A STUDY OF THE RELIABILITY AND VALIDITY OF THE GLOBAL ASSESSMENT OF FUNCTIONING SCALE

By

JACK PHILLIP SCHAEFER

Bachelor of Science
Oklahoma State University
Stillwater, Oklahoma
1969

Master of Science
Oklahoma State University
Stillwater, Oklahoma
1980

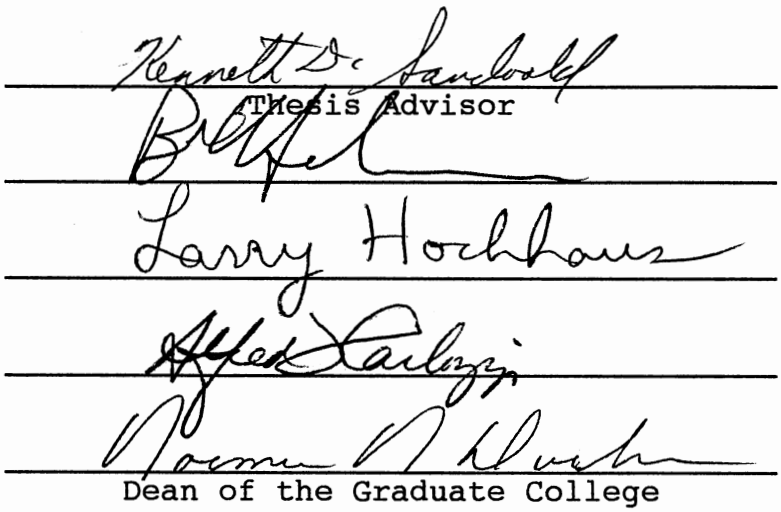
Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the degree of
DOCTOR OF PHILOSOPHY
July, 1991

A STUDY OF THE RELIABILITY AND VALIDITY

OF THE GLOBAL ASSESSMENT OF

FUNCTIONING SCALE

Thesis Approved:



Thesis Advisor

Dean of the Graduate College

ACKNOWLEDGEMENTS

My sincere thanks go out to my colleagues at the Edwin Fair Community Mental Health Center for their participation in this study. Without their help, this project could not have been completed. I am very grateful to Dr. Ken Sanvold for his advice and support in guiding the completion of this study. I am also indebted to the other committee members, Dr. Larry Hochhaus, Dr. Bob Helm, and Dr. Al Carlozzi. Each of them contributed significantly. I am thankful that I met Dr. Lois Pokorny of the Oklahoma Mental Health Research Institute, who provided me with the idea for the study and vital assistance in getting started. Finally, I would like to thank my family, friends and professional colleagues who have encouraged me and have never given up on me as I have struggled to complete my dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Reflecting the growing emphasis on mental health program evaluation and the need for a standard measure of the client's level of psychiatric functioning, the American Psychiatric Association (1987) published the new Global Assessment of Functioning Scale (GAF Scale) in the Diagnostic and Statistical Manual of Mental Disorders (Third Edition - Revised), DSM-III-R. The GAF Scale is used in the DSM-III-R five axis diagnosis process.

The GAF Scale is a revision of the Global Assessment Scale (GAS), published by Endicott, Spitzer, Fleiss and Cohen in 1976. Dr. R. L. Spitzer, who helped develop the GAS and GAF Scale, is also known as the chairperson of the work groups that developed the DSM-III and DSM-III-R.

The GAF Scale is used by clinicians to rate a client's overall functioning level or degree of mental health-illness. The rating is accomplished by assigning the client a number between 1 and 90. A rating of 1 represents the lowest functioning and most severe symptoms, while 90 represents the most superior functioning and lack of symptoms. A rating of 0 is made when there is inadequate information. The GAF Scale is reproduced in Appendix A.

On Axis V of the DSM-III-R, diagnosis ratings on the

GAF Scale are made for two time periods - current (the level of functioning at the time of the evaluation) and the past year (the highest level of functioning for at least a few months during the past year). A GAF Scale rating at the time of discharge is often a standard procedure. To aid the clinician in making the GAF Scale rating, behavioral examples and symptom descriptions are provided for each ten point interval in the scale (See Appendix A).

