Journal of Psychopathology and Behavioral Assessment, Vol. 15, No. 4, 1993

Reliability of Diagnosis in Older Psychiatric Patients Using the Structured Clinical Interview for DSM-III-R

Daniel L. Segal,[1] Michel Hersen,[1,2] Vincent B. Van Hasselt,[1] Robert I. Kabacoff,[1] and Leonard Roth[1]

Accepted: September 24, 1993

We conducted one of the few studies that has examined the reliability of the Structured Clinical Interview for DSM-III-R Axis I (SCID-I) with a mixed inpatient and outpatient population of adults 55 years old and over (range, 56–84 years; mean, 67.33 years). All SCID interviews were videotaped or audiotaped and were administered by Master's-level clinicians working toward their doctorate degrees in clinical psychology. Interrater reliability estimates (kappa and percentage agreement) were calculated for current major depressive episode (47% base rate) and the broad diagnostic categories of anxiety disorders (15% base rate) and somatoform disorders (12% base rate). Kappa values were .70, .77, and 1.0. Respective percentage agreement was 85% for major depression, 94% for anxiety disorders, and 100% for somatoform disorders. Overall percentage agreement was 91%. We conclude that the SCID-I can be effectively administered by relatively inexperienced clinicians to diagnose older psychiatric patients reliably. Directions that future research might take are offered.

KEY WORDS: Structured Clinical Interview for DSM-III-R; DSM-III-R; reliability; diagnosis; older adults.

[1] Center for Psychological Studies, Nova University, 3301 College Avenue, Fort Lauderdale, Florida 33314.

[2] To whom correspondence should be addressed.

0882-2689/93/1200-0347$07.00/0 © 1993 Plenum Publishiing Corporation

INTRODUCTION

Before the introduction of well-operationalized criteria for psychiatric diagnosis in the DSM-III (American Psychiatric Association, 1980) and DSM-III-R (American Psychiatric Association, 1987), attempts to achieve adequate reliability of diagnosis often produced results that were below scientifically acceptable limits (Frank, 1975; Hersen & Bellack, 1988). In a review of early reliability studies of psychiatric diagnosis, this task was described as a "hopeless undertaking," given the uniformly poor results obtained for even major diagnostic categories (Grove, Andreason, McDonald-Scott, Keller, & Shapiro, 1981). Not only have the well-defined and specified criteria of the improved DSM contributed to diagnostic reliability, but the appearance of structured and semistructured interview schedules has greatly facilitated the task (Grove, 1987; Grove *et al.*, 1981; Hersen & Bellack, 1988; Rubinson & Asnis, 1989).

The Structured Clinical Interview for DSM-III (SCID) is the first such interview that was specifically designed on the basis of DSM-III criteria for mental disorders (Spitzer & Williams, 1984). Four years later, the SCID was updated to reflect modifications that appeared in DSM-III-R (Spitzer, Williams, Gibbon, & First, 1988). Since then, this instrument has been widely used in research settings either to select or to describe particular diagnostic groups (see Spitzer, Williams, Gibbon, & First, 1992; Williams *et al.*, 1992). The reliability (interrater or test–retest methods) of the SCID in adult populations with diverse disorders has been evaluated in a number of studies (e.g., Riskind, Beck, Berchick, Brown, & Steer, 1987; Skre, Onstad, Torgerson, & Kringlen, 1991; Williams *et al.*, 1992). A review of these investigations has recently been completed (see Segal, Hersen, & Van Hasselt, in press), indicating a good interrater reliability for broad diagnostic categories as well as specific disorders.

Surprisingly, only one investigation to date has specifically evaluated the reliability of the SCID in older adults. As part of their study comparing several screening scales for depression in the elderly, Stukenberg, Dura, and Kiecolt-Glaser (1990) reported interrater 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 reported, and reliability for problems other than mood disorders was not assessed. Given the recent interest in assessment and treatment of emotional disorders in the elderly, it seems important to ascertain the reliability of the SCID in this still underserved and underresearched population.

Williams *et al.* (1992) astutely point out that reliability of a structured interview is not a static statistic that remains constant from one study or situation to the next. To the contrary, the reliability of interviewer-administered instruments is affected by many factors, such as the characteristics of the interviewers and the subject sample, type of reliability assessed (e.g., interrater or test–retest), and reliability of the diagnostic criteria (Williams *et al.*, 1992). Thus, it cannot be assumed that the SCID automatically will be reliable if administered to older adults or other untested populations. The purpose of the present study, therefore, was to evaluate the interrater reliability of the SCID for diagnosing psychiatric disorders in older adults.

