Assessment and Validation of Diagnostic Interviewing Skills for the Mental Health Professions

S. M. Bögels,[1,4] C. P. M. van der Vleuten,[2] G. Blok,[2] R. Kreutzkamp,[1] R. Melles,[3] and H. G. Schmidt[2]

A behavioral test was developed to assess the quality of diagnostic interviewing skills of (future) mental health professionals. Two aspects of diagnostic interviewing ability are distinguished: process skills, reflecting the interpersonal and communication skills; and content skills, referring to the information-gathering ability of the interviewer. It was found that diagnostic interviewing can be reliably measured with respect to interrater reliability. However, interviewer performance on one case proved to be a poor predictor of performance on other cases. It was concluded that a large number of cases is required to obtain reliable scores of general diagnostic interviewing ability. Validity was supported by the correlational analyses. Process skills were strongly related to patient satisfaction, whereas content skills were related to the amount of relevant information given by the patient and the accuracy of the diagnostic formulation and treatment plan.

---

[1]Department of Experimental Abnormal Psychology, University of Limburg, P.O. Box 616, 6200 MD Maastricht, The Netherlands.
[2]Department of Educational Development and Research, University of Limburg, Maastricht, The Netherlands.
[3]Support Center against Sexual Abuse, Municipal Health Service, Utrecht, The Netherlands.
[4]To whom correspondence should be addressed.

217

INTRODUCTION

The diagnostic interview in mental health care (mostly termed psychiatric interview) constitutes the main tool for assessment and, thereby, for treatment of mental health problems. Many practitioners have documented how to conduct such an interview (e.g., Hersen & Turner, 1985; Shea, 1988). Although it is widely recognized that the development of diagnostic interviewing skill should be one of the main objectives of mental health education (Borus & Yager, 1986; Maguire, 1982; Pollock, Shanley, & Byrne, 1985), there has been little systematic study on the assessment of diagnostic interviewing competence (McCready & Waring, 1986; Rutter & Cox, 1981; Shea & Mezzich, 1988).

In contrast with the lack of empirical studies on the assessment of diagnostic interviewing in the mental health field, a variety of studies has been directed at the measurement of clinical (interviewing) performance in the medical domain [see Van der Vleuten & Swanson (1990) for a review on this area]. Worldwide medical schools have started to use performance-based tests to measure clinical skills. Such tests usually consist of a number of " stations" at which students are required to perform a variety of tasks, including history-taking, physical examination and counseling, using simulated patients. One important finding, consistent in almost all research on clinical performance, is the low correlation between doctors' performance across cases. This has been termed *the case specificity of clinical competence* (Elstein, Shulman, & Sprafka, 1978). As a result, performance-based tests in medical schools usually consist of a large number of cases to judge students' skills reliably.

A few systematic studies were directed to the assessment of psychiatric interviewing. Mumford, Schlesinger, Cuerdon, and Scully (1987) investigated ratings of videotaped simulated patient interviews to evaluate a psychiatric clerkship. They identified two types of skills: process skills, necessary to establish a positive interpersonal relationship; and content skills, referring to an information-gathering ability. Ratings on process and content interviewing skills did not correlate or correlated only slightly with preceptor ratings of data gathering, interaction, and problem-solving skills. Furthermore, no relation was found between performance during the psychiatric interview and a multiple-choice end-of-clerkship psychiatric examination. Kraan and Crijnen (1987) evaluated general practice residents' psychiatric interviewing skills based on consults with two simulated patients. It appeared that the skills were very susceptible to differences between the two cases. A negative relationship was found between process skills and diagnostic accuracy. A positive relation was found between process skills and patient satisfaction, and between content skills and diagnostic accuracy.

Pollock *et al.* (1985) studied what outcome variables were associated with the performance of clinical clerks on a psychiatric interview. Contrary to their expectations, no relation was found between performance during the psychiatric interview and overall performance during the 1-month psychiatric rotation, a multiple-choice psychiatric examination, and patients' evaluation of the empathy of the interviewer.

