the subclassification of affective disorders were the complex composite diagnostic criteria. Whereas some DIS/DSM-III diagnoses are derived in a clear and simple way, once the key symptoms are assessed (e.g., panic disorder and obsessive-compulsive disorder), multiple criteria are needed for some other diagnoses. It was shown that many of the diagnostic discrepancies observed resulted from attempts to assemble composite diagnostic criteria, that is, in situations where age of onset, duration, and impairment criteria and information obtained from questions about required symptoms are all needed together for a positive diagnosis. Another major source of discrepancies was the process of determining the "worst episode" criterion, which is necessary to ascertain the existence of a depressive episode. Although the interview questions assumed that the "worst period" is operationalized as the period with the largest number of symptoms present, patients themselves implicitly tend to use different criteria – possibly subjective criteria of having *felt* "worst". Restructuring this important part of the interview might improve the reliability in this section considerably.

Time frames

With regard to the issue of lifetime versus more current diagnoses, a slightly confusing picture emerged. At first glance, the reliability of the DIS/CIDI with respect to most of the current diagnoses seemed to be considerably lower than for lifetime diagnoses. When analyzing this result more carefully, however, it became clear that differences between concordance rates did not necessarily indicate different reliability. First of all, it has to be considered that in some diagnostic sections there was a dramatic drop in base rates for the 4-week time frame. Therefore, the respective kappa values should be interpreted very cautiously. Yule's Y values, on the other hand, which are supposed to be less sensitive to low base rates, were derived for many diagnostic sections using the pneudo-Bayes estimates of the cell frequencies instead of the observed frequencies in order to deal with "sampling zeros". It is doubtful, hence, whether small differences in Yule's Y values between the four time frames compared indicated true differences with regard to their reliability. Based on these restrictions, differences in the rates between 6-month, 12-month, and lifetime diagnoses might indicate overall only minor differences of the test-retest reliability calculated for lifetime and more current diagnoses. With regard to specific diagnostic categories, there were indications that concordance rates decreased for those disorders having episodic characteristics. In the case of an "active phase of schizophrenia", this trend was rather pronounced, whereas concordance rates for long-lasting, "chronic" disorders, such as phobias or generalized anxiety disorder, remained on the same level or even increased. Further investigations in larger samples and with higher base rates are required to confirm these findings. Results on the reliability of time-related criteria in the DIS/CIDI, an issue very closely related to that dealt with in this paper, will be presented in a later publication (Wittchen et al. in preparation).

Conclusions

After subjecting the DIS/CIDI to a very stringent test-retest design, we can summarize that with one minor exception (generalized anxiety disorder), all major DIS/DSM-III disorders can be assessed with at least an acceptable reliability, so as not to constrain the validity of the diagnoses. Compared with earlier studies conducted with the DIS (Burke 1986), some improvements of the most recent DIS/CIDI version were demonstrated. These positive results are encouraging, given the attempt of the DIS/CIDI version used in this study to score PSE-compatible diagnostic classes as well.

References

American Psychiatric Association Committee on Nomenclature and Statistics (1980) Diagnostic and Statistical Manual of Mental Disorders, 3rd edn. American Psychiatric Association, Washington, DC

Anthony JC, Folstein M, Romanoski AJ, von Korff MR, Nestadt GR, Chadal R, Merchant A, Brown H, Shapiro S, Kramer M, Gruenberg EM (1985) Comparison of the lay diagnostic interview schedule and a standardized psychiatric diagnosis. Arch Gen Psychiatry 42:667–675

Bartko JJ, Carpenter WT (1976) On the methods and theory of reliability. J Nerv Ment Dis 163:307–317

Bishop YMM, Feinberg SE, Holland PW (1975) Discrete multi-variate analysis: Theory and practice. MIT Press, Cambridge

Boyd JH, Robins LN, Holzer III, CE, von Korff M, Jordan KB, Escobar JI (1985) Making diagnoses from DIS data. In: Eaton WW, Kessler LG (eds) Epidemiologic field methods in psychiatry. The NIMH epidemiologic catchment area program. Academic Press, Orlando, San Diego, New York, London, Toronto, Montreal, Sydney, Tokyo, pp 209–231

Burke JD (1986) Diagnostic categorization by the Diagnostic Interview Schedule (DIS): A comparison with other methods of assessment. In: Barrett J, Rose RM (eds) Mental disorders in the community. The Guilford Press, New York, pp 255–285

Burnam NA, Karno M, Hough RL, Escobar JI, Forsythe AB (1983) The Spanish Diagnostic Interview Schedule. Reliability and comparison with clinical diagnoses. Arch Gen Psychiatry 40:1189–1196

Cohen J (1960) A coefficient of agreement of nominal scales. Educ Psychol Meas 20:37–46

Feighner JP, Robins E, Guze SB, Woodruff RA, Winokur G, Munoz R (1972) Diagnostic criteria for use in psychiatric research. Arch Gen Psychiatry 26:57–63

Helzer JE, Robins LN, McEvoy LT, Spitznagel EL, Stoltzman RK, Farmer A, Brockington IF (1985) A comparison of clinical and Diagnostic Interview Schedule diagnoses: Physician reexamination of lay-interviewed cases in the general population. Arch Gen Psychiatry 42:657–666

Klerman GL (1985) Diagnosis of psychiatric disorders in epidemiologic field studies. Arch Gen Psychiatry 42:723–724

Robins LN (1985) Reflections on testing the validity of psychiatric interviews. Arch Gen Psychiatry 42:918–924

Robins LN, Helzer JE, Croughan J, Ratcliff KS (1981) National Institute of Mental Health Diagnostic Interview Schedule: Its history, characteristics and validity. Arch Gen Psychiatry 38:381–389

Robins LN, Helzer JE, Ratcliff KS, Seyfried W (1982) Validity of the Diagnostic Interview Schedule, Version II: DSM-III diagnoses. Psychol Med 12:855–870

Semler G (1983) Deutsche modifizierte Version des Composite International Diagnostic Interview (CIDI). Max-Planck-Institut für Psychiatrie, München

Semler G, Wittchen H-U (1983) Das Diagnostic Interview Schedule: Erste Ergebnisse zur Reliabilität und differentiellen Validität der deutschen Fassung. In: Kommer D, Roehrle B (eds) Gemeindepsychologische Perspektiven (3), Köln, pp 109–117

Spitzer RL, Endicott J, Robins E (1978) Research diagnostic criteria. Rationale and reliability. Arch Gen Psychiatry 35:773–782

Spitznagel EL, Helzer JE (1985) A proposed solution to the base rate problem in the kappa statistic. Arch Gen Psychiatry 42:725–728

Wing JK, Cooper JE, Sartorius N (1974) The description and classification of psychiatric symptoms: An instruction manual for the PSE and CATEGO system. University Press, London

Wittchen H-U, Semler G, von Zerssen D (1985) A comparison of two diagostic methods. Arch Gen Psychiatry 42:677–684

Yule GU (1912) On the methods of measuring association between two attributes. J Roy Statist Soc 75:581–642

Zubin J (1967) Classification of the behavior disorders. Ann Rev Psychol 18:373–401

Received August 22, 1986