The inclusion of the GAF Scale in such an important context as the universally used DSM-III-R diagnosis process reflects the widespread acceptance of its predecessor, the GAS (Appendix B). Dekker (1983), in an exhaustive review of the literature, determined that the GAS had found the most widespread acceptance of all psychotherapy outcome measures. He reported that by 1983, the GAS had been used in over 200 published research studies, and had been adopted by five states as a standard level of functioning measure in their mental health agency management information systems.

Important decisions have been based in part on information gained from GAS scores. The comparison of admission GAS scores with discharge GAS scores provides a quick and easy outcome measure for mental health centers. Information like this has been used in outcome research, treatment evaluation, program planning, budgeting, policy development, and cost effectiveness studies (Ciarlo, 1982; Newman, 1980). Two NIMH publications, issued shortly after the introduction of the GAS and cited by Dekker in 1983, strongly recommended its use in mental health program

evaluation (Hagedorn, Beck, Neubert & Werlin, 1976; Hargreaves, McIntyre, Attkinson, & Siegal, 1977).

Among the uses of the GAF in research, according to Dekker (1983), have been the following: (1) to evaluate the effectiveness of new treatments, (2) to evaluate change due to psychotherapy, (3) to evaluate change due to psychoactive medication, (4) as a criterion measure in validity studies for newly developed scales, (5) to compare the level of disturbance between experimental and control groups, and (6) as a screening instrument to select subjects.

The problem addressed in this study relates to the fact that the GAS has grown in popularity and importance in spite of its questionable validity, and the GAF Scale, which is very similar to the GAS, has not been adequately evaluated for reliability or validity. Dekker's 1983 review found that while the reliability of the GAS was adequate, validity studies were sparse, tended to have methodological problems, and reported only low-to-moderate concurrent validity coefficients. In the years since 1983, only two studies (Holcomb & Otto, 1988; Sohlberg, 1989) have investigated the validity of the GAS, and they provide conflicting results.

With regard to the newer GAF Scale, a review of the literature since its introduction in 1987 revealed no published research studies which have used it or studied its reliability or validity. It appears that the older GAS is still being used by researchers. However, it is reasonable to assume that the supposedly improved GAF Scale will supersede the GAS as a research instrument eventually. Its

inclusion in the DSM-III-R diagnosis system will virtually insure its adoption as an outcome measure in mental health evaluation systems. The need for reliability and validity data on the GAF Scale is obvious.

Purpose of the Research

The present study was designed to empirically test the interrater reliability, concurrent validity, and construct validity of the GAF Scale. Interrater reliability of the scale was examined by comparing independent GAF ratings of psychiatric outpatients by two mental health professionals currently providing ongoing treatment to those outpatients. The concurrent validity of the scale was examined by comparing GAF ratings of psychiatric outpatients with independent ratings of the same outpatients on the Brief Psychiatric Rating Scale (BPRS) (Overall & Gorham, 1962). Supported in the literature as a reliable and valid instrument, the BPRS has been extensively used over the past 28 years in a wide range of research applications.

Construct validity of the GAF Scale was examined by comparing the mean GAF score of a group of clients who have severe diagnoses with the mean GAF score of a group with mild to moderate diagnoses. Also, mean GAF scores of chronically mentally ill (CMI) clients were compared with mean GAF scores of non-chronically mentally ill clients.

Based on previous findings, it was expected that the interrater reliability coefficients for the GAF Scale would be in the acceptable range (.70 and above), and that

concurrent validity coefficients would range from low to moderate. It was also expected that the severe diagnosis group and the CMI group would have significantly lower GAF scores than the comparison groups.

CHAPTER II

REVIEW OF THE LITERATURE

Forerunners of the GAF Scale

Members of the Psychotherapy Research Project of the Menninger Foundation were the first to develop a 100 point global mental health rating scale. Called the Health-Sickness Rating Scale (HSRS) (Luborsky, 1962) it was an attempt to fill a need for "a simple survey instrument to record shorthand judgements of a patient's status - one that would permit recording of changes over time in a single case and easy comparison of one case with another" (Luborsky, 1962, p. 408). They hit upon the idea for a scale that assigned an absolute numerical rating of degree of mental health.