In summary, diagnostic interviewing tests so far seem to lack validity. Low correlations were found between interviewing scores and relevant concurrent and outcome measures. Low correlations could be the result of a low true relationship between interviewing skills and other measures. On the other hand, unreliability of scores caused by case specificity of diagnostic interviewing ability could also account for the low correlations found. Not only interrater reliability, but also inconsistencies of interviewing behavior across different cases (i.e., mental disorders) may be an important source of variability. If this is true, validity studies need to control for case specificity.

The diagnostic interview has two goals (Reiser, 1984). The first goal is establishment and maintenance of a good working alliance, if only to promote compliance. The second goal is collection of information to arrive at a diagnostic formulation and treatment plan. Adequate construct validity of a test of diagnostic interviewing skills is indicated if this test measures interviewing skills that contribute to these goals. Based on the definition of process and content skills, process skills can be expected to contribute mainly to the goal of establishment of a good interviewer–patient relationship, whereas content skills can be expected to contribute mainly to the goal of diagnostic information collection and adequate diagnostic formulation.

The purpose of this study was to examine a test to assess diagnostic interviewing ability. We wished to explore interrater and intercase reliability, as well as the construct validity of the test.

METHOD

Subjects

Twenty graduated students of Mental Health Sciences (University of Limburg, The Netherlands) participated in the experiment. The mean age of the S's was 23.6 years (range, 20–42 years); 15 were female, and 5 were male. Mental Health Sciences is a 4-year full-time problem-based curriculum (Barrows, 1980), in which education is organized in 6-week units around specific themes, such as "mood," "anxiety," "addiction," and "illu-

sions and delusions." During their study students are extensively trained in diagnostic interviewing. They receive basic training in process skills, followed by training in interviewing and diagnosing patients with various mental disorders, such as anxiety disorders, mood disorders, and psychotic disorders. For an extensive description of the training program, see Bögels (1994).

Cases

Four standardized case histories were developed, derived from real patients. The cases concerned a young man with delusional thoughts and an alcohol problem (diagnosis Delusional Disorder, Persecutory Type, and Alcohol Abuse), a young woman with obsessions and panic attacks (diagnosis Panic Disorder with Agoraphobia and Obsessive-Compulsive Disorder), a middle-aged man with social anxiety and a history of avoidance (diagnosis Social Phobia, Generalized, and Avoidant Personality Disorder), and a widow with depressive complaints (diagnosis Major Depression, Single Episode). The cases were enacted by four experienced simulated patients, each patient portrayed one case. They were trained to simulate the case with high fidelity and constant over time. The validity of simulated patients has been demonstrated in research on medical students' and residents' skills (Norman, Tugwell, & Feightner, 1982; Sanson-Fisher & Poole, 1980).

Instruments

Diagnostic Interviewing Rating Scale in Mental Health Care (DIRSM). A rating form was developed to measure the quality of diagnostic interviewing skills (see Table II for a summary of the items). The DIRSM consisted of two subscales, called "process skills" and "content skills." Process scores reflected ratings on 11 aspects of interpersonal skills, such as asking questions, summarizing, thinking aloud, and giving information. Content scores on the DIRSM reflected ratings on 13 aspects of information-gathering technique, including chief complaint, history of the present complaint, and social and family history. Aspects were precisely defined in behavioral terms. Clear criteria for inadequate and adequate behavior were described. Criteria were so described as to be maximal independent of individual patient differences. Each aspect was scored on a 5-point scale, from 1 (insufficient) to 5 (advanced). Finally, raters were asked to provide global judgments of process and content quality of the interviewer, both on a 10-point scale.

