Journal of Clinical Geropsychology, Vol. 1, No. 4, 1995

Update on the Reliability of Diagnosis in Older Psychiatric Outpatients Using the Structured Clinical Interview for DSM IIIR

Daniel L. Segal,[1] Robert I. Kabacoff,[1] Michel Hersen,[1] Vincent B. Van Hasselt,[1] and Christine Freeman Ryan[1]

This study extends previous research designed to evaluate the reliability of the Structured Clinical Interview for DSM IIIR Axis I (SCID-I) with a heterogeneous outpatient population of older adults 55 years old and over (range = 55–87 years; mean = 67.05 years). Forty SCID interviews were administered and audiotaped by masters level clinicians working toward their clinical psychology doctorate degrees. Inter-rater reliability estimates (kappa and percent agreement) were calculated for the broad diagnostic categories of mood disorder, anxiety disorder, somatoform disorder, and psychoactive substance use disorder, as well as specific disorders including major depression, dysthymia, and panic disorder. Inter-rater reliability estimates based on kappa generally were substantial, with five categories over .73. The range was .23 (substance abuse) to .90 (major depression). We conclude that the SCID-I can be effectively administered by relatively inexperienced clinicians to reliably diagnose numerous disorders in older psychiatric outpatients. Suggestions for future research in this area are discussed.

KEY WORDS: SCID; DSM IIIR; reliability; diagnosis; older adults.

Numerous studies over the past several years have evaluated the reliability (inter-rater or test–retest methods) and validity of the original Structured Clinical Interview for DSM III (SCID; Spitzer and Williams, 1984) and its revision devised in 1988 (Spitzer, Williams, Gibbon, and First, 1988) to reflect modifications that appeared in DSM IIIR (American Psychiatric Association, 1987). However, a recent review of these reliability investigations (see Segal, Hersen, and Van Hasselt, 1994) suggests that little attention has been directed to assessing psychometric properties of the SCID in under represented populations, such as the elderly. This void in the literature is significant given that the elderly are the fastest growing segment of our society and appear increasingly in need of psychological evaluation and intervention (Donohue, Hersen, and Van Hasselt, in press). Interest in specialized

[1]Center for Psychological Studies, Nova Southeastern University.

1079-9362/95/1000-0313$07.50/0

assessment and treatment of emotional disorders in older adults has similarly increased in recent years (Cohen, 1989), and it seems important to fully substantiate reliability of the SCID with this burgeoning yet still under-researched population.

To date, only two investigations have specifically evaluated the reliability of the SCID in older adults. As part of their evaluation of the relative efficacy of three depression-screening scales in a nonclinical elderly population, Stukenberg, Dura, and Kiecolt-Glaser (1990) reported inter-rater reliability of the presence or absence of a mood disorder in 75 cases (kappa = .92). However, reliability data for specific mood disorders (e.g., major depression, dysthymia) were not provided and reliability for problems other than mood disorders was not determined. Additionally, base rates for mood disorders were not presented for the reliability sample. As kappas are unstable at extremely high or low base rates (Grove, 1987; Segal *et al.*, 1994), this omission severely limits interpretation of the reported kappa.

Our research group (Segal, Hersen, Van Hasselt, Kabacoff, and Roth, 1993) recently evaluated the reliability of the SCID with a mixed inpatient (N = 10) and outpatient (N = 23) elderly population in a "natural" clinical setting, as part of a comprehensive assessment battery. Specifically, the SCID was administered in the first or second intake session to older adults seeking outpatient treatment for a wide range of psychiatric disorders. Interviewers were second and third year graduate students in clinical psychology, and not highly trained or skilled clinical researchers as has been the norm in many previous reliability investigations. However, our interviewers are more representative of the types of mental health professionals who administer the SCID in clinical settings (e.g., psychiatric inpatient and outpatient facilities, community mental health centers) and who function as "front line" clinicians. Our graduate students had prior coursework in adult psychopathology and basic interviewing strategies and were trained in administration of the SCID primarily via didactic presentations and several role-played SCID interviews with corrective feedback.

