

Monash Interview for Liaison Psychiatry (MILP)

Development, Reliability, and Procedural Validity

DAVID M. CLARKE, M.B.B.S., PH.D., F.R.A.N.Z.C.P.

GRAEME C. SMITH, M.D., F.R.A.N.Z.C.P.

HELEN E. HERRMAN, M.D., F.R.A.N.Z.C.P., DEAN P. MCKENZIE, B.A.

The Monash Interview for Liaison Psychiatry (MILP) is a structured interview designed for use with patients who have physical and psychiatric comorbidity. Linked to a computerized diagnostic algorithm, the MILP is able to establish diagnoses according to DSM-III-R, International Classification of Diseases–10th Edition (ICD–10), and DSM-IV criteria, as well as a range of other criteria relevant to consultation-liaison psychiatry. Interrater reliability was assessed with 54 joint interviews, in which the mean kappa for agreement of items was 0.83 and of diagnoses was 0.68. Comparative procedural validity was tested against DSM-III-R decision-tree diagnoses, ICD–10 checklist diagnoses, and Structured Clinical Interview for DSM-III-R interview diagnoses on another sample of 54 patients. Mean kappas for these comparisons were 0.61, 0.56, and 0.31, respectively. As predicted, the MILP more fully covered the spectrum of somatizing disorders, compared with the other methods for establishing diagnoses.

(Psychosomatics 1998; 39:318–328)

Descriptive psychopathology has been the cornerstone of twentieth-century psychiatry, and classification its preoccupation in recent decades. However, the emphasis on measurement and reliability of diagnosis has tended to overshadow the importance of validity and utility of classifications.^{1,2} This emphasis is particularly evident in the field of consultation-liaison

psychiatry, in which current systems of classification poorly describe the nature, range, and etiological understanding of the psychiatric pathology observed. In depressed patients with a physical illness, for example, there is always a psychosocial stressor, and “organicity” frequently cannot be excluded. In this situation, etiological statements are often overly simplistic and reductionist.^{3,4} Further, the range of depressive phenomena seen in the context of physical illness is different from that seen in general psychiatric practice and subthreshold diagnoses are significant.^{5,6} Despite anxiety being extremely common in the physically ill, diagnoses of anxiety disorders are made infrequently,⁷ due to the difficulty of defining what is pathological⁸ and to the tradition of hierarchy.⁹ Somatization, a common phenomenon,¹⁰ is poorly represented by standard classifications,¹¹ and the attribution

Received January 16, 1998; accepted January 30, 1998. From the Consultation-Liaison Psychiatry Research Unit, Monash University Department of Psychological Medicine, Monash Medical Centre, Melbourne, Australia. Dr. Herrman is from the University of Melbourne, and Hospital and Community Psychiatry Service, St. Vincent's Hospital, Melbourne. Address reprint requests to Dr. Clarke, Monash University, Department of Psychological Medicine, Monash Medical Centre, 246 Clayton Road, Clayton, Victoria 3168, Australia. e-mail: david.clarke@med.monash.edu.au.

Copyright © 1998 The Academy of Psychosomatic Medicine.

of cause and meaning to physical symptoms is always difficult. For all these reasons, there is a need to examine the psychopathology of physical-psychiatric comorbidity without the assumptions made by classifications of mental disorders developed in other populations. Because most structured interviews to date have been developed around the current classificatory systems, these interviews are limited in their usefulness for this task.

We describe a new structured interview, the Monash Interview for Liaison Psychiatry (MILP), designed for the study of psychiatric disorders in the physically ill. The validity of the interview is based on its inquiry of a range of phenomena appropriate to the setting and its emphasis on a careful consideration of the attribution of cause. Compared with the Diagnostic Interview Schedule (DIS)¹² and its descendants, the Composite International Diagnostic Interview¹³ and Structured Clinical Assessment for Neuropsychiatry,¹⁴ the MILP covers a wider range of mood disorders and somatoform disorders and is administered by interviewers competent to make the clinical judgments required. Compared with the Structured Clinical Interview for DSM-III-R (SCID),¹⁵ the MILP is more comprehensive in inquiry, in that it minimizes the use of screening questions and "skips," and covers a wider range of disorders.

DESCRIPTION OF THE MILP

The MILP is a structured interview that takes between 60 and 90 minutes to administer. It includes a systematic symptom inquiry, with the emphasis on current state, defined as being during the past month, although the duration of any symptom is recorded for the application of diagnostic criteria. For symptoms present, a judgment of attribution of cause is coded as one or more of the following: physical illness or injury, medication, drugs or alcohol, psychogenic or "unexplained." Guided by the interview protocol, the interviewer makes this judgment after asking the subject a number of questions, inquiring of the medical staff if necessary. With the assistance of computerized data entry and

scoring algorithm,¹⁶ the interview generates diagnoses according to DSM-III-R,¹⁷ DSM-IV,¹⁸ and International Classification of Diseases (ICD)-10th Edition¹⁹ classifications. In addition, the MILP allows inclusion of a range of other diagnoses relevant to consultation-liaison psychiatry. These include the Stewart *et al.* criteria for depression²⁰; the Endicott criteria²¹; the concepts of atypical depression²²; Newcastle endogenous depression²³; grief²⁴; and diagnoses of "subthreshold" disorders, such as the DSM disorders "not otherwise specified (NOS)," the ICD disorders "unspecified," and "abridged somatization."²⁵ The interview has been acceptable for patients with physical illness but cannot be administered to patients with significant cognitive impairment. Diagnoses of organic mental disorders are not made.