Patient Satisfaction. The Patient Satisfaction with Communication Checklist (Crijnen & Kraan, 1987), originally developed to measure patient satisfaction related to physicians' communication, was modified for use in the psychiatric context. Simulated patients indicated how strongly they agreed or disagreed with 18 statements, on a 5-point scale from 1 (totally disagree) to 5 (totally agree). An example of a statement is, "The interviewer encouraged me to talk about what's bothering me."

Amount of Relevant Information. Quantity of relevant information given by the patient was measured as follows. Of each case a list of case-specific information items was constructed, important for diagnosis and treatment planning. This list consisted of about 30 items per case. Each item was rated in one of two categories, "mentioned" or "not mentioned," by the patient during the interview.

Diagnostic Accuracy. Diagnostic accuracy was measured by means of a semistructured questionnaire (Problem Solving Questionnaire; Kraan & Crijnen, 1987). The interviewer responded to six open-ended questions about diagnosis, differential diagnosis, explanatory hypotheses, and plans for further management.

Knowledge of the DSM-III-R. A test, consisting of six open-ended questions, was developed to measure knowledge of mental disorders, as classified by DSM-III-R. One question concerned knowledge of the DSM-III classification system, the other questions concerned general diagnostic knowledge of the disorders that were impersonated by the simulated patients.

Procedure

Subjects were instructed to conduct a diagnostic interview with the four simulated patients, each for a period of 30 min, to make a diagnosis and management plan The interviews were videotaped. Immediately after each interview *S*'s were asked to complete the Problem Solving Questionnaire concerning the patient they had just seen. The patients simultaneously completed the patient satisfaction checklist. After completion of the four interviews, *S*'s were tested on their DSM-III-R knowledge. The whole test session took 3.5 hr, and the *S*'s interviewed the patients in varying order.

Scoring

Four expert clinical psychologists were asked to rate the videotaped skills of the interviewers. All raters had extensive experience in interviewing and diagnosing patients with mental disorders; their mean professional ex-

perience was 4 years (range, 2 to 6). They were also familiar with teaching these skills to students. Raters were trained in using the DIRSM to a satisfactory interrater reliability level. All videotapes were then independently scored by two raters. The amount of relevant information was rated by two clinical psychology students. Preliminary research showed that one rater provided sufficient reliable ratings [interrater reliability between the two raters, based on their ratings of nine videotapes of different *S*'s and different patients, was .97 (intraclass correlation)]. Each videotape was, therefore, scored by only one rater. To judge the answers of the *S*'s on the Problem Solving Questionnaire, a scoring key was developed by eight psychiatric experts, consisting of accurate, semiaccurate and wrong answers. Two psychologists (other than those who scored the videotapes) independently rated the completed Problem Solving Questionnaires of the *S*'s, based on the scoring key. Yet another psychologist rated the knowledge of the DSM-III-R, by means of a predetermined scoring key of accurate, semiaccurate, and wrong answers.

Total scores across items were calculated for process and content skills of the DIRSM, diagnostic accuracy, patient satisfaction, and knowledge of DSM-III. Because the variable amount of relevant information had different total scores in different cases, a proportional score was calculated.

Statistical Analysis

Reliability was analyzed by using generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson, Webb, & Rowley, 1989). Generalizability theory provides a framework for examining the dependability of behavioral measurements. In contrast to classical test theory, which treats only one source error source at a time (e.g., interrater reliability or test–retest reliability), generalizability theory recognizes that there may be multiple sources of error variance which define how accurately observed scores permit us to generalize about interviewers' behavior in a universe of situations.

By means of generalizability analysis (Brennan, 1983) the relative magnitude of variance caused by interviewers, cases, raters, and their interactions was estimated. A raters-nested-within-cases-by-persons design ("j:ixp" design in generalizability terminology) was used to estimate variance components for process and content skills and diagnostic accuracy, since each case was rated by a fixed pair of raters. For the variables amount of relevant information and patient satisfaction variance components were estimated according to a simple person-by-case design ("pxi"), as, for these instruments, each case was rated by only one rater. For all instruments

generalizability coefficients were calculated with projections to different numbers of cases and—where appropriate—to different numbers of raters. These coefficients express the reproducibility (reliability) of the observed scores generalized to the universe of raters and cases, from which the sample of raters and cases participating in this study is taken.

