Reliability of the Structured Clinical Interview for DSM-III-R: An Evaluative Review

Daniel L. Segal, Michel Hersen, and Vincent B. Van Hasselt

Research evaluating the reliability of the Structured Clinical Interview for DSM-III-R (SCID) is reviewed. Reliability procedures and studies are examined. Several versions of the SCID are covered, including the SCID-I (axis I disorders), SCID-II (axis II disorders), SCID-Positive and Negative Syndrome Scale (SCID-PANSS; functional-dimensional assessment for psychotic disorders), and SCID–Upjohn Version (panic disorder). The SCID has been found to yield highly reliable diagnoses for most axis I and axis II disorders. Suggestions for future research on the SCID are offered, particularly with respect to (1) the lack of studies in which SCID diagnoses are compared with diagnoses from unstructured interviews or other structured-interview formats, and (2) the need for a more natural evaluation of this instrument. Also, the importance of establishing norms and obtaining reliability data for underserved clinical populations is discussed.

THE STRUCTURED Clinical Interview for DSM-III[1] (SCID-P) and DSM-III-R,[2] its revision, have enjoyed widespread popularity as a strategy to obtain reliable and valid psychiatric diagnoses during the course of clinical trials for a variety of nosological entities.[3] However, despite its popularity in academic psychiatry and clinical psychology, the SCID unfortunately is not routinely applied in most clinical settings to yield diagnostic conclusions. The unstructured, albeit time-honored, clinical interview still predominates. Thus, the lag between empirical findings and actual clinical implementation is woefully apparent. However, the heartening news is that there is a growing but still small body of literature on the reliability of this instrument and its variants that should further encourage its application.

In this review, we critically evaluate the extant literature on the SCID. We begin by briefly looking at the emergence of the SCID within its historical context. This is followed by examination of the reliability procedures that have been performed to evaluate the SCID. Next, we consider in some detail the several reliability studies that have recently appeared in the literature. Finally, we outline future research that would enhance the likelihood of a more universal application of the SCID. Specific gaps in the research literature are identified that need to be addressed in subsequent studies.

HISTORICAL CONTEXT

Until the advent of DSM-III[4] and DSM-III-R[5] and the development of structured interviews, there had been significant problems in the formulation of accurate and reliable psychiatric diagnoses. Unfortunately, confusion, uncertainty, and overlap of diagnoses historically have been the rule rather than the exception, with the unacceptable low reliability of even major diagnostic categories reported.[6-9]

The importance of obtaining reliable diagnoses is considerable, especially since reliability is a necessary (albeit insufficient) prerequisite for validity. If a diagnostic interview does not provide reproducible data when readministered under identical conditions (i.e., poor reliability), what the instrument purports to measure is inconsequential.[10] Once acceptable levels of reliability have been achieved, validity studies can then be performed. For example, differences between diagnostic groups can be compared and related to other assessment findings or treatment outcome.

DSM-I[11] and DSM-II[12] had been criticized for their poor reliability and questionable validity. Second, the lack of a clear association between obtained diagnoses and subsequent treatment planning efforts has been cited as a shortcoming of these earlier classification systems.[8] However, reliability was greatly improved with the introduction of operationalized, specified, and standardized criteria for mental disorders in DSM-III,[4] as well as the construction of standardized structured diagnostic interviews.[6,7] In a comprehensive review of

From the Center for Psychological Studies, Nova University, Fort Lauderdale, FL.

Address reprint requests to Michel Hersen, Ph.D., Nova University Center for Psychological Studies, 3301 College Ave, Fort Lauderdale, FL 33314.

0010-440X/94/3504-0005$03.00/0

major reliability studies, Grove[7] notes that "many diagnoses have been made more understandable and more reliable by introducing specified diagnostic criteria and structured interviews" (p. 115).

A number of structured interviews have been used over the past several years in clinical research.[9] These strategies were designed to improve reliability by standardizing questions asked of patients and providing guidelines for categorizing or coding responses. Adoption of such procedures served to reduce variability (and potential sources of error) among interviewers.

Three instruments widely used prior to DSM-III and DSM-III-R were the Schedule for Affective Disorders and Schizophrenia,[13] the Present State Examination,[14] and the National Institute of Mental Health Diagnostic Interview Schedule (DIS).[15] These schedules yielded much improved reliability over earlier efforts, in which unstructured interviews were performed to diagnose psychopathology.[7,8] However, the Schedule for Affective Disorders and Schizophrenia and Present State Examination were developed before DSM-III and its clear-cut standardized criteria, a limitation for which these instruments may be criticized. Although the DIS has been adapted to make diagnostic evaluations based on DSM-III criteria, it is a highly structured interview (i.e., standard questions and a dichotomous yes/no forced-choice response format) designed for administration by nonclinicians; clinical judgment is not needed for administration, and diagnosis is made by computer. However, concerns have been raised as to whether valid diagnoses can be formulated with such a totally structured interview administered by trained lay people and interpreted by computer.[16] Indeed, Spitzer[17] earlier examined this issue and concluded that clinicians were probably necessary to the diagnostic process, thus questioning the apparent value of the DIS.

