Training and quality assurance with the Structured Clinical Interview for DSM-IV (SCID-I/P)

Joseph Ventura,a,b,* Robert P. Liberman,a,b,c Michael F. Green,a,b,c Andrew Shane,a,b,c Jim Mintz,a,b,c

1Intervention Research Center for Major Mental Illness (11648), West Los Angeles V.A Medical Center, 1111 South Wilshire Boulevard, Los Angeles, CA 90073, USA
2Department of Psychiatry and Behavioral Sciences, University of California, Los Angeles, USA
3West Los Angeles Veterans Affairs Medical Center, West Los Angeles, USA

Received 26 May 1997; revised 23 February 1998; accepted 25 March 1998

Abstract

Accuracy in psychiatric diagnosis is critical for evaluating the suitability of the subjects for entry into research protocols and for establishing comparability of findings across study sites. However, training programs in the use of diagnostic instruments for research projects are not well systematized. Furthermore, little information has been published on the maintenance of interrater reliability of diagnostic assessments. At the UCLA Research Center for Major Mental Illnesses, a Training and Quality Assurance Program for SCID interviewers was used to evaluate interrater reliability and diagnostic accuracy. Although clinically experienced interviewers achieved better interrater reliability and overall diagnostic accuracy than neophyte interviewers, both groups were able to achieve and maintain high levels of interrater reliability, diagnostic accuracy, and interviewer skill. At the first quality assurance check after training, there were no significant differences between experienced and neophyte interviewers in interrater reliability or diagnostic accuracy. Standardization of training and quality assurance procedures within and across research projects may make research findings from study sites more comparable. © 1998 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Interrer reliability; Diagnostic accuracy

*Corresponding author. Tel.: +1 310 206 5225; fax: +1 310 268 4783; e-mail: jventura@uch.edu

0165-1781/98/$19.00 © 1998 Elsevier Science Ireland Ltd. All rights reserved.
PII: S0165-1781(98)00038-9
1. Introduction

The importance of reliable assessment of psychiatric diagnosis in clinical research is undisputed (Winokur et al., 1988; Mielkowicz, 1992; Williams et al., 1992; McGorry et al., 1995; van Praag, 1997). Improvements in diagnostic reliability have been obtained through the use of structured diagnostic instruments such as the Present State Exam (PSE; Luria and McHugh, 1974; Luria and Berry, 1979), the Schedule for Affective Disorders and Schizophrenia (SADS; Andreasen et al., 1982; Endicott and Spitzer, 1978), and the Structured Clinical Interview for DSM-III-R (First et al., 1987). It is generally agreed that structured interviewing techniques can be successfully taught to research staff from different disciplines and with varying amounts of prior clinical experience (Flemenebaum and Zimmermann, 1973; Luria and Berry, 1980; Ventura et al., 1993). Additional improvements in diagnostic reliability have been achieved by using operational criteria for diagnostic classification (Feighner et al., 1972; Catego System; Wing et al., 1974; Research Diagnostic Criteria; Spitzer et al., 1978; DSM-III, III-R, and IV: American Psychiatric Association, 1980, 1987, and 1994, respectively), the application of statistical methods to compute interrater agreement (Bartko and Carpenter, 1976; Faraone and Tsuang, 1994), and the use of formal training programs (Tardiff et al., 1978; Luria and Berry, 1980). Yet, even with these important advances, serious questions about diagnostic reliability and validity remain (Williams et al., 1992; McGorry et al., 1995; Mojtabal and Nicholson, 1995; Hill et al., 1996; van Praag, 1997). In fact, discrepant findings across studies of psychiatric patients may be due to differences in the way diagnostic interviews were conducted or a lack of consistency in how criteria were applied, rather than the result of true differences in the phenomena under study (Winokur et al., 1988; Williams et al., 1992; Hill et al., 1996).

Little consensus exists regarding the minimum amount of prior experience needed to begin training in structured diagnostic interviewing, standardization of the diagnostic training process (McGorry et al., 1995), or the criteria that should be used to certify and monitor diagnostic interviewers. Some reports have demonstrated that research staff with varying levels of prior clinical experience can conduct reliable diagnostic assessments after participation in a standardized training program (Flemenebaum and Zimmermann, 1973; Luria and Berry, 1980; Ventura et al., 1993). Even in studies where formal training in structured diagnostic assessment was provided, there are very few descriptions of quality assurance methods used to guard against rater drift.