In our study, inter-rater reliability was investigated in a sample of 33 older adults (mean age = 67.3 years, range = 56-84, SD = 8.4). Audio and videotaped interviews were assessed retrospectively by an independent doctoral level rater. Reliability estimates (kappa and percentage agreement) were calculated for major depressive episode (base rate = 47%), and the broad diagnostic categories of anxiety disorders (base rate = 15%) and somatoform disorders (base rate = 12%). As kappas are unstable at extremely low base rates (Grove, 1987), reliability statistics were not computed for other disorders where less than 10% (N < 3) of the sample was so diagnosed by one of the assessors. Obtained kappas and percent agreements, respectively, were as follows: major depression (.70, 85%), anxiety disorders (.77, 94%), and somatoform disorders (1.0, 100%). We concluded from this pilot study that reliability of the SCID administered by graduate level clinicians in a limited sample of older adults appears promising. However, due to the small sample size, only one specific disorder (major depression) and two general disorders (anxiety and somatoform) were assessed. Clearly, additional research with larger samples and evaluation of more disorders was warranted. An additional limitation was that the sample was composed of both inpatient and outpatients, while the version of the SCID employed was designed for use in outpatient settings and did not include

a module for schizophrenia. Consequently, most of the subjects who did not receive a SCID diagnosis were inpatient subjects suffering from chronic schizophrenic symptomatology that was not tapped in the version of the SCID applied. The purpose of the present study, therefore, was to expand the previous study by focusing specifically on older adult outpatients and increasing sample size to evaluate interrater reliability of the SCID for additional disorders. Further, given that our initial positive findings were generated by relatively inexperienced interviewers in the pilot study, we felt that a replication was warranted.

METHOD

Subjects

The older adult sample consisted of 40 patients (14 males, 26 females) from the Nova Community Clinic for Older Adults, a university-based outpatient facility. Mean age was 67.05 years (SD = 9.87), with a range from 55 to 87 years. Patients were predominantly Caucasian (97%), and resided in private homes (46%), or apartments and condominiums (54%). Fifty-one percent of the patients were married, 18% were divorced, 5% were separated, and 26% were widowed. Social class, determined on the basis of the Hollingshead (1957) Two Factor Index, yielded the following levels: 1 (8%), 2 (21%), 3 (33%), 4 (28%), and 5 (10%).

Diagnostic Assessment Strategy

The SCID-I (patient edition: developed June 1, 1988) (Spitzer, Williams, Gibbon, and First, 1988) was administered, since it was the most current version of this instrument for Axis I diagnoses at the time the study was conducted. The SCID-I was designed to reflect modifications in DSM IIIR and employed to extract information to make diagnostic decisions for 33 of the more commonly diagnosed DSM IIIR Axis I disorders in adults. All interviews were audiotaped or videotaped for *post hoc* review by the reliability assessors (D.L.S. and C.F.R.).

Second- and third-year masters level students working toward their doctoral degrees served as initial SCID interviewers. Two of the authors (D.L.S., a doctoral level psychologist, and C.F.R., a masters level graduate student) alternately served as independent reliability assessor for each taped interview. Using SCID criteria, the assessor independently rated the audiotape or videotape of the initial diagnostic interview in order to reach a diagnostic conclusion. Current DSM IIIR Axis I diagnoses were obtained, and reliability was assessed for the broad diagnostic categories of mood, anxiety, somatoform, and psychoactive substance use disorders, as well as the specific disorders of major depressive episode, dysthymia, and panic disorder.

Procedures

Initial training for the SCID was conducted over a period of several weeks, amounting to a total training time of 16 hours. During training sessions, DSM IIIR criteria for each disorder were reviewed, and a SCID training videotape designed by the developers of the interview schedule was presented and then discussed. Role-played SCID interviews subsequently were conducted and rated. In cases of diagnostic disagreement, a consensus discussion was carried out to maximize training of the assessor and to clarify the intent of some SCID questions, criteria, and procedures. After administration of the first SCID, each assessor was provided corrective feedback as to technical issues and interviewing style.