In response to shortcomings of previous structured interviews, the SCID[1] is the first such approach that was based solely on DSM-III criteria for mental disorders. Consistent with its close ties to DSM-III, formal diagnostic criteria are embedded in the context of the SCID, thus permitting interviewers to make direct queries about specific features that contribute to the overall diagnostic picture. Additionally, the format and sequence of the SCID are designed to approximate the flow-chart and decision trees used by experienced diagnostic interviewers. Although the SCID provides structure to cover criteria for each disorder, it allows for flexibility, in that the interviewer can (1) probe and restate questions, (2) challenge the respondent, and (3) ask for further clarification if necessary to make a determination as to whether a particular symptom of a disorder is present. Currently, the SCID has two standard versions designed for administration with separate populations, patient and nonpatient. These versions primarily differ with respect to the orientation provided to the patient/subject and the extent to which psychotic symptoms are evaluated.[16,18]

The SCID has undergone several revisions and expansions since its original development. First, it has been updated to reflect changes in the DSM-III-R.[2] Second, questions and criteria have been revised to match current standards. Third, the SCID has been broadened to encompass disorders that were not evaluated in the original version. For example, a supplementary version has been constructed to provide DSM-III-R diagnoses of personality disorders (SCID-II).[19] Further, the SCID has been modified to provide a two-tiered categorical and dimensional assessment of psychotic disorders.[20]

Williams et al.[21] note that "The widespread use of the SCID is attested to by more than 100 published studies that have used the instrument to select or describe their study samples" (p. 630). Indeed, several recent investigations have used the SCID-I to identify patients with schizophrenia,[3] major depression,[22,23] panic disorder with agoraphobia,[24,25] posttraumatic stress disorder,[26,27] and schizophrenia.[3]

Further, in recent studies, the SCID-II has been used to assess personality disorders that were validated in a longitudinal design using an inpatient sample,[28] as well as to compare the comorbidity of personality diagnoses with another structured interview, the Personality Disorder Examination,[29] in two investigations.[30,31] In yet another concurrent validity study, Renneberg et al.[32] compared the SCID-II with a self-report personality inventory, the Millon Clinical Multiaxial Inventory.[33]

Overall, the SCID has had a major impact on

psychiatric research in that diagnosis has been made more reliable and potentially more valid. The SCID has also proved to be a promising instrument in transcultural psychiatric research. For example, this instrument has been translated into Chinese to assess differences in diagnostic practices between Western and Chinese psychiatry.[34] Similarly, a Dutch version of the SCID for dissociative disorders (SCID-D) has been used to improve assessment and diagnosis of dissociative symptoms and disorders in The Netherlands.[35]

RELIABILITY PROCEDURES

In general, two strategies have been implemented to assess reliability of the SCID. In the "test-retest method," two (or more) clinicians interview the same patient on separate occasions, with clinicians formulating independent diagnoses. Reliability in test-retest investigations refers to the extent of agreement between multiple raters as to the presence or absence of a disorder. A major source of variance in this type of study is "information variance," in which separate interviewers elicit conflicting information from the same respondent. For example, a subject may deny experiencing suicidal thoughts to one interviewer, but admit such concerns to another. Information variance can be due to differences in the manner in which clinicians phrase questions, probe about symptoms, or form a therapeutic alliance with the patient. Also, the subject may contribute to information variance by responding inconsistently from one interview to the next.

In test-retest reliability investigations, there typically is a specified time by which the second interview must be completed to decrease the probability that the subject will change his/her symptom picture between interviews. Where the test-retest method has been used, the second reliability interview usually has been conducted at least 24 hours but no longer than 2 weeks after the initial interview.[21,36,37] Although test-retest research can focus on longer-term reliability, we were unable to find any reports that had intervals greater than 2 weeks. An advantage of the test-retest method is that it approximates "true clinical practice," in that patients are commonly assessed by several clinicians with distinct orientations and interviewing styles. However, reliability estimates derived from this method often are lower than those obtained in investigations using other reliability procedures.

The second reliability strategy is referred to as the "joint interview" or "simultaneous rating" design. Here, the same interview is scored by at least two different raters who make independent diagnoses. This may be performed "in vivo" when multiple raters simultaneously observe the same interview, or one rater may conduct the interview while additional raters are present and making judgments. A common variation involves audiotaped or videotaped interviews with post hoc analyses. Whether ratings are made live or post hoc, second or additional raters are "blind" to diagnoses obtained on the basis of the "live" SCID, thus yielding an independent diagnostic appraisal. In many cases, the initial interviewer's diagnosis is included as a data point, which then is compared with judgments from at least one rater who made simultaneous live ratings or used taped interviews for post hoc ratings.

The type of variance encountered in simultaneous- or joint-interviewer designs is referred to as "rater variance," since raters are presented with the same responses from subjects but may score responses in different ways. For example, one rater may judge a symptom criterion for depression to be present (e.g., poor concentration), while another may judge the same response to indicate a subthreshold level of the symptom.

Advantages and shortcomings of both reliability strategies have been articulated elsewhere[7] and need not be discussed at length here. However, it should be noted that Grove[7] contends that the ideal procedure would be to combine both methods in one study. This would involve test-retest of subjects with different interviewers in short- and long-term intervals, while other raters make simultaneous or post hoc ratings of each interview. Despite the expense and effort required, the use of such a procedure appears justified by the potential quality of data obtained.

Whether the test-retest or joint-interviewer design is used, interrater reliability is typically determined by statistics of percentage agreement and/or the kappa index.[7] Unlike percent-

age agreement, kappa corrects for chance levels of agreement.[38] Although there are no definitive guidelines for the interpretation of the kappa index, values greater than .70 are considered to reflect good agreement, whereas values from .50 to .70 suggest fair agreement. Kappa values less than .50 indicate poor agreement, and values less than 0.0 reflect less than chance disagreement.