The Structured Clinical Interview for DSM-IV (SCID-I/P; First et al., 1996) was developed for use in research by trained clinicians. The SCID incorporates the use of obligatory questions, operational criteria from the DSM-IV, a categorical system for rating symptoms, and an algorithm for arriving at a final diagnosis. The SCID allows the research clinician to tailor a diagnostic assessment to fit the needs of a particular research protocol or a particular patient (Spitzer et al., 1992). In addition, the SCID instructions encourage the diagnostic interviewer to use all sources of information in rating the presence or absence of a symptom or sign of psychopathology. Thus, the SCID is a good choice as a diagnostic instrument for use in research, because, while containing elements that increase reliability, several of the SCID's features facilitate the training process.

The Diagnosis and Psychopathology Unit (DPU) of the UCLA Intervention Research Center for Major Mental Illnesses has developed a Training and Quality Assurance Program for the SCID-I/P. The training program consists of videotaped SCID interviews with 'Gold Standard' consensus ratings and a team of Research Center senior diagnosticians who are mentors for training interviewers. To meet certification criteria, trainees co-rated videotaped SCID interviews and conducted SCID interviews that were co-rated by senior diagnosticians. After training was completed, a Quality Assurance (QA) program monitored rater drift by requiring trainees to co-rate and conduct one SCID interview annually. We hypothesized, based on prior studies and on our experience with the BPRS (Ventura et al., 1993),
that research staff who were inexperienced in structured interviewing could achieve acceptable levels of interrater reliability, diagnostic accuracy, and interviewer skill.

2. Methods

2.1. Participants

All participants were research faculty, research associates, or visiting scholars at the UCLA Intervention Research Center for Major Mental Illnesses. The participants were trained on the Structured Clinical Interview for DSM-IV, Patient Version 2.0 (SCID-I/P; First et al., 1996). Early training courses used the Structured Clinical Interview for DSM-III-R, Patient version (SCID-P; Spitzer et al., 1990). Interrater reliability data on the SCID were collected from January, 1991 to March, 1997. Data from quality assurance checks on drift of interrater reliability involved SCID interviews that were co-rated or conducted by the interviewer from April, 1992 to September, 1997. The version of the SCID used in the Research Center was modified to include additional items for rating the presence of delusions and hallucinations adapted from the 9th edition of the Present State Examination (PSE; Wing et al., 1974).

2.2. Training program for the SCID

Training and Quality Assurance (QA) was provided by six specially trained Research Center diagnosticians who were Board certified psychiatrists or Ph.D. level psychologists and highly reliable (see Section 2.4) in the use of structured instruments. They had a minimum of five years' experience administering the PSE, three years' experience administering the SCID, and more than five years' experience administering the Brief Psychiatric Rating Scale. Diagnosticians oriented trainees to the SCID manual, co-rated training and quality assurance SCID interviews, evaluated interviewer skill, and provided remediation when research staff failed to meet recertification criteria. The training program for the SCID utilized a library of instructional materials including: (1) glossary of the Present State Exam (Wing et al., 1974); (2) User's Guide for the SCID-I (First et al., 1996); (3) audio tapes (Falloon and Lukoff, 1984) and videotapes on psychiatric interviewing and diagnosis (Reid, 1989; Ciernyansky, 1994); and (4) videotaped SCID interviews with 'Gold Standard' consensus ratings (contact first author regarding availability of training videos). Consensus ratings for the training and QA videos were created by obtaining ratings from a minimum of three senior diagnosticians and reaching a consensus on those symptoms or diagnoses in which disagreements occurred. The following is a description of the SCID training program.

2.2.1. Orientation to psychiatric diagnosis with the SCID

The orientation phase of the training program was tailored to the trainee's prior level of experience using structured diagnostic instruments. Self-instructional training materials, such as audiotapes by Falloon and Lukoff (1984), were used to introduce trainees to the recognition and rating of various psychotic symptoms. Orientation also involved viewing videotapes demonstrating: (1) elicitation and rating of various hallucinations and delusions; (Lukoff and Ventura, 1985); (2) structured interviewing techniques (Kern, 1994); and (3) making clinical diagnoses according to DSM-IV (Ciernyansky, 1994). The trainee was also expected to become familiar with the SCID by reading the User's Guide and observing the SCID being used by a diagnostician. If trainees had extensive prior experience making diagnoses or using structured interviews with psychiatric patients, they began SCID training after becoming familiar with the SCID interview and observing interviews conducted by a diagnostician.