Information given to all assessors prior to administration of the SCID was standardized, consisting of a brief synopsis from a telephone screening of the patient. Informed consent was obtained from all patients prior to their participation. The SCID was administered on either the first or second assessment session, as part of a comprehensive intake assessment battery that included measures of depression, hopelessness, anxiety, assertiveness, marital adjustment, memory functioning, and social support. After administering the SCID, assessors filled out a summary sheet noting all current DSM IIIR Axis I diagnoses. The identical procedure was followed by the reliability assessor, except that ratings were made on the basis of videotapes and audiotapes.

RESULTS

Table I presents all current DSM IIIR Axis I diagnoses based on SCID evaluations for the 40 outpatients. Multiple diagnoses were permitted provided that patients met full criteria for each disorder.

Agreement between assessors was calculated using percentage agreement and the kappa correlation coefficient, which are commonly used statistics in similar reliability studies (see Arntz, van Beijsterveldt, Hoekstra, Eussen, and Sallaerts, 1992; Riskind, Beck, Berchick, Brown, and Steer, 1987) and were calculated in our previous study (Segal *et al.*, 1993). Unlike percentage agreement, kappa corrects for chance levels of agreement (Fleiss, 1981). Although there are no definitive guidelines for the interpretation of the kappa index, Landis and Koch (1977) have provided the following benchmarks: values above .81 are considered almost perfect, values from .61 to .80 suggest substantial agreement, values between .41 and .60 indicate moderate agreement. Kappa values between .21 and .40 denote fair agreement, while values below .20 are slight. Values below 0.0 reflect less than chance disagreement.

As kappas are unstable at extreme base rates, reliability statistics were only computed for those disorders where 10% ($N = 4$) or more of the sample was so diagnosed by *at least one* of the assessors. Thus, reliability coefficients were determined for the broad diagnostic categories of mood disorder (60% base rate), anxiety disorder (25% base rate), somatoform disorder (9% base rate), and psychoactive substance use disorder (9% base rate). Specific disorders evaluated included major

Table I. Current DSM IIIR Axis I Disorders Based on SCID Evaluations (N = 40)

Diagnosis	Number of cases[a]	
	Assessor 1	Assessor 2
Mood disorder	28	28
Major depression	21	21
Dysthymia	4	3
Bipolar disorder-depressed	2	2
Bipolar disorder-manic	0	0
Bipolar disorder-mixed	0	1
Anxiety disorder	9	11
Panic disorder	6	6
Generalized anxiety disorder	3	3
Social phobia	0	1
Agoraphobia without panic	0	1
Somatoform disorder	4	3
Somatization	2	2
Somatoform pain disorder	1	1
Hypochondriasis	1	0
Psychoactive substance use disorder	5	2
Alcohol abuse or dependence	3	1
Drug abuse or dependence	2	1
Eating disorder	0	1
Adjustment disorder	2	2
No DSM IIIR Axis 1 disorder	5	5

[a]Numbers do not add to 40 because patients could have more than one disorder.

depressive disorder (58% base rate), dysthymia (9% base rate), and panic disorder (15% base rate). Base rate was calculated as the percentage of subjects with the disorder out of the total sample *averaged* over both assessors. Estimates of less prevalent individual disorders (i.e., bipolar, generalized anxiety, social phobia, and agoraphobia without panic disorder) were not calculated due to low base rates in our sample; rather, they were subsumed under broader diagnostic categories where applicable. Diagnoses were evaluated on a dichotomous basis: either present or absent. For example, with major depression, a match was obtained if both assessors agreed on presence or absence of this disorder for a patient, regardless of whether other diagnoses also were recorded. For the four broad diagnostic categories (mood, anxiety, somatoform, and psychoactive substance use), agreement on the specific subcategories was required for a match to be considered positive. For example, a patient diagnosed with panic disorder by one assessor and generalized anxiety disorder by another assessor was considered to be a disagreement, even though both disorders fall under the overall rubric of anxiety disorders.