Interrater reliabilities of process skills, content skills, and diagnostic accuracy were estimated by calculating an intraclass correlation on all available scores of rater pairs ("i:p" design in generalizability terminology). This coefficient provides an index of interrater reliability that includes differences in rater stringency as measurement error. Interrater reliability per item of the DIRSM was calculated by means of a weighted Cohen's κ. A κ of .40 is considered moderate, a κ of .60 high/substantial, and a κ of .80 excellent (Pritchard, 1981).

Correlational analysis was used for validity calculations. In order to establish to what extent process and content skills contribute to the goals of establishing rapport with the patient, collection of relevant information, and adequate diagnostic formulation, partial correlations were calculated, controlling for content and process skills, respectively.

RESULTS

In Table I, descriptive statistics of all variables are presented. The cases were judged as very realistic by the S's (mean fidelity score was 4.2, on a scale from 1 = very unrealistic to 5 = very realistic), supporting the fidelity of the role-plays.

Interrater reliability of total scores was good to excellent; the intraclass correlation was .79 for process skills, .72 for content skills, and .84 for diagnostic accuracy. The interrater reliabilities of item scores of the DIRSM ranged from fair to excellent (see Table II).

Table I. Descriptive Statistics of All Variables (Averaged Across Raters and Across Cases)

	Mean	SD	Scale min.	Scale max.
Process skills	43.9	5.6	11	60
Content skills	39.4	3.6	14	75
Diagnostic accuracy	7.1	1.3	0	12
Relevant information (%)	38.5	4.6	0	100
Patient satisfaction	65.8	6.0	18	90
DSM-III knowledge	6.0	2.1	0	12

Table II. Interrater Reliability (Weighted Cohen's κ) per Item of the Diagnostic Interviewing Rating Scale in Mental Health Care (DIRSM)

Process skills	κ
1. Attending behavior	.56
2. Open-ended questions	.54
3. Close-ended questions	.40
4. Concreteness	.48
5. Reflection of feelings	.41
6. Paraphrasing of content	.59
7. Summarizing	.78
8. Giving information	.61
9. Asking for feedback	.74
10. Thinking aloud	.47
11. Structuring	.48

Content skills	κ
12. Clarification of presenting problems	.53
13. Exploration of other problems	.57
14. Reason for encounter, help request	.71
15. History of main problem	.57
16. Analyzing causal factors	.69
17. Exploration of duration and course	.64
18. Exploration of consequences for daily life	.69
19. History of prior treatments	.75
20. Psychiatric examination	.89
21. Personal history	.85
22. Assessment of premorbid personality	.75
23. Exploration of social relations	.47
24. Occupational, financial, and housing conditions	.66

Table III contains the variance components of process and content skills. The person component represents the desirable true score variance. It indicates how much subjects vary in ability. The case component reflects absolute differences between cases and is an index of case difficulty. The rater-within-case component reflects systematic differences in rater stringency. The person-by-case component represents inconsistencies in the rank-ordering of persons across cases, the so-called "case specificity problem." The last component refers to the (residual) measurement error. For both process and content skills, the person-by-case component is the largest source of variance. The "true" person variance is also large for process skills, but is much lower for content skills. Content skills appear quite sensitive to absolute differences between cases, whereas case variance accounts for only 8% of the variance of process skills. The general error term is a substantial source of variance in both process and content skills. Variance components of the other variables are also given.