2.2.2. Calculation of interrater reliability for training and quality assurance

A minimum of six SCID training videos which were accompanied by consensus ratings were used for computing interrater reliability and in providing the trainee with feedback. We used χ statistics to compare the presence or absence of each symptom or critical SCID item, rather than to evaluate classification of the diagnosis. Minimum
standards of acceptable symptom agreement were an overall $k$ of 0.75, specificity of 0.75 and sensitivity of 0.75. The sensitivity and specificity $k$ values are chance-corrected indices of two well-known statistics in medical epidemiology. Sensitivity refers to the rater's ability to detect symptoms judged present, and specificity refers to the rater's ability to recognize that certain symptoms are rated as absent in accordance with the Gold Standard. Weighted $k$ statistics measure the degree of association between a test and a criterion ('gold standard') and are conceptually linked to, but distinct from, the well-known agreement $k$ discussed by Cohen (1960) and Bloch and Kraemer (1989).

These two statistics are discussed at length by Kraemer (1992), who refers to them as the quality of sensitivity $k(0.01)$ and quality of specificity $k(0.0)$, respectively (p. 94), and recommends their use in the evaluation of test performance. Unlike the familiar sensitivity and specificity indices, these weighted $k$ values are rescaled so that random performance is 0% and optimal performance is 100%. Our use of these statistics is unconventional in one important way. Rather than judging individual people as 'cases' or 'non-cases' against an external standard, trainees in our setting judged symptoms and items elicited by the SCID as 'present' or 'absent', with the consensus judgment serving as the gold standard reference. Because separate items within a scale are not independent, conventional significance testing of these weighted $k$ values was not appropriate. However, we used them only descriptively and the trainee’s judgments about the presence or absence of a symptom were not tested for statistical significance.

The trainee's diagnostic accuracy was evaluated using three levels of correctness: (1) correct; (2) partially correct; and (3) incorrect. A diagnosis was 'partially correct' if the criteria for the diagnosis closely overlapped with the consensus diagnosis. Diagnoses that were partially correct with schizophrenia included 'schizophreniform' and in some cases 'schizoaffective disorder' when the differential diagnosis was difficult because of lack of clarity regarding the overlap between mood and psychosis. Major depressive disorder NOS was partially correct when the consensus diagnosis was major depressive disorder without psychotic features. Alcohol abuse was partially correct when the consensus diagnosis of alcohol dependence was made using the minimum number of symptoms required. The trainee's diagnosis had to agree with the consensus diagnosis 90% of the time. For diagnoses that were partially correct or incorrect, immediate remediation was provided. All of the SCID modules were administered by the trainee and used to calculate inter-rater agreement. If the diagnostic and symptom interrater reliability criteria were met after viewing six videotapes of varying degrees of difficulty, the trainee continued to the next step in the training.

2.2.3 Interviewer-conducted SCID assessment

After viewing the training videotapes, trainees were required to conduct four SCID interviews on actual patients. To be certified, the trainee had to demonstrate 'Good' to 'Excellent' interrater reliability on psychiatric symptoms (mean overall $k$ of at least 0.75) and 90% accuracy in diagnosis. Symptom elicitation skill was assessed with the DPU Checklist of Interviewer Behaviors (Appendix A; Ventura et al., 1993). If the minimum criteria were not met during training, additional SCIDs were conducted until two consecutive interviews ('moving window') met all of the reliability criteria. At the time of certification, the SCID interviewer and his or her principal investigator received written feedback (see Appendix B). The report detailed the overall $k$ (and $k$ for sensitivity and specificity), diagnostic accuracy, and provided feedback on interviewer skill.