RELIABILITY

Methods

To examine interrater reliability, an observed interview model was chosen, as used in other similar studies.²⁶⁻²⁸ This method removes information and occasion variance but requires a commitment on the part of raters to remain independent in their assessments.^{29,30} Fifty-four joint interviews were conducted by the first author (DMC), a psychiatrist, together with one of two psychologists, raters alternating between interviewing and observing. The sample size was chosen to be consistent with other studies.^{26,28,31-33} Patients were selected from the medical and surgical wards of Monash Medical Centre, a university-affiliated, suburban general hospital. During the period of study (1992-1993), all available patients were screened upon admission with the 36-item General Health Questionnaire³⁴ scored in the "chronic" manner,³⁵ using a 20/21 cutoff.³⁶ To create a "clinical" sample, the patients scoring above the cutoff were invited to participate until the desired sample size was achieved. The mean age \pm standard deviation (SD) of the patients interviewed was 46.0 \pm 18.0 years; 57% were male. Following the interview, data were en-

tered, and the diagnoses were generated by computer. Agreement between the raters was determined for both diagnoses and items. The primary measure of agreement used was the kappa statistic, which gives a chance-corrected measure of agreement between two raters.³⁷⁻³⁹ Because kappa is unstable in the presence of low base rates, it is not reported in instances in which the base rate was less than 10% for items (set conservatively) and 5% for diagnostic categories (to include a wider range).⁴⁰ Overall agreement, which is not chance-corrected, is also reported, defined as agreed cases plus agreed noncases expressed as a percentage of total subjects.⁴¹ In line with suggested practice, kappas above 0.40 were taken to indicate at least “moderate” agreement, whereas kappas below 0.40 were considered unsatisfactory.⁴² All statistics were calculated by a FORTRAN program written by one of the authors (DMcK).

Results

Reliability measures for interview items are summarized in Table 1. The mean kappa for all items was 0.83. One symptom item had a kappa less than 0.04—“rapport with grief”—an item involving subjective clinical judgment. Two “attribution” items had kappas less than 0.04. Both these items had a large proportion of empty cells in the analysis table (94% and 92%), a situation that makes kappa unstable in a way similar to a low base rate.⁴³ The judgments of whether depressive and anxiety reactions were “in excess of what might be expected” (a criterion required for the diagnosis of adjustment disorder) also had kappas less than 0.40.

Mean kappa for agreement on DSM-III-R diagnoses was 0.68 (see Table 2). One diagnosis, anxiety disorder NOS, had a kappa less than 0.40. Mean kappa for DSM-IV diagnoses was 0.63. Anxiety disorder NOS and pain disorder had kappas less than 0.40. Mean kappa for ICD-10 diagnoses was 0.66. Depressive disorder unspecified and adjustment disorders with anxiety had kappas less than 0.40. Mean kappa for diagnoses not included in these three systems was

0.91. Overall mean kappa for diagnostic agreement was 0.68.

PROCEDURAL VALIDITY

The concern here is not the validity of the underlying constructs or diagnoses (construct validity, predictive validity etc.), but procedural validity, defined as “the extent to which the new diagnostic procedure yields results similar to the results of an established diagnostic procedure that is used as a criterion” (p. 595).⁴⁴ Because there is no established “gold standard,” comparisons were made with a number of other diagnostic procedures. It is, in a sense, a comparative procedural validity that is being measured. Three comparisons were made: 1) MILP vs. a DSM-III-R decision-tree diagnosis, 2) MILP vs. an ICD-10 checklist, and 3) MILP vs. the SCID.

Methods

Patients were screened and recruited, as for the reliability study. One day per week, one pa-

TABLE 1. Interrater reliability for Monash Interview for Liaison Psychiatry (MILP)

Items	Kappa		
	Mean	Minimum	Maximum
Demographic, past and social history (13 items)	0.81	0.50	0.97
Symptoms (142 items) ^a	0.91	-0.20	1.0
Attribution of symptoms (31 items) ^b	0.60	0	0.91
Duration of symptoms (36 items) ^b	0.82	0.53	1.0
Currency of symptoms (21 items) ^b	0.84	0.66	1.0
Total (all items)	0.83		

^aItems with a base rate of less than 10% or greater than 90% have not been included in the analysis.

^bIf the symptom was “not present,” these questions were not answered. Items coded in less than 40% of subjects (22) have not been included in the analyses.