Along its scoring continuum from 0 (any condition which, if unattended, would quickly result in the patient's death) to 100 (an ideal state of complete functioning), there were 8 unequally distributed anchor points with behavioral descriptions and diagnostic examples. Thirty four sample case vignettes were also provided which were tied to ratings at least every five points up the scale. Using the scale examples given, it appeared that individuals diagnosed as schizophrenic could not be rated above 50 and

"closed ward" chronic schizophrenics were to be rated about 10.

Eighteen studies of the reliability and validity of the HSRS were summarized by Luborsky and Bachrach (1974). Interrater reliability between independent judges ranged from .65 to .94. It was concluded that the HSRS was a reasonably reliable instrument. Concurrent validity coefficients were reported ranging from .32 to .84, but the higher coefficients were obtained by comparing ratings made by one judge using the HSRS and another instrument to rate the same person. This made it impossible to determine the extent to which the correlation was inflated due to contamination. In other words, when a rater uses two different measuring scales to rate the same subject, the rater may have a tendency to link the two ratings or make them consistant with one another in reflecting the rater's underlying opinion about the subject characteristics in question. This may cause a spuriously high correlation between the two ratings. When independent ratings were obtained, the HSRS correlated lower with other rating scales, from .32 to .55. The HSRS became very popular and was frequently used in research studies during the 1960's and 1970's. However, in the 1980's it was superseded by a newer and supposedly improved scale, the GAS (Dekker, 1983).

Reputed to be an advancement in the development of these global rating scales, the Global Assessment Scale (GAS) was published in 1976 (Endicott et al., 1976). Like the HSRS, the GAS was a 100-point scale calling for a single

numerical rating of a subject's psychological or psychiatric health. The scale values ranged from 1, the sickest individual, to 100, the healthiest. The GAS differed from the HSRS in that it had 10 anchor points distributed at equal intervals and the anchor points were defined by only behavioral or symptom descriptions, not diagnostic labels. The authors explained that tying diagnostic labels to anchor points increased the difficulty of the rating task. For instance, they found that often schizophrenics in remission satisfied the behavioral criteria for HSRS scores above 50, yet the diagnostic examples implied that they should not be rated above 50. They decided to "eliminate all diagnostic constraints and instead provide for each interval a number of specific behavioral descriptions exemplifying that range" (Endicott et al., 1976, p. 767). The GAS also did away with the sample case vignettes because they added needlessly to the complexity of the task and because in actual practice the vignettes were rarely used.

According to Dekker (1983), the interrater reliability of the GAS was evaluated in 13 published studies and 4 unpublished studies prior to 1983. These studies yielded 31 interrater reliability coefficients ranging from a low of .33 to a high of .98, with a median interrater reliability of .80, an acceptable level. Twenty six of these reliability coefficients were obtained using trained raters and five were obtained using untrained raters. The average correlation coefficient of trained raters was .78 and the average of untrained raters was .59. The highest

correlations were obtained using highly educated raters and the lowest were obtained using the least educated raters.

A review of research studies using the GAS since 1983 yielded only one which specifically examined the interrater reliability of the scale. Sohlberg (1989) reported tests of interrater reliability for the GAS. Adult patients who had originally presented with eating disorders were seen at follow-up two and five years after first presentation. Control judges were blind to the identity and GAS of the patients as scored by other judges. All judges were reportedly trained in the use of the GAS. Coefficients ranged from .83 to .92, which are consistent with reliability data obtained in earlier studies.

While reliability must be established first in evaluating the psychometric properties of an instrument, validity is also a crucial factor. If the instrument does not measure what it is supposed to measure, then the conclusions based upon its utilization must be questioned.