Inter-rater reliability estimates for current broad and specific DSM IIIR diagnostic categories with adequate base rates are presented in Table II. Results for the broad diagnostic group of somatoform disorder (kappa = .84, percent agreement = 98) suggests almost perfect agreement. Agreement was slightly lower, but still substantial, for mood disorder (kappa = .79, percent agreement = 90) and anxiety disorder (kappa = .73, percent agreement = 90). Psychoactive substance

Table II. SCID Inter-rater Reliability (Kappa), Percentage Aggreement, and Base Rates for SCID Diagnoses

Diagnosis	Kappa	Percentage agreement	Base rate
Mood disorder (general)	.79	90	60%
Major depressive disorder	.90	95	58%
Dysthymia	.53	93	9%
Anxiety disorder (general)	.73	90	25%
Panic disorder	.80	95	15%
Somatoform disorder (general)	.84	98	9%
Psychoactive substance use (general)	.23	88	9%

use disorder (kappa = .23, percent agreement = 88) had the lowest agreement, reflecting only fair agreement.

For specific disorders, almost perfect agreement was obtained for major depressive disorder (kappa = .90, percent agreement = 95) and panic disorder (kappa = .80, percent agreement = 95), while agreement for dysthymia (kappa = .53, percent agreement = 93) was moderate.

DISCUSSION

The present investigation is an extension and modification of an initial study conducted to evaluate the reliability of the SCID in older adults. Our findings confirm that this instrument leads to highly reliable classification for several broad and specific DSM IIIR Axis 1 diagnoses in an older psychiatric sample. Additionally, we successfully enhanced results from our previous study providing further evidence that the SCID can be effectively administered by masters level clinicians who are representative of the types of clinicians providing services in many applied settings. Indeed, the semi-structured format of the SCID makes it an ideal instrument for beginning clinicians who may not: (1) know the criteria for many DSM IIIR diagnoses, (2) have easy access to questions to assess presence or absence of symptoms, and (3) have grasped the intricacies of differential diagnosis and diagnostic interviewing. Clearly, a thorough understanding of symptoms, questions, and diagnostic interviewing processes is required to make reliable and valid diagnoses, and the SCID provides the novice clinician with structure and training in all three areas.

Our results are quite similar to agreement rates found in our pilot study (Segal *et al.*, 1993) with a smaller mixed inpatient and outpatient sample of older adults. Segal *et al.* (1993) reported reliability rates for only three disorders. Obtained kappas and percent agreements, respectively, were as follows: major depression (.70, 85%), anxiety disorders (.77, 94%), and somatoform disorders (1.0, 100%). In contrast, inter-rater agreement rates in the present study were much higher for major depression (.90, 95%), almost identical for anxiety disorders (.73, 90%), and somewhat lower but still reflecting almost perfect agreement for somatoform disorders

(.84, 98%). Our results are also comparable to other SCID reliability investigations with younger patient samples, described below by diagnostic category.

Mood Disorders

Our inter-rater agreement for the general class of mood disorder was substantial (kappa = .79). For specific disorders, almost perfect agreement was obtained for major depressive disorder (kappa = .90, percent agreement = 95), while agreement for dysthymia (kappa = .53, percent agreement = 93) was moderate. Interestingly, major depression was the most frequent diagnosis in our sample (base rate = 60%), suggesting high prevalence of this disorder in the aged. Riskind *et al.* (1987) videotaped 75 psychiatric outpatients to assess inter-rater reliability of DSM III major depression (MDD; N = 25, base rate = 32%), and found a kappa of .72. Similarly, our results for major depression and dysthymia are higher than data obtained by developers of the SCID in their extensive multisite, test–retest reliability project (Williams *et al.*, 1992). In a sample of 390 psychiatric patients, kappa for current major depression and dysthymia were .64 and .40, respectively. Our kappas with a much smaller sample of patients was higher for major depression (.90) and dysthymia (.53). Even more elevated values for major depression (.93) and dysthymia (.88) were obtained by Skre, Onstad, Torgerson, and Kringlen (1991), who assessed inter-rater reliability using 54 audiotaped SCID interviews of adults participating in a larger Norwegian twin study of mental illness. Clearly, though, our results with older individuals are comparable to these previous studies.