TABLE 2. Interrater reliability for diagnostic categories

	Base Rate ^a Rater 1 (%)	Base Rate Rater 2 (%)	Kappa	% Agreement
DSM-III-R diagnoses				
Major depression	16.7	16.7	0.87	96.3
Dysthymia	11.1	7.4	0.56	92.6
Adjustment disorder with depressed mood ^b	14.8	22.2	0.64	88.9
Depressive disorder not otherwise specified (NOS)	16.7	25.9	0.62	87.0
Generalized anxiety disorder	7.4	11.1	0.78	96.3
Agoraphobia	3.7	5.6		98.1
Obsessive-compulsive disorder	7.4	7.4	1.00	100
Adjustment disorder with anxious mood	5.6	9.3	0.46	92.6
Depersonalization disorder	13.0	7.4	0.70	94.4
Posttraumatic stress disorder (PTSD) syndrome ^c	16.7	18.5	0.81	94.4
Anxiety disorder NOS	20.4	22.2	0.39	79.6
Somatoform pain disorder	3.7	7.4		92.6
Undifferentiated somatoform disorder	16.7	20.4	0.76	92.6
Somatoform disorder NOS	11.1	13.0	0.74	94.4
Psychological factors affecting physical condition	11.1	9.3	0.49	90.7
Alcohol dependence or abuse	11.1	5.6	0.64	94.4
Other drug dependence or abuse	38.9	42.6	0.77	88.9
DSM-IV diagnoses				
Major depressive disorder	14.8	16.7	0.79	94.4
Dysthymic disorder	5.6	5.6	0.65	96.3
Adjustment disorder with depressed mood	27.8	40.7	0.56	79.6
Depressive disorder NOS	3.7	7.4	—	92.6
Generalized anxiety disorder	7.4	11.1	0.78	96.3
Agoraphobia	3.7	5.6	—	98.1
Obsessive-compulsive disorder	7.4	7.4	1.00	100
Adjustment disorder with anxiety	14.8	25.9	0.55	85.2
Depersonalization disorder	13.0	9.3	0.81	96.3
PTSD syndrome	16.7	18.5	0.81	94.4
Anxiety disorder NOS	11.1	11.1	0.06	81.5
Pain disorder with psychological factors	7.4	11.1	0.34	88.9
Undifferentiated somatoform disorder	5.6	11.1	0.64	94.4
Somatoform disorder NOS	25.9	24.1	0.56	83.3
Psychological factors affecting medical condition	9.3	9.3	0.56	92.6
Alcohol dependence or abuse	11.1	5.6	0.64	94.4
Other drug dependence or abuse	40.7	40.7	0.77	88.9
International Classification of Disease-10 diagnoses				
Depressive episode	11.1	13.0	0.74	94.4
Dysthymia	14.8	7.4	0.63	92.6
Adjustment disorder: depressive reactions	51.9	46.3	0.67	83.3
Depressive episode unspecified	5.6	18.5	0.24	83.3
Generalized anxiety disorder	16.7	18.5	0.68	90.7
Agoraphobia	1.9	3.7		98.1
Obsessive-compulsive disorder	7.4	7.4	1.0	100
Adjustment disorder with anxiety	9.3	7.4	0.39	90.7
Depersonalization: derealization syndrome	7.4	7.4	0.46	92.6
PTSD syndrome	14.8	16.7	0.79	94.4
Anxiety disorder unspecified	31.5	44.4	0.50	79.9
Neurasthenia	9.3	13.0	0.81	96.3
Persistent somatoform pain disorder	3.7	7.4	—	92.6
Undifferentiated somatoform disorder	22.2	22.2	0.68	88.9
Somatoform disorder unspecified	11.1	13.0	0.74	94.4

continued

TABLE 2. Interrater reliability for diagnostic categories (continued)

	Base Rate ^a Rater 1 (%)	Base Rate Rater 2 (%)	Kappa	% Agreement
Psychological and behavioral factors associated	7.4	3.7	—	92.6
Alcohol dependence or abuse	7.4	3.7	—	96.3
Other drug dependence or abuse	37.0	38.9	0.88	94.4
Other classifications				
Newcastle endogenous depression ^d	5.6	3.7	—	98.1
Endicott depression ^e	22.2	22.2	1.0	100
Stewart depression ^f	22.2	20.4	.94	98.1
Grief ^g	44.4	48.2	0.93	96.3
Abridged somatization disorder ^h	11.1	7.4	0.78	96.3
Mean			0.68	92.5

^aBase rate is the proportion of diagnosed cases expressed as a percentage of the sample ($n = 54$).

^bAdjustment disorders with mixed anxiety and depression are included in both adjustment disorder with depressed mood and adjustment disorder with anxious mood.

^cPTSD syndrome is posttraumatic stress disorder without the "exceptional stressor" criterion.

^dNewcastle criteria described by Carney et al.²³

^eDescribed by Endicott.²¹

^fDescribed by Stewart et al.²⁰

^g"Grief" defined by the authors with the following criteria: 1) mood disturbance understood to be a reaction to a loss and 2) at least one of the following: recurrent thoughts of loss; recurrent memories, dreams, or mental pictures; pangs of grief; yearning or pining; crying about the loss.