RELIABILITY STUDIES

Despite increasing use of the SCID in recent clinical research, surprisingly few reliability studies on this instrument have been conducted to date. Table 1 presents investigations that reported reliability data for various versions of the SCID-I. For many of these efforts, estimations of reliability were only a part of larger validity studies. The most extensive study was performed by the original developers of the SCID.[21] This project included test-retest reliability interviews of 592 subjects in four patient and two nonpatient sites in the United States and one patient site in Germany. There were 25 interviewers, all mental health professionals, who were trained in administration of the SCID.

Table 1. Reliability Studies for the SCID-I (axis I disorders)

Reference/SCID Version	Method	N	Disorder	Kappa
Riskind et al.[39] (1987)/SCID-I	SR	75	Major depression	.72
			Generalized anxiety	.79
Malow et al.[36] (1989)/SCID-I	T-RT	29	Current diagnosis:	
			Cocaine dependence	.89
			Opioid dependence	.73
			Lifetime diagnosis:	
			Cocaine dependence	.84
			Opioid dependence	.80
Stukenberg et al.[23] (1990)/ SCID-NP (nonpatients)	SR	75	Mood disorders (general category)	.92
Skre et al.[40] (1991)/SCID-I	SR	54	Schizophrenia	.94
			Major depression	.93
			Dysthymia	.88
			Cyclothymia	.80
			Bipolar disorder	.79
			Generalized anxiety	.95
			Panic disorder	.88
			Posttraumatic stress	.77
			Social phobia	.72
			Simple phobia	.70
			Anxiety disorder (NOS)	.59
			Obsessive-compulsive	.40
			Agoraphobia without panic	.32
			Alcohol abuse or dependency	.96
			Other substance abuse	.85
			Adjustment disorder	.74
			Somatoform disorder	−.03
Williams et al.[21] (1992a)				
SCID-P (patients)	T-RT	390	21 disorders	.61*
SCID-NP (nonpatients)	T-RT	202	16 disorders	.37*
Williams et al.[37] (1992b)/ SCID–Upjohn version	T-RT	72	Panic disorder	.87
			Subtypes:	
			Uncomplicated	.73
			With limited phobic avoidance	.61
			Agoraphobia with panic attacks	.66
Segal et al.[47] (1993)/SCID-I	SR	33	Major depression	.70
			Anxiety disorders	.77
			Somatoform disorders	1.00

Abbreviations: T-RT, test-retest; SR, simultaneous rating, including live, audiotaped, and videotaped interviews.

*Kappas reflect overall weighted means for all disorders combined. Kappas for some of the individual disorders are substantially higher.

Randomly matched pairs of two professionals independently evaluated and rated the same subject within a 2-week period. All SCID interviews were audiotaped and reviewed retrospectively. The sample included 390 patient and 202 nonpatient subjects for whom levels of agreement for current and lifetime disorders were presented. In both the patient and nonpatient groups, reliability data (in the form of kappa index) were reported for disorders that occurred with at least minimal frequency, defined by a sample of 10 subjects for any particular diagnosis. Base rates were noted for all disorders. Categories that were rarely diagnosed included somatization disorder and adjustment disorders.

Results indicated that in the patient sample, kappas for current disorders ranged from a low of .40 (dysthymia; base rate, 6%) to a high of .86 (bulimia nervosa; base rate, 8%). Kappas for some common current disorders were as follows: bipolar disorder (.84; base rate, 14%), major depression (.64; base rate, 31%), schizophrenia (.65; base rate, 11%), alcohol abuse/dependence (.75; base rate, 5%), drug abuse/dependence (.84; base rate, 16%), panic disorder (.58; base rate, 9%), and generalized anxiety disorder (.56; base rate, 2%). Combining all disorders yielded an overall weighted kappa of .61 for current disorders and .68 for lifetime disorders. Reliability generally was poor in the nonpatient sample, with an overall weighted kappa of .37 for current disorders and .51 for lifetime disorders. Kappas for current disorders were reported for only four categories that were diagnosed a minimum of 10 times altogether by either rater, major depression (.42; base rate, 4%), panic disorder (.59; base rate, 2%), social phobia (.41; base rate, 3%), and simple phobia (.48; base rate, 4%). Still, it is not surprising that these lower base rates corresponded to poorer kappas. Williams et al.[21] compared SCID kappa values with reliability data from other structured interviews such as the Schedule for the Assessment of Depression and Schizophrenia/Research Diagnostic Criteria and DIS. These investigators concluded that the reliability of the SCID "is roughly similar, across categories, to that obtained with other major diagnostic instruments" (p. 636). Even though reliability generally was adequate and comparable to that of other instruments, Williams et al.[21] noted two factors that may have precluded attainment of even higher reliability values. First, the SCID was being revised and updated during the initial 6 months of the study. This was necessary to mirror changes in DSM-III that were taking place during its revision to DSM-III-R. It was suggested that higher reliability coefficients may have emerged had the final version of the SCID been used throughout the investigation, rather than less specific (and possibly less adequate) versions. Second, reliability was examined for a wide range of disorders, rather than for a preselected and more limited range of problems. Williams et al.[21] suggest that the higher base rates and possible specialized training or experience of interviewers in more focused studies may have increased reliability in those endeavors. However, in the absence of empirical verification, it is unclear whether results would have been improved had such shortcomings in the design been eliminated.

Smaller-scale reliability investigations also have been performed. For example, Riskind et al.[39] videotaped 75 psychiatric outpatients to assess interrater reliability (kappa index) of DSM-III major depressive disorder (n = 25, 32% base rate) and generalized anxiety disorder (n = 24, 33% base rate). Results showed that the SCID produced reliable diagnoses for major depressive disorder (kappa, .72) and generalized anxiety disorder (.79). A strength of this study is that the full SCID sections for the two disorders were administered, without allowing interviewers to omit parts of a section if initial symptoms were absent. This procedure reduced a potential source of cuing by the initial interviewer that might artificially inflate agreement rates. Other positive aspects of the investigation include (1) presentation of standard screening information to both the interviewers and videotape raters, thus reducing information variance; (2) rotation of raters between the roles of initial interviewer and videotape rater; and (3) presentation of base rate data to clarify kappa values.