Anxiety Disorders

Our agreement was substantial for the broad diagnostic category of anxiety disorder (kappa = .73). Skre *et al.* (1991) obtained similar excellent inter-rater agreement for the broad anxiety disorder group (.82). For specific disorders, we found almost perfect agreement for panic disorder (kappa = .80). This value is comparable to the kappa of .82 noted by Skre *et al.* (1991). Similarly, Williams, Spitzer, and Gibbon (1992), in their extensive test–retest reliability study of the Upjohn version of the SCID, reported that agreement (kappa) on the diagnosis of panic disorder (N = 62) was very good (.87), although their base rate was very high (86%). In the large multisite, test–retest reliability investigation for the full SCID (Williams *et al.*, 1992), kappa for panic disorder was somewhat lower (.58, base rate = 9%)

Somatoform Disorders

Results for the broad diagnostic group of somatoform disorder (kappa = .84) suggest almost perfect agreement, similar to our earlier study (Segal *et al.*, 1993). In sharp contrast, virtually no agreement was found for the general category of somatoform disorder (—.03) by Skre *et al.* (1991). Despite our low base rate (9%) for such

disorders, raters disagreed on only one of the four cases, resulting in a very high kappa. It is possible that somatoform disorders are more easily identified in older patients, who may be more obviously somatically preoccupied than their younger counterparts. Further research is needed to clarify the appropriateness of criteria in this diagnostic category for diverse age groups with larger samples and higher base rates.

Psychoactive Substance Use Disorders (PSUD)

Psychoactive substance use disorder (PSUD) (kappa = .23, percent agreement = 88) had the lowest agreement of all disorders in our study, reflecting only fair agreement. These comparatively lower values are contrasted with the almost perfect reliability for the broad diagnostic category of PSUD (kappa = .93) found by Skre *et al.* (1991) in their Norwegian twin study, and high agreement rates for alcohol abuse (.75) and drug abuse (.84) reported in the large multisite SCID study (Williams *et al.*, 1992). However, both of these studies were conducted primarily on middle-aged or younger subjects, not older adults. It has been suggested that many diagnostic criteria for substance abuse may not be applicable for older individuals, given their unique social, psychological, and physiological states (see review by King, Hersen, Segal, and Van Hasselt, 1994 for further discussion of diagnostic issues with elderly substance abusers).

Several limitations of the present study should be noted. While the present sample size was larger, more homogeneous, and provided reliability data for more disorders than our earlier investigation, other additional individual disorders (e.g., generalized anxiety disorder, bipolar disorder, eating disorders) still were not evaluated due to low base rates in our clinics. Reliability of such diagnoses should be assessed in older samples with higher base rates, as might be encountered in speciality clinics for those disorders. Also, a more stringent test–retest strategy may result in somewhat reduced reliability coefficients. Further replication of this study with even larger sample sizes of older adults clearly is indicated. Future work involving older subjects should also focus on evaluating the psychometric properties of the most recent version of the SCID (First, Spitzer, Gibbon, and Williams, 1994) modified to reflect recent advances in DSM IV (American Psychiatric Association, 1994).

On the basis of the present study and our earlier pilot study with older individuals, in conjunction with our clinical experience in assessment, it is apparent that the SCID is a valuable assessment tool that enhances the diagnostic interviewing skills of its users. Indeed, we replicated the positive findings that were generated by relatively inexperienced interviewers in the pilot study. We therefore recommend use of the SCID for further clinical/research applications with older adults. Additionally, it may be advantageous to augment the substance abuse module when interviewing elderly clients.

REFERENCES

American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders* (3rd Ed., revised), Author, Washington, D.C.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th Ed.), Author, Washington, D.C.