After the training was complete, diagnosticians continued to supervise diagnostic interviewers to ensure that diagnosis and symptom ratings were done reliably. The Diagnostic Procedures Checklist, which contains a list of criteria from DSM-IV for several major Axis I disorders, was used systematically to determine that all necessary diagnostic and symptom criteria were met. The project diagnosticians completed and signed the Diagnostic Procedures Checklist indicating acceptance of responsibility for the final diagnosis.
2.3. The QA program

The Quality Assurance (QA) program insured standardization and consistency of symptom assessment, diagnosis, and interviewer style across projects and over time. Two types of QA checks (passive and active) were scheduled for each interviewer annually. During the passive check, all certified SCID interviewers rated a live or videotaped SCID interview conducted by a senior Research Center diagnostician. During the active check, diagnostic interviewers conducted a live SCID interview co-rated by a diagnostician. Immediate verbal feedback was provided on reliability of symptom ratings, accuracy of diagnosis, and the strengths and weaknesses of interviewer behavior. The interviewers' reliability was evaluated using the same methods employed during SCID training. A written report was provided to the interviewer and project PI detailing the overall \( \kappa \) (and \( \kappa \) for sensitivity and specificity), accuracy of diagnosis, and interviewer style (for live QA checks only).

2.3.1. Reliability standards for QA

Reliability standards for the SCID QA Program were identical to the requirements set for initial training, but were based on a single SCID assessment. If the diagnostic or symptom rating discrepancies resulted in an overall \( \kappa \) lower than 0.75 or if the standards for interviewer skill during active QA checks were not met, the diagnostic interviewer was given a remediation session. In remediation sessions diagnostic interviewers reviewed relevant sections of the DSM-IV criteria, the PSE glossary of symptom definitions, or viewed additional SCID training videotapes that included specific symptoms of difficulty (e.g. negative symptoms). When we encountered poor interviewer style, the rater was given immediate and specific feedback regarding areas of difficulty (e.g. failure to use sufficient follow-up probes) or was asked to conduct another SCID interview co-rated by a member of the DPU.

2.4. Diagnosticians' interrater reliability and diagnostic agreement

Interrater reliability of diagnosticians on the SCID was evaluated by using videotapes with 'consensus' ratings from the Training and Quality Assurance Program and found to be in the excellent range: an overall \( \kappa \) of 0.83, specificity of 0.81 and sensitivity of 0.87. Diagnostic agreement was also evaluated for the six diagnosticians who were in full agreement 86% of the time across 29 SCID interviews. The diagnosticians' ability to reliably make Axis I or Axis II diagnoses was evaluated through classification of 20 case vignettes. The diagnosticians achieved good to excellent agreement with the research project's team diagnosis (mean \( \kappa = 0.78 \)) and with a consensus diagnosis (mean \( \kappa = 0.71 \)). These findings suggest that the Research Center diagnosticians were able to achieve good to excellent levels of diagnostic reliability.

3. Results

Analyses of SCID interviewers' interrater reliability were calculated at two time points: (1) at the completion of SCID training; and (2) at the first QA check. The SCID interviewers' diagnosis was also compared to the consensus diagnosis at the same two time points. Only the 30 diagnostic interviewers who were employed in Research Center projects and who were entered into the QA program were included in the analysis. At the completion of training, these 30 SCID interviewers demonstrated excellent agreement on SCID symptoms (overall \( \kappa = 0.85 \), range = 0.71–0.97) and very good diagnostic accuracy (82%). The \( \kappa \) on symptom agreement at the first QA check was lower than that achieved during training (overall \( \kappa = 0.76 \), range = 0.51–1.00, \( n = 30 \), but would be considered good to excellent (Fleiss, 1973). Diagnostic accuracy at the first QA check was also very good (83%).