Besides examining the reliability of the SCID in an elderly population, the present study differs from previous reliability investigations along an important dimension: the level of training of the initial interviewers. Spitzer and Williams (1983), in their instruction manual for the SCID, suggest that this instrument should be administered by trained interviewers who have a background in psychopathology and DSM criteria. After all, the SCID was designed to approximate the diagnostic flowchart used by experienced diagnostic interviewers. Given its flexible, semistructured format, proper administration often requires that interviewers probe, restate, or clarify questions in ways that are sometimes not clearly outlined in the manual to judge accurately if particular symptom criteria have been met. The task, therefore, requires that the SCID assessor have a working knowledge of psychopathology and DSM-III-R, as well as basic clinical and interviewing skills. However, Rubinson and Asnis (1989) note that reliability estimates obtained in several recent investigations may be spuriously high in light of the selection of highly trained and committed clinical researchers with M.D. and Ph.D. degrees.

Interviewers in the present study were nine second- or third-year Master's-level graduate students in clinical psychology, whose training consisted of coursework in psychopathology, clinical interviewing, and role-played SCID interviews. However, such Master's-level clinicians are prototypical of the mental health professionals who should be administering the SCID in clinical settings (e.g., psychiatric inpatient and outpatient facilities, community mental health centers). Basically, the clinical and practical utility of the SCID will be further enhanced if reliability of the instrument is documented with somewhat less experienced and less rigorously trained clinicians.

Finally, it should be noted that rather than collecting our data in a highly structured clinical-research setting, we used the SCID in two ongoing clinical settings (an outpatient clinic for community-dwelling older adults and an intermediate-term residential treatment facility for chronically mentally ill older adults) during the course of a comprehensive diagnostic evaluation and assessment. Overall, administration of the SCID by Master's-level clinicians during the course of clinical work should provide a strong test of its application to a specialized patient population of older adults.

METHOD

Subjects

Two groups of patients 55 years of age and older served as subjects. Included were patients from the Nova Community Clinic for Older Adults, a university-based outpatient facility (*N* = 24; 10 males, 14 females), and the Nova Geriatric Institute, an intermediate-term residential university-based psychiatric facility for more severely disturbed chronic patients (*N* = 9; 5 males, 4 females). The final sample (*N* = 33) consisted of 15 men (45%) and 18 women (55%). The mean age was 67.33 years (*SD* = 8.40 years), with a range from 56 to 84 years. Patients were predominantly White (97%) and resided in private homes (34%), apartments and condominiums (45%), or group boarding homes (21%). Forty-five percent of the patients were married, 27% were divorced, 14% were separated, and 14% were widowed. Social class, characterized by Hollingshead (1957) Two Factor Index scores, yielded the following levels: 2 (35%), 3 (35%), 4 (20%), and 5 (10%).

Assessment Strategy

The SCID-I (patient edition: developed June 1, 1988) (Spitzer *et al.*, 1988) was administered in our study, since it is the most current version of this instrument for Axis I diagnoses and has been revised to reflect modifications in DSM-III-R. The SCID-I is designed to extract information to make diagnostic decisions for 33 of the more commonly diagnosed DSM-III-R Axis I disorders in adults. All interviews were audiotaped or videotaped for post hoc review by the reliability assessor (D.L.S).

Nine second- and third-year Master's-level students working toward their doctoral degrees served as initial SCID interviewers. The senior author, a doctoral-level psychologist (D.L.S), served as the independent reliability assessor for each taped interview. Using SCID-I criteria, he independently rated the audiotape or videotape of the initial diagnostic interview to reach a diagnostic conclusion. Current DSM-III-R Axis I diagnoses were obtained, and reliability was assessed for major depressive episode, as well as the broad diagnostic categories of anxiety and somatoform disorders.

Procedures

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

Information provided 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 in either the first or the second assessment session, as part of a comprehensive intake assessment battery that included measures of depression, hopelessness, anxiety, marital adjustment, sexual attitudes, memory functioning, and social support. After administering the SCID, assessors filled out a summary sheet detailing all DSM-III-R current 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

Current DSM-III-R Axis I diagnoses based on SCID evaluations are presented in Table I for the 33 patients. Multiple diagnoses were permitted provided that patients met the full criteria for each disorder. As can be seen, 19 patients were diagnosed with mood disorders by one assessor, while only 15 were similarly diagnosed by the second assessor. First and second assessors diagnosed four and six anxiety disorders, respectively, while both agreed on the presence of four somatoform and two adjustment disorders. Several patients had two disorders, while only one patient was rated as meeting criteria for three disorders.