Arntz, A., van Beijsterveldt, B., Hoekstra, R., Eussen, M., and Sallaerts, S. (1992). The interrater reliability of a Dutch version of the Structured Clinical Interview for DSM-III-R Personality Disorders. *Acta Psychiat. Scand.* 85: 394-400.

Cohen, G. D. (1989). The movement toward subspeciality status for geriatric psychiatry in the United States. *Int. Psychogeriat.* 1: 201-205.

Donohue, B., Hersen, M., and Van Hasselt, V. B. (in press). Historical perspectives in clinical geropsychology. In Hersen, M., and Van Hasselt, V. B. (eds.), *Psychological Treatment of Older Adults: An Introductory Textbook,* Plenum Press, New York.

First, M. B., Spitzer, R. L., Gibbon, M., and Williams, J. B. W. (1994). *Structured Clinical Interview for Axis I DSM-IV Disorders—Patient Edition (SCID-I/P, Version 2.0)*, Biometrics Research Department, New York State Psychiatric Institute, New York.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions* (2nd Ed.), John Wiley and Sons, New York.

Grove, W. M. (1987). The reliability of psychiatric diagnosis. In Last, C. G., and Hersen, M. (eds.), *Issues in Diagnostic Research,* Plenum Press, New York, pp. 99-119.

Hollingshead, A. B. (1957). *Two Factor Index of Social Position* (Available from A. B. Hollingshead), 1965 Yale Station, New Haven, CT 06520.

King, C. J., Van Hasselt, V. B., Segal, D. L., and Hersen, M. (1994). Diagnosis and assessment of substance abuse in older adults: Current strategies and issues. *Addict. Behav.* 19: 41-55.

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

Riskind, J. H., Beck, A. T., Berchick, R. J., Brown, G., and Steer, R. A. (1987). Reliability of DSM-III-R diagnoses for major depression and generalized anxiety disorder using the Structured Clinical Interview for DSM-III-R. *Arch. Gen. Psychiat.* 44: 817-820.

Segal, D. L., Hersen, M., and Van Hasselt, V. B. (1994). Reliability of the Structured Clinical Interview for DSM-III-R: An evaluative review. *Comprehen. Psychiat.* 35: 316-327.

Segal, D. L., Hersen, M., Van Hasselt, V. B., Kabacoff, R. I., and Roth, L. (1993). Reliability of diagnosis in older psychiatric patients using the Structured Clinical Interview for DSM-III-R. *J. Psychopathol. Behav. Assess.* 15: 347-356.

Skre, I., Onstad, S., Torgerson, S., and Kringlen, E. (1991). High interrater reliability for the Structured Clinical Interview for DSM-III-R Axis I (SCID-I). *Acta Psychiat. Scand.* 84: 167-173.

Spitzer, R. L., and Williams, J. B. W. (1984). *Structured Clinical Interview for DSM-III Disorders (SCID-P 5/1/84)* (revised), Biometrics Research Department, New York State Psychiatric Institute, New York.

Spitzer, R. L., Williams, J. B. W., Gibbon, M., and First, M. B. (1988). *Structured Clinical Interview for DSM-III-R-patient version (SCID-P 6/1/88)*, Biometrics Research Department, New York State Psychiatric Institute, New York.

Stukenberg, K. W., Dura, J. R., and Kiecolt-Glaser, J. K. (1990). Depression screening scale validation in an elderly, community dwelling population. *Psychol. Assess.* 2: 134-138.

Williams, J. B. W., Gibbon, M., First, M. B., Spitzer, R. L., Davies, M., Borus, J., Howes, M. J., Kane, J., Pope, H. G., Rounsaville, B., and Wittchen, H. (1992). The Structured Clinical Interview for DSM-III-R (SCID): Multisite test–retest reliability. *Arch. Gen. Psychiat.* 49, 630-636.

Williams, J. B. W., Spitzer, R. L., Gibbon, M. (1992). International reliability of a diagnostic intake procedure for panic disorder. *Am. J. Psychiat.* 149: 560-562.