In a similar study evaluating a broader range of DSM-III-R axis I disorders, Skre et al.[40] assessed interrater reliability using 54 audiotaped SCID interviews of adults participating in a larger Norwegian twin study of mental illness.

Base rates were reported for each diagnosis, and kappas were reported for all diagnoses represented by two or more cases. Excellent interrater agreement was obtained for the broad diagnostic categories of psychotic disorder (kappa, 1.00), mood disorder (.93), anxiety disorder (.82), and psychoactive substance use disorder (.93); virtually no agreement was found for the general category of somatoform disorder (−.03). Reliability was generally lower for specific diagnoses, but was still high for schizophrenia (kappa, .94), major depression (.93), dysthymia (.88), generalized anxiety disorder (.95), panic disorder (.88), alcohol use disorder (.96), and other nonalcohol psychoactive substance use disorders (.85). Moderate agreement was found for cyclothymia (.80), posttraumatic stress disorder (.77), social phobia (.72), simple phobia (.70), bipolar disorder (.79), and adjustment disorder (.74). Poor reliability was indicated for obsessive-compulsive disorder (.40) and agoraphobia without history of panic disorder (.32). However, base rates were extremely low (i.e., $<10\%$, $n < 5$) for bipolar disorder ($n = 2$), agoraphobia without history of panic disorder ($n = 2$), posttraumatic stress disorder ($n = 4$), somatoform disorder ($n = 4$), and adjustment disorder ($n = 4$). As such, kappas are unstable for disorders with extremely low base rates, and, of consequence, should be interpreted with caution.[7]

In innovative fashion, Skre et al.[40] also evaluated reliability for combinations of diagnoses (e.g., mood and anxiety disorders). Reliability was acceptable to good for combinations of two disorders (range, .53 to 1.00), and poorer for combinations of three disorders (range, .38 to .87). However, this is not surprising, given that combinations of two diagnoses are less reliable than single diagnoses, and that combinations of three are less reliable than combinations of two. Diminished reliability with increased numbers of diagnostic combinations is predictable from psychometric theory, and is analogous to why Minnesota Multiphasic Personality Inventory two-point code types are less reliable than either Minnesota Multiphasic Personality Inventory scale alone. A small difference on one scale (e.g., short by one symptom) can affect reliability of one diagnosis. A small difference on either of two scales (i.e., symptom counts) can affect a dual diagnosis, and this is likely to occur quite often. Similarly, with three diagnoses, the chances of disagreement on any one scale are even further increased.

Using the test-retest method, Malow et al.[36] examined reliability of the SCID-I on a sample of 29 inpatients diagnosed with either cocaine or opioid dependence. Second SCID interviews were performed within 48 hours of initial assessment. Kappa values for current cocaine or opioid dependence were .89 and .73. Kappas of .84 and .80 were obtained for lifetime cocaine or opioid dependence. Regrettably, base rate data for this sample were not reported; this is critical, given that base rates are always helpful in understanding kappa values. As was reported by Grove, the lower the base rate (or higher, if $>.5$), the poorer the kappa.[7]

With a similar test-retest design, an extensive reliability study of the Upjohn version of the SCID was conducted at 13 international sites.[37] Training for SCID administrators was brief, consisting of a 2-day workshop format using training videotapes and role-played administrations. All patients were prescreened over the telephone to ensure the presence of some panic symptoms. Of the 72 patients who participated in the investigation, 52 (72%) were retested within 1 week of the initial administration. Agreement (kappa index) on the diagnosis of panic disorder ($n = 62$) was very good (.87), although this base rate was very high (86%). However, agreement was only fair to good for subtypes of panic disorder, i.e., uncomplicated ($n = 9$; kappa, .73), panic disorder with limited phobic avoidance ($n = 24$; kappa, .61), and agoraphobia with panic attacks ($n = 15$; kappa, .66).

The aforementioned investigations were conducted on versions and editions of the SCID-I, and targeted axis I disorders. The SCID-II has been developed to facilitate diagnosis of DSM-III-R personality disorders.[19] Table 2 presents studies that have reported reliability data for the SCID-II.

Fogelson et al.[41] evaluated interrater reliability of the SCID-II in a sample of 45 first-degree relatives of probands with schizophrenia, schizoaffective or bipolar disorder. These investigators reported intraclass correlation coefficients between two raters for five personality

Table 2. Reliability Studies for the SCID-II (axis II disorders)

Reference	*Method*	N	*Personality Disorder*	*Kappa*
Malow et al[36] (1989)	T-RT	29	Borderline	.87
			Antisocial	.84
O'Boyle and Self[45] (1990)	T-RT	5	Any personality disorder	.74
Wonderlich et al.[43] (1990)	SR	14	Obsessive-compulsive	.77
			Histrionic	.75
			Borderline	.74
			Dependent	.66
			Avoidant	.56
Brooks et al.[42] (1991)	SR	30	Paranoid	.77
			Schizotypal	.89
			Borderline	.64
			Histrionic	.43
			Narcissistic	.78
			Avoidant	.56
			Dependent	.84
			Obsessive-compulsive	.66
			Passive-aggressive	.50
			Self-defeating	.63
			Schizoid, antisocial*	
Fogelson et al.[41] (1991)	SR	45	Borderline	.82
			Schizotypal	.73
			Paranoid	.70
			Schizoid	.60
			Avoidant	.84
Arntz et al.[44] (1992)	SR	70	Avoidant	.82
			Dependent	1.00
			Obsessive-compulsive	.72
			Passive-aggressive	.66
			Self-defeating	1.00
			Paranoid	.77
			Schizotypal	.65
			Histrionic	.85
			Narcissistic	1.00
			Borderline	.79
			Total	.80
			Schizoid, antisocial, sadistic†	
Renneberg et al.[32] (1992)	SR	32	Avoidant	.81
			Obsessive-compulsive	.71
			Paranoid	.61
			Borderline	.63
			Any personality disorder	.75

Abbreviations: T-RT, test-retest; SR, simultaneous rating including live, audiotaped, and videotaped interviews.