Agreement between assessors was calculated using the kappa correlation coefficient and percentage agreement, which are commonly used statistics in similar reliability studies (see Arntz, van Beijsterveldt, Hoekstra, Eussen, & Sallaerts, 1992; Riskind *et al.*, 1987). Unlike percentage agreement, kappa corrects for chance levels of agreement (Fleiss, 1981). While there are no definitive guidelines for the interpretation of kappa

Table I. Current DSM-III-R Axis I Disorders Based on SCID Evaluations (N = 33)

	Number of cases[a]	
Diagnosis	Assessor 1	Assessor 2
Mood disorder	19	15
Bipolar disorder manic	0	0
Bipolar disorder depressed	3	1
Major depression	14	13
Dysthymia	2	1
Anxiety disorder	4	6
Panic disorder	3	3
Generalized anxiety disorder	1	1
Social phobia	0	2
Somatoform disorder	4	4
Somatization	2	2
Somatoform pain disorder	2	2
Psychoactive substance use disorder	2	1
Alcohol abuse or dependence	2	1
Adjustment disorder	2	2
No DSM-III-R Axis 1 disorder	8	11

[a] Numbers do not add to 33 because patients could have more than one disorder.

coefficients, the following guidelines are often used: values above .75 are considered excellent, while values from .60 to .75 suggest good agreement. Kappa values between .50 and .60 indicate moderate agreement, while values below .50 are poor. Values below .0 indicate less than chance disagreement.

As kappas are unstable at extremely low base rates, reliability statistics were computed only for those disorders where more than 10% (N > 3) of the sample was so diagnosed by one of the assessors. Thus, reliability coefficients were determined for the major depressive episode (47% base rate) and the broad diagnostic categories of anxiety disorders (15% base rate), and somatoform disorders (12% base rate). Base rate

Table II. SCID Interrater Reliability and Percentage Agreement

Diagnosis	Kappa	Percentage agreement
Major depressive disorder	.70[a]	85
Anxiety disorder (general)	.77[a]	94
Somatoform disorder (general)	1.00[a]	100

[a] $p < .001$.

was calculated as the percentage of subjects with the disorder out of the total sample averaged over both assessors. Estimates of other individual disorders (i.e., bipolar, dysthymia, panic, generalized anxiety, social phobia, and alcohol abuse) were not calculated due to low base rates in our sample; rather, they were included under broader diagnostic categories where applicable. Diagnoses were evaluated on a dichotomous basis: either present or absent. For example, with depression, a match was obtained if both assessors agreed on the presence or absence of this disorder for a patient, regardless of whether other diagnoses also were recorded. For the broad diagnostic categories (anxiety and somatoform), 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.

Interrater reliability estimates for current disorders with adequate base rates are shown in Table II. Results indicate high interrater reliability as evaluated by kappa (.70 to 1.00) and percentage agreement (85 to 100%). Overall agreement for all diagnoses was 91%.

DISCUSSION

The present study is one of the first to evaluate the reliability of the SCID in older adults. Our results suggest that this instrument can be applied reliably with this population during the natural course of clinical services. Indeed, we obtained very good diagnostic reliability for major depressive disorder, anxiety disorders, and somatoform disorders. Also, of some importance, we found that newly trained Master's-level clinicians were able to administer the SCID appropriately to older adults. In fact, the use of graduate-level assessors, who attained reliability estimates comparable to those of more seasoned clinical researchers, certainly underscores the value of the SCID.

The results we obtained are quite comparable to agreement rates found in similar SCID reliability investigations with younger adult patients. For example, Riskind *et al.* (1987) videotaped 75 psychiatric outpatients to assess the interrater reliability of DSM-III major depression (kappa = .72, percentage agreement = 82). For major depression, we obtained a kappa of .70 and a percentage agreement of 85.

Similarly, our results for major depression are comparable to data obtained by developers of the SCID in their extensive test–retest reliability project (Williams *et al.*, 1992). Williams *et al.* (1992) included 592 subjects in four patient and two nonpatient sites in the United States and one patient site in Germany. Randomly matched pairs of two professionals independently evaluated and rated the same patient within a 2-week period. Levels of agreement for current and lifetime disorders were presented for many disorders. In the patient sample (N = 390), the kappa for current major depression was .64; our kappa with a much smaller sample of patients suffering from depression was similar (.70). Additionally, our overall percentage agreement rate (91%) was slightly higher (83%) than that obtained by Riskind *et al.* (1987). We also obtained a kappa of .77 for anxiety disorders, which is comparable to the kappa of .82 found by Skre *et al.* (1991) in their audiotaped reliability study of Norwegian twins.