^hAs defined by Escobar and Canino.²⁵

tient was randomly selected from those already interviewed with the MILP. A SCID interview was performed within 48 hours of the MILP by a "blind" psychologist trained and satisfactorily rated in the use of the SCID but not otherwise involved in the study. Fifty-four interviews were completed, and this group of patients became the sample for each of the three comparisons. The mean age \pm SD of this group was 46.8 ± 16.7 years; 69% were male.

The data were entered into a computer, and DSM-III-R and ICD-10 diagnoses were generated. In addition, and before generating the MILP diagnoses, the interviewer created DSM-III-R diagnoses, aided by a computerized decision-tree program,^{45,46} and ICD-10 diagnoses by using the ICD-10 Research Criteria Checklist provided by the World Health Organization (Janca A, personal communication, 1992). Because no assumptions were made about the "gold standard," the measures of agreement used were kappa and overall agreement as previously described.³⁹ Because reliability assumes that error is random,⁴⁷ exact

McNemar's tests were used to look for systematic bias between ratings—specifically to determine the presence of any significant difference in base rates between the two methods of diagnosis under consideration.⁴³ Exact probabilities, based on the binomial distribution, were calculated by using the algorithm of Berry et al.⁴⁸ Results are only considered for the diagnostic categories that mostly had base rates greater than 5%. All diagnoses are "current."

Results

The results are presented in Table 3, Table 4, and Table 5. Table 3 shows good agreement for the MILP vs. DSM-III-R decision-tree diagnoses for all groups, except the somatizing disorders, in instances in which there was a tendency (McNemar's $P < 0.10$, NS) for the MILP to yield more diagnoses. Anxiety disorder NOS also had a kappa less than 0.40.

In the MILP vs. ICD checklist comparison, there was a similarly less than satisfactory agree-

ment for the somatizing disorders (Table 4), with significantly different base rates evident.

In the MILP vs. SCID analyses (Table 5), there were four groups with overall poor agreement and significant base rate differences.

DISCUSSION AND CONCLUSIONS

Reliability

As expected with a structured interview, interrater reliability results overall are satisfactory, influenced by the joint-interview study design, which minimized occasion and information variance. The mean kappa for the presence or absence of symptoms was above 0.80, and there was only one symptom in which there was poor agreement—rapport with grief. This item requires no question of the patient, and this assessment is based solely on the judgment made by the interviewer and observer. This lack of agreement may be partly explained by the differences in training and experience among the raters. The judgments about the attribution of

symptoms were also mostly satisfactory, with only two items having kappas less than 0.40, and both of these have high numbers of “empty cells.”

For the diagnoses, measures of agreement are mostly moderate or better. Only 4 out of 49 categories had kappas less than 0.40, and these were all the less specific categories—three NOS categories and an adjustment disorder. These diagnoses have few criteria and are dependent mostly on the exclusion of other diagnoses. It is possible that the disagreements in the other, for example, anxiety disorders, had a cumulative effect on the residual category of anxiety disorder NOS. This theory is suggested by the satisfactory agreement for the ICD-10 diagnosis of anxiety disorder unspecified, which has exactly the same symptom criteria as its DSM-III-R and DSM-IV counterparts.

Procedural Validity

Because there is no true gold standard, the aim in the validity studies was to compare the

TABLE 3. Comparative procedural validity: Monash Interview for Liaison Psychiatry (MILP) vs. DSM-III-R decision tree (DTREE)

	Base Rate MILP, %	Base Rate DTREE, %	Kappa	% Agreement
Major depression	27.8	35.2	0.74	88.9
Dysthymia	3.7	3.7	—	100
Adjustment disorder with depressed mood	31.5	24.1	0.73	88.9
Depressive disorders not otherwise specified (NOS)	16.7	25.9	0.58	85.2
Minor mood disorders (all)	35.2	29.6	0.79	90.7
Generalized anxiety disorder	7.4	7.4	1.00	100
Agoraphobia	11.1	14.8	0.84	96.3
Obsessive-compulsive disorder	7.4	7.4	1.00	100
Adjustment disorder with anxious mood	9.3	13.0	0.44	88.9
Anxiety disorder NOS	20.4	29.6	0.07	64.8
Anxiety disorders (all)	31.5	38.9	0.68	85.2
Somatoform pain disorder	5.6	0	—	94.4
Undifferentiated somatoform disorder	25.9	16.7	0.18	72.2
Somatoform disorder NOS	11.1**	1.9**	—	90.7
Psychological factors affecting physical condition	9.3	13.0	0.44	88.9
Somatizing disorders (all)	50.0*	35.2*	0.26	63.0
Alcohol dependence or abuse	3.7	5.6	—	98.1
Other drug dependence or abuse	35.2	38.9	0.84	92.6
Mean			0.61	88.2

Note: Exact McNemar's

* $P < 0.10$, NS; ** $P < 0.05$.