*Kappas could not be computed due to insufficient between-subject variation for antisocial and schizoid personality disorders.

†Kappas could not be computed because no patients met criteria for schizoid, antisocial, and sadistic personality disorders.

disorders that are part of schizophrenia and affective-spectrum disorders: borderline (.82), schizotypal (.73), paranoid (.70), schizoid (.60), and avoidant (.84). Fogelson et al.[41] noted that values were in the acceptable range for personality disorders and were higher than previously reported results obtained with use of other structured interviews. Although prevalence rates for these disorders were not specifically reported, it was indicated that the rates were quite low. Given that lower base rates generally yield lower correlation coefficients,[7] one may predict from this study that correlation coefficients would be at least as high as the ones reported for generalization to clinical samples with higher base rates.

Two similar reliability investigations using variants of the joint-interviewer method evaluated personality disorders in anxious outpatients. Renneberg et al.[25] administered the SCID-II to 32 patients seeking treatment at an outpatient anxiety disorder clinic. Audiotaped

interviews were scored post hoc by a second rater to evaluate interrater reliability. Reliability estimates (kappa coefficients) were calculated for personality disorder categories where 10% or more ($n > 3$) of the sample were diagnosed as having the disorder in question by at least one rater. Disorders with low prevalence rates included schizoid, schizotypal, antisocial, histrionic, narcissistic, dependent, passive-aggressive, and self-defeating personality disorder. Kappas were reported for the following disorders: avoidant (.81), obsessive-compulsive (.71), paranoid (.61), and borderline (.63), although exact base rates were not provided. Interrater reliability for diagnosis of any personality disorder was .75. Overall, reliability of the SCID-II was considered "promising," with excellent results found for avoidant, obsessive-compulsive, and presence or absence of any personality disorder. Further, Renneberg et al.[32] suggest that the moderate levels of agreement for borderline and paranoid personality disorders may be attributable to the "low prevalence rates" of these two disorders, since kappa varies according to the occurrence rates of the disorders evaluated.

Interrater reliability (kappa and percentage agreement) of the SCID-II also was investigated in a sample of 30 outpatients with panic disorder with agoraphobia.[42] Audiotaped and videotaped SCID-II interviews were assessed retrospectively by two additional independent raters. Results revealed adequate agreement for 10 personality disorders examined, with generalized kappas ranging from .43 for histrionic to .89 for schizotypal personality disorders. Base rates were reported for each disorder, with obsessive-compulsive (27%), avoidant (27%), and paranoid (20%) personality disorder among the highest. Low prevalence rates (i.e., $< 10\%$) were revealed for schizoid (0%), antisocial (0%), schizotypal (0%), and histrionic (7%) personality disorder. Kappas were not computed for schizoid and antisocial personality disorder, as neither disorder was diagnosed by any rater. Confusingly, kappa was reported for schizotypal personality disorder, although its prevalence was noted to be 0%. A unique feature of this study is that the interrater agreement was also evaluated for every symptom or criterion item for each personality disorder. Results showed acceptable agreement (alpha $< .05$) for 110 of 112 criteria.

We should note at this point that low prevalence per se is no bar to computing kappas. Two issues are identified. First, for a given sensitivity and specificity, very low prevalences lead to low kappas. These values are still the best estimates of reliability, but they may mislead people who do not understand that kappas from a clinic with one base rate may not generalize well to a situation with another much lower or higher base rate. Second, with small sample sizes, kappa becomes badly restricted at low base rates. That is, if there are just two patients with a disorder out of 50, then a change of diagnosis on one person by just one rater produces a huge variance in kappa, which is undesirable. Thus, with these guidelines in mind, it is suggested that kappas be reported for low-prevalence disorders, or if they are not reported, then reasons for this should be described.

Focusing on personality disorders in 14 patients diagnosed with eating disorders, Wonderlich et al.[43] investigated interrater reliability by retrospectively rating audiotaped interviews. Kappas were obtained for five personality disorders that were diagnosed five or more times in a larger sample of 46 patients, i.e., obsessive-compulsive (.77), histrionic (.75), borderline (.74), dependent (.66), and avoidant (.56). Whereas base rates were provided for these disorders in the larger sample, no such data were reported for the smaller reliability sample, nor were kappas indicated for disorders with lower prevalence rates. And without base rate data, the stability of reported kappas cannot be ascertained.

A limitation of the above three investigations is that all included subjects with a narrow range of axis I disorders—anxiety disorders in two studies and eating disorders in one. This limited emphasis may have increased reliability coefficients, as base rates for selected disorders most likely were quite high. Additionally, samples were small and prevalence rates of certain disorders were low. Thus, some disorders were not diagnosed and reliability was not determined.