We evaluated interrater reliability of SCID interviewers with and without previous experience in structured interviewing. The experienced group (\( n = 16 \)) had previous training in structured interviewing (using the Present State Exam or Schedule for Affective Disorder and Schizophrenia) and the neophyte group (\( n = 14 \)) did not. Most experienced staff (66%) were doctoral level psychologists or psychiatrists and many of the inex-
perceived interviewers (64%) were graduate stu-
dents in doctoral programs or individuals with a
Master’s degree in psychology. Experienced inter-
viewers were able to achieve a modestly higher
level of agreement than neophyte interviewers
(see Table 1). However, both groups achieved
excellent interrater reliability at the completion
of SCID training. One interviewer who had no
previous experience with structured interviews
had to conduct two additional SCIDs to achieve
certification compared to none of the experienced
interviewers. In addition, a multiple regression
analysis showed that prior experience with struc-
tured interviewing was a better predictor of good
interrater reliability than having a doctoral de-
gree (F(27) = 5.16, P = 0.03), even though most
doctoral staff had some prior experience. But, by
the time of the first QA check, no difference in
interrater reliability on symptoms was found
between the two groups. Regarding diagnostic
accuracy, 75% of experienced interviewers met
the criteria for diagnostic accuracy without need-
ing additional training, compared to 35% of inex-
perienced raters (χ² = 4.69, P = 0.03). However,
at the first QA check on diagnostic accuracy,
there were no significant differences between ex-
perienced and neophyte interviewers (75% vs.
93%, respectively; χ² = 2.25, P = 0.325).
We compared previously trained and neophyte
diagnostic interviewers in five areas of interviewer
competency: (1) rapport; (2) probing; (3) es-
tablishing the time frame; (4) symptom frequency
and functional impairment; and (5) organizing
and structuring the interview. An overall rating of
interviewer skill was also made on a 1 = ‘not at all’
to 5 = ‘always’ scale. Data obtained by diag-
nosticians using the DPU Checklist of Interviewer
Behaviors during training and QA checks were
combined for the purpose of this analysis. The
number of diagnostic interviewers in this analysis
n = 20) was smaller than in analyses of interrater
reliability and diagnostic accuracy because the
DPU’s Checklist of Interviewer Behaviors was de-
veloped after the training and QA program had
begun. Global ratings on the Checklist of Intervi-
ewer Behaviors showed that both experienced
(mean = 4.37, S.D. = 0.52, n = 9) and neophyte
(mean = 4.21, S.D. = 0.37, n = 11, P < 0.04) diag-
nostic interviewers performed within the ‘Good’
to ‘Excellent’ range.

4. Discussion

Interrater reliability on symptom ratings was
significantly higher during the training program
than during QA checks for both experienced and
neophyte interviewers. The decrease in reliability
at the first QA check may reflect the introductory
level of the initial training videos, which may have
artificially inflated interrater agreement on diag-
nosis and symptom ratings. Introductory SCID
training videos featured expert SCID interviewers
conducting assessments with patients who can
describe their symptoms well. In contrast, many
of the patients chosen for the SCID QA checks
were not preselected as good training cases.
Rather, they were acutely ill and difficult to inter-
view. In spite of the decrease in reliability, both
experienced and neophyte interviewers main-
tained good to excellent interrater reliability.
Several limitations to the current study should
be discussed. As mentioned, the χ² statistics re-

Table 1
A comparison of interrater reliability using kappa for interviewers who were experienced vs. neophytes in the use of structured assessments at the completion of SCID training and at the first Quality Assurance check

<table>
<thead>
<tr>
<th>Level of prior training</th>
<th>Experienced (n = 16)</th>
<th>Neophyte (n = 14)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (S.D.)</td>
<td>Range</td>
</tr>
<tr>
<td>Completion of SCID training</td>
<td>0.67 (0.06)</td>
<td>0.77–0.96</td>
</tr>
<tr>
<td>First Quality Assurance check</td>
<td>0.76 (0.13)</td>
<td>0.56–0.94</td>
</tr>
</tbody>
</table>
ported in this study were not used in the traditional fashion to represent agreement in sorting cases into diagnostic groups, e.g., schizophrenia vs. mood disorder. Furthermore, $k$ statistics of symptom agreement for co-rated interviews were derived from SCID assessments done when both judges were present. This is a less stringent test compared to the method of interviewing the patient separately by two judges, which typically yields lower interrater agreement. Additionally, many patients used in live interviews for training and QA had been pre-screened for having a major psychiatric disorder rather than selected at random from all possible diagnostic groups. Although these factors may have increased group interrater reliability, they probably did not systematically favor the experienced or the neophyte raters.

As shown in Table 2, there have been a variety of structured diagnostic assessment instruments available to national and international researchers. Few research studies describe the training program for diagnostic interviewers or the standards used to determine certification of interrater reliability. Even fewer studies describe quality assurance procedures designed to ensure that diagnostic interviewers have not drifted from original training standards. Standardization of training and Quality Assurance procedures across research centers may be advantageous because it would allow comparison of research findings.