MILP with a variety of known standards, thus placing the MILP in relation to these standards, in a form of “triangulation”—a term borrowed from “orienteeing.”⁴⁹ In these studies, there was close agreement between the DSM-III-R diagnoses of the MILP made by the computerized algorithm compared with the decision tree in all groups, except the somatizing disorders, in which the MILP yielded more diagnoses. There was a similar result in the comparison of ICD-10 diagnoses made by the MILP and the ICD-10 checklist.

Greater disagreements occurred in the MILP vs. SCID comparison. The SCID yielded no diagnoses of somatizing disorders, which is not surprising. The SCID has a small range of somatic diagnoses, only proceeds with the somatic section of the interview if the interviewer judges it to be relevant after screening questions, and is noted to make somatoform diagnoses infrequently.⁵⁰ These differences contrast with the

exhaustive symptom inquiry of the MILP. Similarly, with other drug use disorders, the MILP yielded 13 diagnoses, compared with the SCID’s one. Anxiety disorders and minor mood disorders each had “fair” agreement,⁴² with kappas of 0.37 and 0.30, respectively. Although no significant bias was recorded and the actual number of cases was small, generalized anxiety disorder, phobias, and obsessive-compulsive disorder were all diagnosed more often with the MILP than with the SCID. Four patients diagnosed by the MILP as having minor mood disorder were diagnosed by the SCID as having major depression, consistent with the lower rate of diagnosis of major depression by the MILP compared with the other standards. However, seven of those patients given a diagnosis of minor mood disorder by MILP were given no diagnosis by the SCID. This difference suggests that the MILP has a lower threshold for diagnosis of minor mood disorders. With the exception of major depres-

TABLE 4. Comparative procedural validity: Monash Interview for Liaison Psychiatry (MILP) vs. International Classification of Diseases (ICD)-10th Edition Checklist

	Base Rate MILP, %	Base Rate ICD Checklist, %	Kappa	% Agreement
Depressive episode	14.8*	25.9*	0.44	81.5
Dysthymia	11.1*	5.6*	0.64	94.4
Adjustment disorders: depressive reactions	46.3	46.3	0.63	81.5
Depressive episode: unspecified	16.7	14.8	0.37	83.3
Minor mood disorders (all)	57.4	53.7	0.55	77.8
Generalized anxiety disorder	16.7**	31.5**	0.51	81.5
Agoraphobia	9.3	7.4	0.88	98.1
Obsessive-compulsive disorder	7.5	7.4	1.00	100
Adjustment disorder with anxiety	9.3	13.0	0.63	92.6
Anxiety disorder unspecified	38.9***	14.8***	0.17	64.8
Anxiety disorders (all)	42.6	51.9	0.60	79.6
Persistent somatoform pain disorder	5.6	3.7	—	90.7
Undifferentiated somatoform disorder	16.7***	0***	—	83.3
Dissociative anesthesia	7.4	9.3	0.64	94.4
Neurasthenia	13.0**	1.9**	—	85.2
Somatoform autonomic dysfunction	7.4	13.0	0.10	83.3
Psychological and behavioral factors associated . . .	1.9**	11.1**	—	90.7
Somatizing disorders (all)	44.4	33.3	0.15	59.3
Alcohol dependence or harmful use	3.7	5.6	0.79	98.1
Other drug dependence or harmful use	35.2	37.0	0.80	90.7
Mean			0.56	85.5

Note: Exact McNemar’s.

* $P < 0.10$, NS; ** $P < 0.05$; *** $P < 0.01$.

sion, the MILP tends to be overinclusive compared with the SCID.

CONCLUSIONS

The studies of comparison described here examined different possible sources of variance. The comparison with the SCID is the most susceptible to error. Although unstructured assessments have been used as the standard for validity studies,⁵¹ this method was not used here, as it provides no standardization of the process. The use of a structured interview such as the SCID minimizes information and interpretation variance of the standard; however, since the SCID was conducted at a time separate from the MILP, the possibility of occasion variance exists. This possibility was minimized by conducting the SCID interview very soon after the MILP.

There are limitations in the study and in the interpretation of results. Because the interview is so broad, leading to a large number of diagnoses, the base rate for some items and diagnoses was small, affecting the kappa statistic adversely. The study sample was "enriched" by screening and was thus similar to clinical pop-

ulations seen in the consultation-liaison setting. Still some diagnoses were rarely made. Validity of these diagnoses will need to be tested in clinical samples with higher prevalences for these disorders. In line with principles established in studies of the DIS,⁵² it was considered important to test the interview in the type of population in which it is to be used and to examine the disorders commonly encountered in this setting.