In a larger-scale project performed in Holland by Arntz et al.,[44] two raters observed the same interview to compare diagnoses of 70 outpatients suffering primarily from anxiety dis-

orders. Interrater reliability (kappa) of this Dutch version of the SCID-II ranged between .65 for schizotypal and 1.00 for dependent, self-defeating, and narcissistic personality disorders. However, these investigators noted that the number of patients with personality disorders was small, with the exception of patients with avoidant, paranoid, and obsessive-compulsive personality disorders; thus, reported agreement levels may be somewhat unstable. Similar to the procedure used by Brooks et al.,[42] Arntz et al.[44] also reported interrater reliability for each of the criteria for all personality disorders. Out of a total of 116 criteria, results (using intraclass correlation coefficient) suggested "excellent" reliability for 84 ($r > .75$) and "good" reliability for 14 ($r > .65$); only six criteria showed disappointing levels ($r < .60$).

Using a test-retest design, O'Boyle and Self[45] examined reliability of personality disorders in a subsample of five inpatients participating in a larger validity study. The mean interval between administrations was 1.7 days. Agreement (kappa) for presence of any personality disorder was .74. However, reliability data were not collected for individual disorders due to the extremely small sample size. Also using a test-retest design (second interview within 48 hours), Malow et al.[36] reported reliability of borderline (kappa, .87; total sample base rate, 16%) and antisocial (.84; total sample base rate, 15%) personality disorders in a subsample of 29 out of 117 male veteran inpatients diagnosed with either cocaine or opioid dependence. Surprisingly, although paranoid personality disorder was diagnosed in 7% of patients in the total sample, interrater agreement for this disorder was not reported. Limitations of this study include a small sample size and questionable generalizability due to exclusive inclusion of male veterans who were cocaine or opiate users.

A comprehensive test-retest–reliability investigation of the SCID-II recently has been conducted with 284 subjects in six diverse settings.[46] At all sites, two raters independently evaluated the same subject at separate times with intervals ranging from 1 day to 2 weeks. A report detailing the data collected in this extensive study is expected soon, and should add greatly to our current knowledge regarding reliability of the SCID-II.

Although several investigations have assessed interrater or test-retest reliability for the SCID-I and SCID-II in a variety of adult populations, only two investigations to date have specifically evaluated the reliability of the SCID in older adults. Yet, given the upsurge of interest in assessment and treatment of emotional disorders in older adults, it seems important to ascertain the reliability of the SCID with this underserved population. As part of their study comparing several screening scales for depression in the elderly, Stukenberg et al.[23] 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 provided, and reliability for problems other than mood disorders was not determined. Additionally, base rates for mood disorders were not reported for the reliability sample.

Our research group[47] evaluated the reliability of the SCID with an elderly population in a "natural" clinical outpatient setting, as part of a comprehensive assessment battery. Specifically, the SCID was administered in the first or second intake session for older adults seeking outpatient treatment for a wide range of disorders. It is important to point out that interviewers were second- and third-year graduate students in clinical psychology, and not highly trained or skilled clinical researchers as has been the norm. 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 did have prior course work in psychopathology and general interviewing strategies. Moreover, they were trained to criteria with the SCID, primarily via role-played SCID interviews.

In our study, interrater reliability was investigated in a sample of 33 older adults (age, 67.3 ± 8.4 years [mean ± SD]). Audiotaped and videotaped interviews were assessed retrospectively by an independent rater at the doctoral level. Reliability estimates (kappa and percentage agreement) were calculated for major depressive episode (47% base rate) and the broad diagnostic categories of anxiety disorders (15%

base rate) and somatoform disorders (12% base rate). Since kappas are unstable at extremely low base rates,[7] 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 percentage agreements, respectively, are as follows: major depression (.70, 85%), anxiety disorders (.77, 94%), and somatoform disorders (1.0, 100%). Thus, reliability of the SCID-I administered by graduate-level clinicians in a limited sample of older adults appears very promising, although additional research with larger samples of obviously is warranted.

CONCLUSIONS

Our review indicates that although the SCID has been widely and successfully used in controlled research for several years now, there are still gaps in the research data. For example, there are few studies in which SCID diagnoses are compared with non-SCID diagnoses for reliability. Further, even if recent SCID studies are compared with old pre–DSM-III or pre–Research Diagnostic Criteria diagnostic reliability studies, there are confounding factors in this research. The earlier studies not only did not use the SCID (or any structured interview), they did not use clear and operationalized diagnostic criteria either. Thus, any increase in reliability is difficult to interpret; improved reliability could be a function of the structured interview, better diagnostic criteria, or some combination of both. To fully assess how much the SCID contributes to increased reliability, there is a need for studies that use DSM-III-R (and eventually) DSM-IV criteria and compare the SCID with an unstructured interview or another structured interview such as the DIS-R.

Several additional shortcomings and concerns about the SCID and their implications for its application in clinical practice warrant mention and further empirical investigation. A major issue in research on reliability of the SCID is the degree to which many studies address the topic of clinical and practical utility. For example, most studies in this area have been conducted in an "artificial" research context rather than in an ongoing clinical setting. SCID interviewers usually are experienced clinical researchers who are well-trained and committed to the development and success of the instrument. Moreover, these raters have exceptional familiarity with DSM-III-R criteria and the most advanced training in psychopathology, from both research and clinical perspectives. Unfortunately, this is not the typical situation in standard clinical practice. In most mental health centers, for example, clinicians who might use the SCID are not as committed or as well trained as clinical researchers participating in reliability studies. At this point, the effect of using standard clinical-service providers (as compared with clinical researchers) on SCID reliability is unknown. Can this instrument be used to formulate reliable diagnoses with the type and level of front-line clinicians who work in mental health centers? Although results from a study conducted by our research group[47] are promising, these findings await further empirical support from research with larger and more diverse patient populations.