From the variance components in Table III, generalizability coefficients were calculated. In Table IV generalizability coefficients are presented as a function of the number of cases and raters. Based on four cases and two raters, the sample size of the present experiment, generalizability is acceptable for process skills. Content skills require a larger number of cases (at least 10) to reach acceptable generalizability of scores, mainly as a result of the lower "true" person variance. Generalizability coefficients of the other variables are also shown. The reliability (internal consistency) of knowledge of DSM-III-R is low (α = .38), the internal consistency of the patient satisfaction checklist is high (α = .90).

Table V contains the correlations between the different variables. It demonstrates that process skills correlate strongly with patient satisfaction. Content skills correlate substantially with the amount of relevant information given by the patient and moderately with diagnostic accuracy. Contrary to expectations, content skills do not correlate with knowledge of DSM-III-R. As different variables were rated by different judges (except for process and content skills), the relationship between the different variables is not due to the method of data collection (or halo effect). Moreover, taken into account the generally low correlations between behavioral measures and questionnaires, observed correlations are fairly high.

Partial correlations were calculated, controlling for process skills or content skills. After partialling out content skills, process skills correlated .32 with patient satisfaction, .08 with relevant information gathered, -.29 with diagnostic accuracy, and -.25 with DSM-III knowledge. Content skills, partialling out process skills, correlated .34 with diagnostic accuracy, .24 with relevant information gathered, .17 with patient satisfaction, and -.03 with DSM-III knowledge.

DISCUSSION

Results of this study show that diagnostic interviewing skills can be reliably measured with respect to interrater reliability. However, differences in rank-ordering of interviewers across cases, generally referred to as *the content specificity problem*, threatened overall reliability of the assessment of diagnostic interviewing performance. This is found in virtually all studies to assess clinical competence (Van der Vleuten & Swanson, 1990). What causes these variability of performance across cases? With regard to content skills, differences in knowledge and experience between interviewers concerning specific mental disorders can explain this phenomenon. Interestingly, however, process skills suffered even more from person-by-case interaction. Differences in proficiency to handle the interviewing process

Table III. Absolute and Proportional Variance Components of Process and Content Skills, Diagnostic Accuracy, Relevant Information, and Patient Satisfaction (SE in Parentheses)

Source	Process skills	Content skills	Diagnostic accuracy	Relevant information	Satisfaction
P (persons)	19.8 (8.3) 34%	7.1 (3.9) 15%	1.6 (0.7) 27%	.03 (.07) 3%	23.3 (11.5) 26%
C (cases)	4.6 (4.1) 8%	13.6 (9.8) 28%	1.8 (1.2) 31%	.09 (.08) 10%	13.6 (10.3) 15%
R:C (rater-within-case)	0.9 (0.9) 2%	1.9 (1.4) 4%	0.1 (0.4) 2%		
PC (person-by-case)	21.6 (5.1) 37%	14.5 (3.9) 30%	1.5 (0.4) 25%	.74 (.14) 86%	52.3 (9.6) 59%
PR:C (error)	11.5 (1.8) 20%	11.8 (1.9) 24%	0.9 (0.2) 15%		

Table IV. Reproducibility of Process and Content Skills, Diagnostic Accuracy, Relevant Information, and Patient Satisfaction as a Function of the Number of Cases and the Number of Raters

| N of cases | Process skills | | Content skills | | Diagnostic accuracy | | Relevant information[a] | Patient Satisfaction[a] |
	One rater	Two raters	One rater	Two raters	One rater	Two raters	One rater	One rater
1	.38	.42	.21	.26	.40	.45	.03	.31
2	.55	.59	.35	.41	.57	.62	.07	.47
4	.71	.74	.52	.58	.73	.77	.14	.64
10	.86	.88	.73	.78	.87	.89	.29	.82

[a]For a single rating only, since only single ratings were available.