Although the training and Quality Assurance program we described took place in the context of a large research center, the model can be applied in smaller research projects. For example, we have trained principal investigators at our Research Center who then became project diagnosticians. Each project diagnostican was responsible for training and maintaining the quality of diagnosis and symptom ratings for the project's diagnostic interviewers using the videotape training program provided by our Center. We check the project diagnostican's reliability on an annual basis by having him or her travel to our Center. We have also used this model in a multi-national study on optimal treatment of schizophrenia (PI: Ian Falloon, M.D.). For this study, diagnostic interviewers from various study sites (e.g., Italy, Sweden, Norway) were trained at a central European city. The interrater reliability of SCID interviewers was checked by circulating videotapes or through quality assurance checks.

<table>
<thead>
<tr>
<th>Structured diagnostic instruments</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Schedule for Clinical Assessment in Neuropsychiatry (SCAN)</td>
<td>World Health Organization, Geneva, Switzerland</td>
</tr>
<tr>
<td>(2) Schedule for Affective Disorder and Schizophrenia (SADS)</td>
<td>New York State Psychiatric Institute, New York, New York</td>
</tr>
<tr>
<td>(3) Comprehensive Assessment of Symptoms and History (CASH)</td>
<td>Nancy Andreasen, M.D., Department of Psychiatry, University of Iowa, College of Medicine, Iowa City, Iowa</td>
</tr>
<tr>
<td>(4) Composite International Diagnostic Interview-Revised (CIDI-R)</td>
<td>Washington University School of Medicine, St. Louis, Missouri</td>
</tr>
<tr>
<td>(5) Diagnostic Interview Schedule (DIS)</td>
<td>Washington University School of Medicine, St. Louis, Missouri</td>
</tr>
</tbody>
</table>
conducted at a central European city. Thus, these improvements in standardization of training and quality assurance procedures contributed to greater reliability and permit pooling of comparable data across multiple research sites.

Acknowledgements

The authors thank Daniel Gutkind, B.A., Elizabeth Gilbert, B.A., and Mark McGee, B.A., for their assistance. The data analyses were conducted by Sun Soon Hwang, M.S., M.P.H. of the Methodological and Statistical Support Unit of the UCLA Intervention Research Center for Major Mental Illnesses. We would also like to thank Miriam Gibbon, M.S.W., and the New York State Psychiatric Institute for expert consultation and the loan of SCID training videos to the Research Center’s Diagnosis and Psychopathology Unit. We are greatly indebted to Helena C. Kraemer, Ph.D., of Stanford University, for her valuable statistical consultation when the Diagnosis and Psychopathology Unit’s SCID training program was initiated.

Appendix A

Checklist of Interviewer Behaviors
UCLA Research Center for Severe Mental Illnesses
Diagnosis and Psychopathology Unit

Name of Trainee: ___________________________ Date: ___________________________
Name of Evaluator: ___________________________ Type of Interview: ___________________________

INSTRUCTIONS: This form lists 10 interviewer behaviors, each to be rated on 5-point scale from 'not at all' to 'always'. Circle the number which best describes the interviewer’s performance.

1 = not at all 2 = sometimes 3 = moderately 4 = most of the time 5 = always

RAPPORT:
1. Interviewer maintains good nonverbal contact with the patient (e.g., eye contact, leans forward) 1 2 3 4 5
2. Interviewer maintains nonjudgmental attitude towards the patient’s psychotic material. 1 2 3 4 5
3. Interviewer is sufficiently empathetic (e.g., makes emotionally congruent remarks). 1 2 3 4 5

PROBING:
4. Interviewer probes sufficiently after receiving negative responses to nonpsychotic items. 1 2 3 4 5
5. Interviewer follows up on all potentially positive responses to nonpsychotic items. 1 2 3 4 5
6. Interviewer probes sufficiently after receiving negative responses to psychotic items. 1 2 3 4 5
7. Interviewer follows up on all potentially positive responses to psychotic items. 1 2 3 4 5

TIME FRAME:
8. Interviewer informs patient of the time frame being used at appropriate junctures. 1 2 3 4 5

IMPAIRMENT & FREQUENCY
9. Interviewer clarifies symptom frequency,
INTerview Structure:
1. Interview process is too rapid so that necessary information is elicited.