The content validity of the GAS was questioned by Ciarlo, Edwards, Kiresuk, Newman & Brown (1981) who pointed out that the scale's behavioral descriptions in the upper half anchor points are less well elaborated than in the lower half, thus making the scale more sensitive in rating psychotic disorders. An analysis of the GAS shows that there are approximately twice as many behavioral descriptions and examples in the range from 0 to 50 than in the range from 51 to 100.

Few studies have examined the concurrent validity of

the GAS, and there is not strong support for the scale's validity among them. The most extensive evaluation was presented by Endicott et al. (1976) in their original publication of the scale. The GAS was administered to psychiatric inpatients by therapists and research interviewers. Independently administered to the same inpatients were the Mental Status Examinations Record (MSER), the Psychiatric Status Schedule (PSS), and the Family Evaluation Form (FEF). All these instruments were administered at admission and again six months later when many subjects were no longer patients.

The MSER is a seven-point global rating scale for overall severity of psychiatric illness ranging from "not ill at all" to "among the most extremely ill". The PSS is a structured interview schedule providing scores on five aspects of client psychopathology. The FEF provides a total score of patient psychopathology based on structured interviews of family members of the patients. Since more severe pathology is indicated by lower scores on the GAS and higher scores on other instruments, expected correlations were in the negative direction.

The interrater correlations were low when taken at the time of admission. The highest admission correlation of -.44 was with the MSER, but as pointed out by Dekker (1983), the similarity between the GAS and MSER as global rating scales suggests that the correlation between the two may be closer to a reliability coefficient than a validity coefficient. The mean admission correlation between the GAS

and FEF was -.22, and the mean correlation between the GAS and five PSS scales was also -.22.

The interrater correlations between scores taken at a six month followup were stronger, presumably because the raters knew much more about the patients upon follow-up than they knew after the admission intake interview. The GAS scores correlated -.62 with the MSER, a mean of -.48 with the FEF, and a mean of -.37 with the PSS.

In summarizing the validity data, the authors viewed the correlations obtained both at admission and six months later as "moderate" and further stated that "since the GAS was developed to improve on these and other global procedures, very high correlations are neither desired nor expected" (Endicott et al., 1976, p. 777).

In his 1983 review, Dekker mentioned five other published studies which potentially shed light on the validity of the GAS, although none were designed specifically for that purpose. All used the GAS to establish validity for newly developed scales. Although the correlations reported were mostly in an acceptable moderate range, four of the studies, Battista (1982), Lefkovit, Morrison, & Davis (1982), Sorenson, Hargreaves & Friedlander (1982), and Stone (1979), used the same rater for both the GAS and the new scale they were trying to validate. This made it impossible to determine the degree to which correlations were inflated. In the one study which did use independent raters, Fawcett, Clark, Schneftner, & Gibbons (1983) obtained a validity coefficient of -.12 when GAS

scores were correlated with scores on the new Pleasure Scale, which was used to measure the intensity of pleasurable responses to normally enjoyable situations.

In Dekker's 1983 study, GAS and MMPI data were collected from adult outpatients. GAS ratings by therapists were correlated with four overall severity indexes from the MMPI, as well as with the individual MMPI scales. Correlations with the overall indexes ranged from .29 to .36. Correlations with the clinical scales were all under .36 and indicated that the GAS scores are more related to psychotic symptoms than to neurotic ones.

More recently, Zheng, Zhao, Phillips, Liu, Cai, Sun, & Huang (1988) examined the concurrent validity of the Chinese Hamilton Depression Rating Scale (CHDS) by correlating CHDS scores with GAS scores. A strong negative correlation (-.83) was found. The ratings on the two scales were made by independent, trained raters.

Since 1983, only two published studies could be found which directly examined the concurrent validity of the GAS. Holcomb and Otto (1988) correlated GAS ratings with several measures of mental status and psychopathology. Clinicians at rural counseling centers administered a comprehensive intake battery which included the GAS, the Missouri Department of Mental Health Adult Mental Status Examination, with 118 items under 10 major headings. A DSM-III diagnosis and a list of presenting problems were determined for each client. Correlations between the GAS and the mental status examination heading scores ranged from the low +.003

(antisocial) to moderate -.35 (disorientation). There was a coefficient of only -.03 between the GAS scores and the presence of hallucinations. The only diagnostic category significantly related to the GAS scores was Conditions Not Attributed to a Mental Disorder. A significant but low correlation (-.27) was found between GAS scores and the number of presenting problems. These low to moderate correlations were obtained using a method by which the same clinician rated clients on all measures, again constituting possible contamination and inflated coefficients.