Table V. Observed Correlations Between Process Skills, Content Skills, Relevant Information, Diagnostic Accuracy, Patient Satisfaction, and DSM-III-R Knowledge

	Content skills	Relevant information	Diagnostic accuracy	Patient Satisfaction	DSM-III-R Knowledge
Process skills	.71***	.37	.07	.68***	.00
Content skills		.57**	.45*	.45*	.02
Relevant information			.40	.42	−.02
Diagnostic accuracy				.40	.13
Patient satisfaction					.16

*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$

with patients with different communication styles, for example, to structure a very talkative patient or to stimulate a withdrawn patient, may account for the variability of process skills across cases. Moreover, content skills appeared quite sensitive to absolute differences in cases. This may be the result of variations between cases in terms of differential-diagnostic difficulty.

These findings have important implications for the measurement of diagnostic interviewing. The variability of diagnostic interviewing performance across cases/content has to be taken into account in studies on diagnostic interviewing. For examination purposes, a large number of cases is required to obtain reliable scores of general diagnostic interviewing ability. Examination of trainees may lead to unreliable inferences, if they are based on single cases. Finally, results suggest that it is generally unnecessary to use more than one rater per interview, at least if many cases are used and different raters are used per case.

The validity of the DIRSM was supported by the correlational analyses. Process skills, supposed to be necessary to establish a good relationship with the patient, were strongly related to the quality of the interviewer-patient relationship as experienced by the patient. Content skills, supposed to reflect the information-gathering ability of the interviewer, were substantially connected with the amount of relevant information given by the patient. Content skills also moderately correlated with accuracy of the diagnostic formulation and treatment plan, whereas process skills did not. Partial correlations (controlling for the correlation between process and content skills) confirmed that process skills are related to patient satisfaction, whereas content skills are related to diagnostic accuracy and relevant information gathered. Clearly, both skill aspects are, in a different way, important to meet the purposes of a diagnostic interview: building up a

relationship, collecting the necessary diagnostic information, and making correct diagnostic inferences.

Although knowledge of classification of mental disorders is assumed to be a precondition for adequately interviewing a psychiatric patient, with respect to content issues, knowledge of DSM-III-R did not correlate with content skills. A first explanation for this unexpected finding concerns the way DSM-III-R knowledge was assessed. We used a test, consisting of six open-ended questions. The low reliability of the test may have influenced the results. Another possible explanation for the lack of correlation between content skills and DSM-III-R knowledge is that we measured knowledge of facts, rather than *application* of this knowledge in real situations. Research by Glaser and colleagues (Glaser, 1984) suggests that the knowledge of experts includes knowledge about the application of what they know. Novices, on the other hand, may lack this "application knowledge." It can be hypothesized that *application* knowledge of DSM-III-R will correlate with content skills.

The DIRSM seems appropriate to measure diagnostic interviewing ability. Future research on this test should concentrate on different areas. First, by means of latent structural analysis, it has to be investigated whether the two aspects of diagnostic interviewing (process and content) are indeed distinguishable factors. Second, ways to obtain reproducible results within a limited amount of testing time need to be explored, for example, by shortening interviewing time per case. Recent research by Bögels, Van Mourik, and Van der Vleuten (1995) showed that shorter interviews (5 or 10 min) produced higher "true" person variance and, as a result, higher generalizability coefficients than longer interviews (20 min). Finally, another validity question seems relevant: Can the test discriminate between novices and experts in this field? With respect to the teaching of diagnostic interviewing, more research is needed to evaluate diagnostic interview training. Studies by Gask, McGrath, Goldberg, and Millar (1988), Goldberg, Steele, Smith, and Spivey (1980), and Maguire, Fairbairn, and Fletcher (1986) showed that medical students and general practitioners can be effectively trained in psychiatric interviewing. The effectiveness of diagnostic interview training programs for the mental health professions still needs to be established.

REFERENCES

Barrows, H. S. (1980). *Problem-based learning: An approach to medical education.* New York: Springer.

Bögels, S. M. (1994). A structured training approach to teaching diagnostic interviewing. *Teaching of Psychology, 21,* 144-150.