GLOBAL Rating of INTERVIEWER Behavior*
1 = unacceptable 2 = poor 3 = fair 4 = good 5 = excellent
*only global ratings of 'good' or 'excellent' meet criteria for a 'pass' on Quality Assurance checks.

COMMENTS:

Appendix B
To: John Doe, M.D.
From: Joseph Ventura, Ph.D., Diagnosis and Psychopathology Unit
Date: February 25, 1997
Re: SCID Training

SCID Training Feedback Report
UCLA Research Center for Severe Mental Illnesses
Diagnosis and Psychopathology Unit

Standards for SCID Training Reliability:
1. Overall kappa coefficient of .75 or greater.
2. 90% agreement on the SCID diagnosis.
3. Demonstration of good interviewing skills as evaluated by the DPU Checklist of Interviewer Behaviors and individual report.

Definition of Kappa
1. Overall Kappa: Degree of agreement with consensus rating. Values greater than .75 are taken to represent excellent agreement. Values between .40 and .75 are taken to represent fair to good agreement. Values below .40 are taken to represent poor agreement.
2. Sensitivity: Agreement on symptoms present as determined by consensus.
3. Specificity: Agreement on symptoms absent as determined by consensus.

<table>
<thead>
<tr>
<th>Training Videotape</th>
<th>Co-Rated Line</th>
<th>Interviewer Conducted</th>
<th>Overall Kappa</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Diagnostic Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tape 1</td>
<td></td>
<td>95</td>
<td>1.00</td>
<td>.97</td>
<td>Correct</td>
<td></td>
</tr>
<tr>
<td>Tape 2</td>
<td></td>
<td>80</td>
<td>.85</td>
<td>.74</td>
<td>Correct</td>
<td></td>
</tr>
<tr>
<td>Tape 3</td>
<td></td>
<td>84</td>
<td>.84</td>
<td>.87</td>
<td>Correct</td>
<td></td>
</tr>
<tr>
<td>Tape 4</td>
<td></td>
<td>80</td>
<td>.80</td>
<td>.80</td>
<td>Correct</td>
<td></td>
</tr>
<tr>
<td>Tape 5</td>
<td></td>
<td>80</td>
<td>.80</td>
<td>.80</td>
<td>Correct</td>
<td></td>
</tr>
<tr>
<td>Tape 6</td>
<td></td>
<td>80</td>
<td>.80</td>
<td>.83</td>
<td>Correct</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B.S.</td>
<td>73</td>
<td>.73</td>
<td>.73</td>
<td>Correct</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E.B.</td>
<td>74</td>
<td>.74</td>
<td>.74</td>
<td>Correct</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>67</td>
<td>.67</td>
<td>.77</td>
<td>Correct</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M.S.</td>
<td>91</td>
<td>.87</td>
<td>.94</td>
<td>Correct</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F.J.</td>
<td>81</td>
<td>.81</td>
<td>.81</td>
<td>Correct</td>
<td></td>
</tr>
</tbody>
</table>

Mean Overall Kappa: .80 .77 .83
Comments on reliability: The reliability of symptom ratings and diagnosis was generally very good.

Comments on interview behaviors: Good job of establishing rapport. Conducted a very thorough, well structured and well paced interview. Good initial and follow-up probes and organization. Good use of a gentle confrontation technique to resolve discrepancies.

Suggestions/Recommendations: Be sure not to get into too much detail during the overview section of the SCID. Remember to probe for impairment in functioning. If you have any questions regarding this report or the use of the SCID, please contact the DPJU. Keep up the good work!

Date of Certification: February 12, 1997

cc: Andrew Shaner, M.D.
Robert F. Liberman, M.D.

References
Luria, R.E., McHugh, P.R., 1974. Reliability and clinical utility of the 'Wing' Present State Examination. Archives of General Psychiatry 30, 866–872.
Miklowitz, D.J., 1992. Longitudinal outcome and medication...
Noncompliance among manic patients with and without mood-incongruent psychotic features. Journal of Nervous and Mental Disease 180, 703-711.