An additional concern is that many of the reliability studies targeted a patient sample with a rather narrow range of axis I and II disorders. For example, two investigations[25,42] exclusively included patients with anxiety disorders, and one[43] only assessed eating-disordered individuals. This limited focus most likely served to increase reliability coefficients. In addition, sample sizes and prevalence rates of certain disorders were low in many reports, thus precluding determination of diagnoses and reliability for some disorders, or reporting unstable kappas due to extreme base rates for other disorders. Future reliability studies may be enhanced by increasing the sample size and expanding the range of axis I disorders in the reliability sample to include mood, psychotic, and other axis I disorders. Otherwise, reliability estimates may be misleading, unstable, or spuriously high.

Finally, additional investigative endeavors are needed to evaluate reliability and establish norms with the SCID for use with underserved groups such as African-Americans, older adults, and physically and developmentally disabled persons. Yet another issue deserving attention is that the SCID will soon have to undergo major revision to correspond with some of the new diagnostic criteria that will appear in DSM-IV. Although current research still is evaluating

psychometric properties of the SCID based on DSM-III-R criteria, it cannot be assumed that these findings will apply to a revised version of this instrument based on yet a new modification of the diagnostic system. Thus, reliability data as to future versions of the SCID inevitably will lag behind establishment of DSM-IV criteria.

Despite these problems we have identified, there are many positive aspects of the SCID. First, the SCID has the potential to facilitate diagnosis and treatment planning, in that it can be efficiently administered as a standard part of a more extensive intake assessment. Second, structured interviews such as the SCID have been shown to generate improved diagnostic reliability relative to less structured "psychosocial" intake formats that currently prevail in clinical practice.[7] Third, the SCID facilitates DSM-III-R–based psychiatric diagnosis and covers the major axis I and all axis II disorders. Fourth, for less behaviorally oriented practitioners, the SCID will enable them to assess the frequency, duration, and intensity of specific symptoms that subsequently can be targeted for treatment. Indeed, the SCID is structured in such a manner that it prompts interviewers to evaluate specific symptom dimensions and also guides patients to describe symptoms in a most detailed fashion. Assessment and clarification of symptoms are highlighted by a unique feature of the SCID, since DSM-III-R diagnostic criteria are embedded in the interview schedule. Consequently, such information is immediately available for examination by interviewers.

REFERENCES

1. Spitzer RL, Williams JBW. Structured Clinical Interview for DSM-III Disorders (SCID-P 5/1/84). Rev. New York: Biometrics Research Department, New York State Psychiatric Institute, 1984.
2. Spitzer RL, Williams JBW, Gibbon M, First MB. Structured Clinical Interview for DSM-III-R–Patient Version (SCID-P 6/1/88). New York: Biometrics Research Department, New York State Psychiatric Institute, 1988.
3. Schooler NR, Keith SJ, Severe JB, Matthews S. Acute treatment response and short term outcome in schizophrenia. Psychopharmacol Bull 1989;25:331-335.
4. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Ed. 3. Washington, DC: American Psychiatric Association, 1980.
5. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Ed. 3. Rev. Washington, DC: Author, 1987.
6. Grove WM, Andreason NC, McDonald-Scott P, Keller MB, Shaprio RW. Reliability studies of psychiatric diagnosis. Arch Gen Psychiatry 1981;38:408-413.
7. Grove WM. The reliability of psychiatric diagnosis. In: Last CG, Hersen M (eds): Issues in Diagnostic Research. New York, NY: Plenum, 1987:99-119.
8. Hersen M, Bellack AS. DSM-III and behavioral assessment. In: Bellack AS, Hersen M (eds): Behavioral Assessment: A Practical Handbook. New York, NY: Pergamon, 1988:67-84.
9. Rubinson EP, Asnis GM. Use of structured interviews for diagnosis. In: Wetzler S (ed): Measuring Mental Illness: Psychometric Assessment for Clinicians. Washington, DC: American Psychiatric, 1989:43-66.
10. Magnusson D. Test Theory. Stockholm, Sweden: Almqvist & Wiksell, 1966.
11. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Washington, DC: Author, 1952.
12. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Ed. 2. Washington, DC: Author, 1968.
13. Endicott J, Spitzer RL. A diagnostic interview: the Schedule for Affective Disorders and Schizophrenia. Arch Gen Psychiatry 1978;35:837-844.
14. Wing JK, Birley JLT, Cooper JE, Graham P, Isaacs A. Reliability of a procedure for measuring and classifying 'present psychiatric state'. Br J Psychiatry 1967;113:499-515.
15. Robins LN, Helzer JE, Croughan J, Ratcliff KS. National Institute of Mental Health Diagnostic Interview Schedule: its history, characteristics, and validity. Arch Gen Psychiatry 1981;38:381-389.
16. Hasin DS, Skodol AE. Standardized diagnostic interviews for psychiatric research. In: Thompson C (ed): The Instruments of Psychiatric Research. New York, NY: Wiley, 1989:19-57.
17. Spitzer RL. Psychiatric diagnoses: are clinicians still necessary? Compr Psychiatry 1983;24:399-411.
18. Spitzer RL, Williams JB, Gibbon M, First MB. The Structured Clinical Interview for DSM-III-R (SCID). Arch Gen Psychiatry 1992;49:624-629.
19. Spitzer RL, Williams JBW, Gibbon M. Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II). Washington, DC: American Psychiatric, 1990.
20. Kay SR, Opler LA, Spitzer RL, Williams JBW, Fiszbein A, Gorelick A. SCID-PANSS: two-tier diagnostic system for psychotic disorders. Compr Psychiatry 1991;32: 355-361.
21. Williams JBW, Gibbon M, First MB, Spitzer RL, Davies M, Borus J, et al. The Structured Clinical Interview for DSM-III-R (SCID): multisite test-retest reliability. Arch Gen Psychiatry 1992;49:630-636.
22. Jacobs S, Hansen F, Berkman L, Kasl S, Ostfeld A. Depressions of bereavement. Compr Psychiatry 1989;30:218-224.
23. Stukenberg KW, Dura JR, Kiecolt-Glaser JK. Depression screening scale validation in an elderly, community dwelling population. Psychol Assess 1990;2:134-138.
24. Rosenbaum JF, Biederman J, Gersten M, Hirshfeld DR, Meminger SR, Herman JB, et al. Behavioral inhibition