Because there is no true "gold standard," what is particularly important is that the interview has adequate reliability. The results demonstrate that it does. However, compared with other diagnostic methods, the results show less than moderate agreement for some diagnostic groups. These results are mostly explained by systematic bias reflecting different thresholds for diagnosis. As Carey and Gottesman⁴⁷ have explained, "When two raters are rating the same phenomenon, using two different thresholds for doing so, the crucial issue is to determine which of the raters has a better, more valid, threshold" (p. 1458). If MILP overdiagnoses somatizing disorders compared with other standards, the validity of these ratings will need to be tested in

TABLE 5. Comparative procedural validity: Monash Interview for Liaison Psychiatry (MILP) vs. Structured Clinical Interview for DSM-III-R (SCID)

	Base Rate MILP, %	Base Rate SCID, %	Kappa	% Agreement
Major depression	27.8	37.0	0.54	79.6
Dysthymia	3.7	0	—	96.3
Adjustment disorders	31.5	24.1	0.27	70.4
Minor mood disorders (all)	35.2	24.1	0.30	70.4
Generalized anxiety disorder	7.4*	0*	—	92.6
Agoraphobia	11.1	7.4	0.34	88.9
Obsessive-compulsive disorder	7.4	3.7	—	96.3
Adjustment disorder with anxious mood	9.3	11.1	-0.11	79.6
Anxiety disorders (all)	29.6	20.4	0.37	75.9
Somatoform pain disorder	5.6	0	—	94.4
Undifferentiated somatoform disorder	25.9***	0***	—	74.1
Psychological factors affecting physical condition	9.3**	0**	—	90.7
Somatizing disorders (all)	35.2***	0***	—	64.8
Alcohol dependence or abuse	3.7	7.4	0.65	96.3
Other drug dependence or abuse	18.5***	1.9***	0.15	88.3
Mean	—	—	0.31	83.9

Note: Exact McNemar's

* $P < 0.10$, NS; ** $P < 0.05$; *** $P < 0.01$.

some further way, such as an examination of predictive or criterion validity. Such an approach will be a movement away from establishing the validity of the procedure toward establishing the validity of the constructs.

Nevertheless, it is important to note the differences between the MILP and other diagnostic methods. The MILP appears to be stricter than other measures in diagnosing major depression. Reliability of this diagnosis is almost perfect. It may be that the MILP is "tougher" in its judgment concerning the attribution of cause and the exclusion of items judged to be of organic origin. In the area of somatizing disorders and drug use disorders, the inquiry of the MILP is more intensive than that of the SCID. In addition, however, the comparisons of the MILP with the DSM decision tree and the ICD checklist suggest there may be problems with the procedural validity with respect to the somatizing disorders. Low kappas for these comparisons are only partly explained by systematic bias. Again, this may be related to difficulties in the process of attribution of cause. This will need further examination.

It is interesting to compare these results with the reliability and validity studies of other instruments. The DIS has probably had the most investigation. The earliest study by Robins et al. yielded the best results (kappas ranging from 0.40 to 0.86), but the study was done with a psychiatric population in which the base rates for diagnoses were high.⁵³ A similar study by Helzer et al. of a community sample showed weaker agreements (kappa 0.24 to 0.68).⁵¹ In both these studies, the standard was a second DIS interview administered by a psychiatrist independently of the first lay interview, a design close to being one of test-retest reliability. Although one would expect almost perfect agreement in this situation, the relatively modest agreement achieved (mean kappa 0.69),⁵³ despite using a structured interview and computer-generated diagnoses, demonstrates how error can be introduced by the interviewer through observation and interpretation variance. In a similar vein, during the DSM-IV trials, McGorry et al. demonstrated significant disagreements be-

tween diagnostic procedures using identical diagnostic criteria.⁵⁴ DIS studies using checklists and clinical interviews as comparisons have also produced particularly poor agreements.^{51,52,55}

The SCID does not yet have published studies of procedural validity, but there have been a number of reports of reliability. Two studies of interrater reliability, one using audiotaped interviews of patients with a range of disorders³¹ and the other joint interviews of depressed patients,⁵⁶ found mean kappas for agreement on diagnosis of 0.76 and 0.68, respectively. Examination of test-retest reliability of the SCID found a mean kappa of 0.61 in a patient sample and 0.37 in a nonpatient sample in which the base rates of diagnoses were lower.⁵⁰ Interestingly, the base rates for the somatizing disorders in both samples were so low (from 0% to 2%) that no measures of agreement were calculated. The Standardized Polyvalent Psychiatric Interview,²⁸ a composite interview built around the Clinical Interview Schedule,⁵⁷ found kappas for interrater reliability mostly in the range of 0.70 to 0.90.

The MILP has been designed specifically for use in the physically ill and has been acceptable to patients and interviewers. Its inquiry is more extensive than other similar instruments in the areas of somatizing disorders, drug use disorders, and subthreshold disorders. The reliability and procedural validity of the MILP is comparable to other structured interviews, although the results highlight the different thresholds for diagnosis compared with the SCID, and potential difficulties with some of the subthreshold disorders, those on the boundary of normality, and those requiring a degree of judgment. As with the SCID, the importance of paying "careful attention to adequate training of interviewers" (p. 636)⁵⁰ is affirmed. The results confirm the need for more study of the classification of psychiatric disorders in the physically ill and the testing of construct and predictive validity of diagnostic categories.

The authors thank Anne Silbereisen, Kevan Pitcher, and Lisa Henry, who conducted the interviews; Paul Low for data management; and

Roland Yap for writing the PROLOG software for the diagnostic algorithm. The study was supported by the National Health and Medical Research Council of Australia.