Bögels, S. M., Van Mourik, T. G. C., & Van der Vleuten, C. P. M. (1995). Authentic assessment of interviewing and counselling skills: Effect of testing time per station on reproducibility and validity. *Teaching and Learning in Medicine, 7*, 155-162.

Borus, J. F., & Yager, J. (1986). Ongoing evaluation in psychiatry: The first step toward quality. *American Journal of Psychiatry, 143*, 1415-1419.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: The American College Testing Program.

Crijnen, A. A. M., & Kraan, H. F. (1987). Measuring patient satisfaction with the communication. In H. F. Kraan & A. A. M. Crijnen (Eds.), *The Maastricht History-taking and Advice Checklist: Studies of instrumental utility,* Thesis, pp. 145-171. Amsterdam: Velden van den Hazelaar.

Cronbach, L. J., Gleser, C. G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley & Sons.

Elstein, A., Shulman, L., & Sprafka, S. (1978). *Medical problem solving.* Cambridge, MA: Harvard University Press.

Gask, L., Goldberg, D., Lesser, A. L., & Millar, T. (1988). Improving the psychiatric skills of the general practice trainee: An evaluation of a group training course. *Medical Education, 22*, 132-138.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychology, 39*, 93-104.

Goldberg, D. P., Steele, J. J., Smith, C., & Spivey, L. (1980). Training family doctors to recognize psychiatric illness with increased accuracy. *Lancet, 2*, 521-523.

Hersen, M., & Turner, S. M. (1985). *Diagnostic interviewing.* New York, Plenum Press.

Kraan, H. F., & Crijnen, A. A. M. (1987). *The Maastricht History-taking and Advice Checklist: Studies of instrumental utility,* Thesis. Amsterdam: Velden van den Hazelaar.

Maguire, P., Fairbairn, S., & Fletcher, C. (1986). Consultation skills of young doctors: I. Benefits of feedback training in interviewing as students persists. *Medical Education, 292*, 1573-1576.

Maguire, P. (1982). Psychiatrists also need interview training (Comments). *British Journal of Psychiatry, 141*, 423-424.

McCready, J. R., & Waring, E. M. (1986). Interviewing skills in relation to psychiatric residency. *Canadian Journal of Psychiatry, 31*, 317-322.

Mumford, E., Schlesinger, H., Cuerdon, T., & Scully, J. (1987). Ratings of videotaped simulated patients and four other methods of evaluating psychiatric clerkship. *American Journal of Psychiatry, 3*, 316-322.

Norman, G. R., Tugwell, P., & Feightner, J. W. (1982). A comparison of resident performance on real and simulated patients. *Journal of Medical Education, 57*, 708-715.

Pollock, D. C., Shanley, D. F., & Byrne, P. N. (1985). Psychiatric interviewing and clinical skills. *Canadian Journal of Psychiatry, 30*, 64-68.

Pritchard, M. J. (1981). Temporal reliability of a questionnaire measuring psychological responses to illness. *Journal of Psychosomatic Research, 25*, 63-66.

Reiser, D. E. (1984). The psychiatric interview. In H. G. Goldman (Ed.), *Review of general psychiatry,* pp. 197-205. Los Altos, CA: Lange.

Rutter, M., & Cox, A. (1981). Psychiatric interviewing techniques: I. Methods and measures. *British Journal of Psychiatry, 138*, 273-282.

Sanson-Fisher, R. W., & Poole, A. D. (1980). Simulated patient and the assessment of medical students' interpersonal skills. *Medical Education 14*, 249-253.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 6*, 922-932.

Shea, S. C. (1988). *Psychiatric interviewing: The·art of understanding.* Philadelphia: Saunders.

Shea, S. C., & Mezzich, J. E. (1988). Contemporary psychiatric interviewing: New directions for training. *Psychiatry, 15*, 385-397.

Van der Vleuten, C. P. M., & Swanson D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*, 58-76.