Despite our encouraging results with a relatively small sample of patients, limitations of the present study should be noted. Our small sample size, of course, precluded evaluation of reliability for many individual disorders (e.g., dythymia, bipolar disorder, panic disorder) that could not be assessed due to low base rates in our clinics. Also, the simultaneous rating design that was employed generally results in somewhat inflated reliability coefficients compared to the more stringent test–retest strategy. However, the simultaneous rating design fits well with a normal flow of patients and does not cause any sort of clinical disruption, as would be the case with a test–retest design.

There is no doubt that replication of this study with larger sample sizes of older adults clearly is warranted. Also, future work should focus on evaluating the psychometric properties of additional versions of the SCID, such as the SCID-II [personality disorders (Spitzer & Williams, 1986)] with older patients.

Although structured interviews have been criticized on the grounds that they appear mechanical and rigid, thus interfering with the development of rapport (see Rubinson & Asnis, 1989), we found very few instances when the SCID was rejected by our population of older adults. This occurred only when the patients presented with such acute distress that no formal assessment was possible until initial crises were resolved. In contrast, most patients seemed pleased about the comprehensive assessment provided by the SCID. More importantly, administration of the SCID by our graduate-level assessors resulted in exploration of fruitful areas that might have been glossed over or neglected altogether in an unstructured evaluation. Similarly, the SCID was frequently found to be an excellent training experience, in that it facilitated understanding and memorization of diagnostic criteria for all disorders it covers.

From our recent clinical experience in employing the SCID with older adults, it is clear that this assessment tool merits increased application in many clinical settings. While further documentation of its reliability with an older adult psychiatric population is warranted, we contend that the SCID should be routinely applied in order to insure that appropriate questions are raised and answered during the course of the initial diagnostic appraisal. Given its clinical and heuristic value, clinicians in both inpatient and outpatient settings serving older patients would be remiss if they failed to include this instrument in their assessment battery.

REFERENCES

American Psychiatric Association (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: APA.

American Psychiatric Association (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Washington, DC: APA.

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

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley & Sons.

Frank, G. (1975). *Psychiatric diagnosis.* Oxford: Pergamon Press.

Grove, W. M. (1987). The reliability of psychiatric diagnosis. In C. G. Last & M. Hersen (Eds.), *Issues in diagnostic research* (pp. 99-119). New York: Plenum Press.

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

Hersen, M., & Bellack, A. S. (1988). DSM-III and behavioral assessment. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (pp. 67-84). New York: Pergamon Press.

Hollingshead, A. B. (1957). *Two factor index of social position* (available from A. B. Hollingshead, 1965 Yale Station, New Haven, CT 06520).

Riskind, J. H., Beck, A. T., Berchick, R. J., Brown, G., & 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. *Archives of General Psychiatry, 44,* 817-820.

Rubinson, E. P., & Asnis, G. M. (1989). Use of structured interviews for diagnosis. In S. Wetzler (Ed.), *Measuring mental illness: Psychometric assessment for clinicians* (pp. 43-66). Washington, DC: American Psychiatric Press (in press).

Segal, D. L., Hersen, M., & Van Hasselt, V. B. (in press). Reliability of the Structured Clinical Interview for DSM-III-R (SCID): An evaluative review. *Comprehensive Psychiatry.*

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

Spitzer, R. L., & Williams, J. B. W. (1983). *Instruction manual for the Structured Clinical Interview for DSM-III (SCID).* New York: Biometrics Research Department, New York State Psychiatric Institute.

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

Spitzer, R. L., & Williams, J. B. W. (1986). *Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II-R, 2/1/86).* New York: New York State Psychiatric Institute.

Spitzer, R. L., Williams, J. B. W., Gibbon, M., & First, M. B. (1992). The Structured Clinical Interview for DSM-III-R (SCID). *Archives of General Psychiatry, 49,* 624-629.

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

Stukenberg, K. W., Dura, J. R., & Kiecolt-Glaser, J. K. (1990). Depression screening scale validation in an elderly, community dwelling population. *Psychological Assessment, 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., & Wittchen, H. (1992). The Structured Clinical Interview for DSM-III-R (SCID): Multisite test–retest reliability. *Archives of General Psychiatry, 42,* 630-636.