References

- Vaillant GE: The disadvantages of DSM-III outweigh its advantages. *Am J Psychiatry* 1984; 141:542–545
- Berrios GE: Phenomenology and psychopathology: was there ever a relationship? *Compr Psychiatry* 1993; 34:213–220
- Leigh H, Price L, Ciarcia J, et al: DSM-III and consultation-liaison psychiatry: towards a comprehensive medical model of the patient. *Gen Hosp Psychiatry* 1982; 4:283–289
- Lipowski ZJ: Is “organic” obsolete? *Psychosomatics* 1990; 31:342–344
- Snaith RP: The concepts of mild depression. *Br J Psychiatry* 1987; 150:387–393
- Wells KB, Stewart A, Hays RD, et al: The functioning and well-being of depressed patients: results from the Medical Outcome Study. *JAMA* 1989; 262:914–919
- McKegney FP, McMahon T, King J: The use of DSM-III in a general hospital consultation-liaison service. *Gen Hosp Psychiatry* 1983; 5:115–121
- Creed F: Anxiety in general medical patients, in *Handbook of Anxiety*, Vol. 2: Classification, Etiological Factors and Associated Disturbances, edited by Noyes R, Roth M, Burrow GD. Amsterdam, The Netherlands, Elsevier Science Publishers, 1988, pp. 239–268
- Foulds GA, Bedford A: Hierarchy of classes of personal illness. *Psychol Med* 1975; 5:181–192
- Katon W, Lin E, VonKorff M, et al: Somatization: a spectrum of severity. *Am J Psychiatry* 1991; 148:34–40
- DeGruy F, Crider J, Hashimi DK, et al: Somatization disorder in a university hospital. *J Fam Pract* 1987; 25:579–584
- Robins LN, Helzer JE, Croughan J, et al: National Institute of Mental Health Diagnostic Interview Schedule: its history, characteristics, and validity. *Arch Gen Psychiatry* 1981; 38:381–389
- Robins LN, Wing J, Wittchen H-U, et al: The Composite International Diagnostic Interview: an epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Arch Gen Psychiatry* 1988; 45:1069–1077
- Janca A, Ustun TB, Sartorius N: New versions of World Health Organization instruments for the assessment of mental disorders. *Acta Psychiatr Scand* 1994; 90:73–83
- Spitzer RL, Williams JBW, Gibbon M, et al: The Structured Clinical Interview for DSM-III-R (SCID) I: history, rationale and description. *Arch Gen Psychiatry* 1992; 49:624–629
- Yap RHC, Clarke DM: An expert system for psychiatric diagnosis using the DSM-III-R, DSM-IV and ICD-10 classifications. Paper published in the Proceedings of the Annual Fall Symposium of the American Medical Informatics Association, Washington DC, October 1996, pp. 229–233
- American Psychiatric Association: *Diagnostic and Statistical Manual of Mental Disorders*, 3rd Edition, Revised. Washington, DC, American Psychiatric Association, 1987
- American Psychiatric Association: *Diagnostic and Statistical Manual of Mental Disorders*, 4th Edition. Washington, DC, American Psychiatric Association, 1994
- World Health Organization: *The ICD-10 classification of mental and behavioural disorders. Diagnostic criteria for research*. Geneva, Switzerland, World Health Organization, 1993
- Stewart MA, Drake F, Winokur G: Depression among medically ill patients. *Dis Nerv Syst* 1965; 26:479–485
- Endicott J: Measurement of depression in patients with cancer. *Cancer* 1984; 53:2243–2247
- Davidson NRT, Miller RD, Turnbull CD, et al: Atypical depression. *Arch Gen Psychiatry* 1982; 39:527–534
- Carney MWP, Roth M, Garside RF: The diagnosis of depressive syndromes and the prediction of ECT response. *Br J Psychiatry* 1965; 3:659–674
- Vargas LA, Loya F, Hodde-Vargas J: Exploring the multidimensional aspects of grief reactions. *Am J Psychiatry* 1989; 146:1484–1488
- Escobar JI, Canino G: Unexplained physical complaints: psychopathology and epidemiological correlates. *Br J Psychiatry* 1989; 154:24–27
- McGorry PD, Singh B, Copolov DL, et al: Royal Park Multidiagnostic Instrument for Psychosis: Part II. Development, reliability and validity. *Schizophr Bull* 1990; 16:517–536
- Wittchen H-U, Robins LN, Cottler LB, et al: Cross-cultural feasibility, reliability and sources of variance of the Composite International Diagnostic Interview (CIDI). *Br J Psychiatry* 1991; 159:645–653
- Lobo A, Campos R, Perez-Echeverria M-J, et al: A new interview for the multi-axial assessment of psychiatric morbidity in medical settings. *Psychol Med* 1993; 23:505–510
- Helzer JE, Robins LN, Taibleson M, et al: Reliability of psychiatric diagnosis, I: A methodological review. *Arch Gen Psychiatry* 1977; 34:129–133
- Grove WM, Andreasen NC, McDonald-Scott P, et al: Reliability studies of psychiatric diagnosis. *Arch Gen Psychiatry* 1981; 38:408–413