in children of parents with panic disorder and agoraphobia. Arch Gen Psychiatry 1988;45:463-470.

25. Renneberg B, Chambless DL, Gracely EJ. Prevalence of SCID-diagnosed personality disorders in agoraphobic outpatients. J Anxiety Disord 1992;6:111-118.

26. Pitman RK, Altman B, Macklin ML. Prevalence of posttraumatic stress disorder in wounded Vietnam veterans. Am J Psychiatry 1989;146:667-669.

27. Hovens JE, Falger PRJ, Op Den Velde W, Schouten EGW, de Groen JHM, van Duijn H. Occurrence of current post traumatic stress disorder among Dutch World War II Resistance veterans according to the SCID. J Anxiety Disord 1992;6:147-157.

28. Skodol AE, Rosnick L, Kellman D, Oldham JM, Hyler SE. Validating structured DSM-III-R personality disorders assessments with longitudinal data. Am J Psychiatry 1988;145:1297-1299.

29. Loranger AW, Susman VL, Oldham JM, Russakoff LM. The Personality Disorder Examination: a preliminary report. J Pers Disord 1987;1:1-13.

30. Oldham JM, Skodol AE, Kellman HD, Hyler SE, Rosnick L, Davies M. Diagnosis of DSM-III-R personality disorders by two structured interviews: patterns of comorbidity. Am J Psychiatry 1992;149:213-220.

31. Skodol AE, Oldham JM, Rosnick L, Kellman HD, Hyler SE. Diagnosis of DSM-III-R personality disorders: a comparison of two structured interviews. Int J Methods Psychiatr Res 1991;1:13-26.

32. Renneberg B, Chambless DL, Dowdall DJ, Fauerbach JA, Gracely EJ. The Structured Clinical Interview for DSM-III-R, Axis II, and the Millon Clinical Multiaxial Inventory: a concurrent validity study of personality disorders among anxious outpatients. J Pers Disord 1992;6:117-124.

33. Millon T. Manual for the MCMI-II. Minneapolis, MN: National Computer Systems, 1987.

34. Wilson LG, Young DH. Diagnosis of severely ill inpatients in China: a collaborative project using the Structured Clinical Interview for DSM-III (SCID). J Nerv Ment Dis 1988;176:585-592.

35. Boon S, Draijer N. Diagnosing dissociative disorders in The Netherlands: a pilot study with the Structured Clinical Interview for DSM-III-R dissociative disorders. Am J Psychiatry 1991;148:458-462.

36. Malow RM, West JA, Williams JL, Sutker PB. Personality disorders classification and symptoms in cocaine and opioid addicts. J Consult Clin Psychol 1989;57:765-767.

37. Williams JBW, Spitzer RL, Gibbon M. International reliability of a diagnostic intake procedure for panic disorder. Am J Psychiatry 1992;149:560-562.

38. Fleiss JL. Statistical Methods for Rates and Proportions. Ed. 2. New York, NY: Wiley, 1981.

39. Riskind JH, Beck AT, Berchick RJ, Brown G, Steer RA. Reliability of DSM-III-R diagnoses for major depression and generalized anxiety disorder using the Structured Clinical Interview for DSM-III-R. Arch Gen Psychiatry 1987;44:817-820.

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

41. Fogelson DL, Nuechterlein KH, Asarnow RF, Subotnik KL, Talovic SA. Interrater reliability of the Structured Clinical Interview for DSM-III-R, Axis II: schizophrenia spectrum and affective spectrum disorders. Psychiatry Res 1991;39:55-63.

42. Brooks RB, Baltazar PL, McDowell DE, Munjack DJ, Bruns JR. Personality disorders co-occurring with panic disorder with agoraphobia. J Pers Disord 1991;5:328-336.

43. Wonderlich SA, Swift WJ, Slotnick HB, Goodman S. DSM-III-R personality disorders in eating-disorder subtypes. Int J Eating Disord 1990;9:607-616.

44. Arntz A, van Beijsterveldt B, Hoekstra R, Hofman A, Eussen M, Sallgerts S. The interrater reliability of a Dutch version of the Structured Clinical Interview for DSM-III-R Personality Disorders. Acta Psychiatr Scand 1992;85:394-400.

45. O'Boyle M, Self D. A comparison of two interviews for DSM-III-R personality disorders. Psychiatry Res 1990; 32:85-92.

46. First MB, Spitzer RL, Gibbon M, Williams JBW, Davies M, Borus J, et al. The Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II). II. Multisite test-retest reliability. J Pers Disord. In press.

47. Segal DL, Hersen M, Van Hasselt VB, Kabacoff RI, Roth L. Reliability of diagnoses in older psychiatric patients using the Structured Clinical Interview for DSM-III-R. J Psychopathol Behav Assess. 1993;15:347-356.