Interview for Liaison Psychiatry

31. Skre I, Onstad S, Torgersen S, et al: High interrater reliability for the Structured Clinical Interview for DSM-III-R Axis I (SCID-I). *Acta Psychiatr Scand* 1991; 84:167–173
32. Janca A, Robins LN, Bucholz KK, et al: Comparison of Composite International Diagnostic Interview and clinical DSM-III-R criteria checklist diagnoses. *Acta Psychiatr Scand* 1991; 84:167–173
33. Janca A, Robins LN, Cottler LB, et al: Clinical observation of assessment using the Composite International Diagnostic Interview (CIDI): an analysis of the CIDI field trials—wave II at St. Louis site. *Br J Psychiatry* 1992; 160: 815–818
34. Goldberg DP: The detection of psychiatric illness by questionnaire, Maudsley Monograph No. 21. London, UK, Oxford University Press, 1972
35. Goodchild ME, Duncan-Jones P: Chronicity and the General Health Questionnaire. *Br J Psychiatry* 1985; 146:55–61
36. Clarke DM, Smith GC, Hermann HE. A comparative study of screening instruments for mental disorders in general hospital patients. *Int J Psychiatry Med* 1993; 23:323–337
37. Cohen J: A coefficient of agreement for nominal scales. *Educational Psychological Measure* 1960; 20:37–46
38. Streiner DL: Learning how to differ: agreement and reliability statistics in psychiatry. *Can J Psychiatry* 1995; 40:60–66
39. Langenbucher J, Labouvie E, Morgenstern J: Measuring diagnostic agreement. *J Consult Clin Psychol* 1996; 64:1285–1289
40. Spitznagel EL, Helzer JE: A proposed solution to the base rate problem in the kappa statistic. *Arch Gen Psychiatry* 1985; 42:725–728
41. Baldessarini RJ, Finklestein S, Arana GW: The predictive power of diagnostic tests and the effect of prevalence of illness. *Arch Gen Psychiatry* 1983; 40:569–573
42. Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 1977; 33:159–174
43. Fleiss JL: *Statistical Methods for Rates and Proportions*, 2nd Edition. New York, Wiley, 1981
44. Spitzer RL, Williams JBW: Classification of Mental Disorders and DSM-III, in *Comprehensive Textbook of Psychiatry*, 4th Edition, edited by Kaplan HI, Sadock BJ. Baltimore, MD, Williams & Wilkins, 1985, pp. 591–613
45. First MB, Williams JBW, Spitzer RL: *DTREE: the electronic DSM-III-R (computer software, user's guide, case workbook)*. Washington, DC, American Psychiatric Press, Inc., 1989
46. First MB, Opler LA, Hamilton RM, et al: Evaluation in an inpatient setting of DTREE: a computer-assisted diagnostic assessment procedure. *Compr Psychiatry* 1993; 34:171–175
47. Carey G, Gottesman II: Reliability and validity in binary ratings. *Arch Gen Psychiatry* 1978; 35:1454–1459
48. Berry KJ, Mielke PW, Helmericks SG: An algorithm to generate discrete probability distributions: binomial, hypergeometric, negative binomial, inverse hypergeometric and poisson. *Behavior Research Methods, Instruments, and Computers* 1994; 26:366–367
49. Campbell DT, Fiske C: Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 1959; 56:81–105
50. Williams JBW, Gibbon M, First MB, et al: The Structured Clinical Interview for DSM-III-R (SCID) II: Multisite test–retest reliability. *Arch Gen Psychiatry* 1992; 49:630–636
51. Helzer JE, Robins LN, McEvoy MA, et al: A comparison of clinical and Diagnostic Interview Schedule diagnoses: Physician re-examination of lay-interviewed cases in the general population. *Arch Gen Psychiatry* 1985; 42:657–666
52. Anthony JC, Folstein M, Romanoski AJ, et al: Comparison of the lay Diagnostic Interview Schedule and a standardized psychiatric diagnosis. *Arch Gen Psychiatry* 1985; 42:667–675
53. Robins LN, Helzer JE, Ratcliff KS, et al: Validity of the Diagnostic Interview Schedule, Version II: DSM-III diagnoses. *Psychol Med* 1982; 12:855–870
54. McGorry PD, Mihalopoulos C, Henry L, et al: Spurious precision: procedural validity of diagnostic assessment in psychotic disorders. *Am J Psychiatry* 1995; 152:220–223
55. Burnam MA, Karno M, Hough RL, et al: The Spanish Diagnostic Interview Schedule. *Arch Gen Psychiatry* 1983; 40:1189–1196
56. Weiss MG, Raguram R, Channabasavanna SN: Cultural dimensions of psychiatric diagnosis: a comparison of DSM-III-R and illness explanatory models in South India. *Br J Psychiatry* 1995; 166:353–359
57. Goldberg DP, Cooper B, Eastwood MR, et al: A standardized psychiatric interview for use in community surveys. *Br J Prev Soc Med* 1970; 24:18–23