Sohlberg (1989) challenged the results of Holcomb & Otto (1988) and conducted an investigation of the concurrent validity of the GAS. GAS ratings correlated strongly with independent measures of psychopathology administered to people at one, two, and five year follow up visits. The GAS scores correlated with the Beck Depression Inventory (-.87), the Symptom Checklist-90 (-.69), the Washington University Sentence Completion Test of ego development (.59), the Eating Disorder Inventory (-.55), the Eating Attitudes Test (-.72) and the presence of a DSM-III eating disorder diagnosis (-.76).

In comparing these two studies it must be noted that Holcomb and Otto (1988) administered all instruments at the time of intake, while Sohlberg (1989) concentrated on measures taken at followup visits. As noted above, past studies indicate that the GAS may produce higher validity scores when it is administered further into treatment.

Another recent study which concluded the GAS is a valid

measure of psychiatric impairment was conducted by Vaillant and Schnurr (1988). The GAS and five other models of psychiatric impairment were correlated with independently obtained measures of psychosocial impairment such as subjects' reported subjective distress, alcohol abuse, maximum earned income, and subjective adjustment to aging. With correlation coefficients ranging from -.66 to -.11 the authors conclude that the GAS is valid and is an effective predictor of adult adjustment in late midlife.

In addition to the evidence concerning the concurrent validity of the GAS cited above, there is also evidence concerning its construct validity. The construct validity of a rating scale such as the GAS can be examined by several methods. First, one may look to see if some form of psychiatric or psychological treatment, which is expected to produce change, is accompanied by change in GAS scores (Dekker, 1983).

In his 1983 review, Dekker summarizes five studies of hospitalized psychiatric patients, primarily schizophrenics, which reported a significant rise in GAS scores from admission to discharge and follow-up (Larkin, 1979; Herz, Endicott, & Spitzer, 1975; Goldstein, 1980; Curran, Miller, Monti, Zwick, & Stout, 1980; Gudeman, Dickey, Rood, Hellman, & Grinspoon, 1981). Dekker (1983) notes two outpatient studies which found GAS increases in patients after treatment (Stone, 1983; Rehm, 1981).

More recently, Husby (1985) reported that the GAS was effective in reflecting change in neurotic patient

characteristics from the start of short-term dynamic psychotherapy to the termination of therapy. Youssef (1987) found that 15 hospitalized psychiatric patients with schizoaffective disorders who received twice weekly patient-family education classes showed more significant improvement on the GAS than controls. Stone (1987) conducted a 10-23 year followup of 254 borderline patients who had been treated with analytically oriented expressive psychotherapy and found that 66% of them had a good GAS outcome and 40% showed a "clinical recovery" according to the GAS. Hunt, Carr, Dagadakis, Christos, & Walker (1985) found that the GAS reflected a faster rate of improvement for patients matched with cognitively similar therapists for 12 sessions as compared to unmatched pairs. In a study of 70 male and 174 female psychiatric patients, Kirshner and Johnston (1983) found that women showed significantly greater responsiveness to treatment compared to the men as measured by the GAS. An outcome study of primal therapy by Dahl and Waal (1983) found that subjects who completed treatment showed moderate improvement on the GAS.

Finally, a drug study by Goldberg (1986) casts doubt on the GAS. Borderline and schizotypal personality disorder subjects were given either thiothixene or a placebo over a 12 week period. Although subjects taking thiothixene were observed to show significant reductions in illusions, ideas of reference, psychoticism, phobic anxiety, and obsessive-compulsivity compared to the placebo group, no significant change was found on the GAS.

Another method that has been used to obtain information about the construct validity of the GAS is to examine studies where the GAS and other instruments are used to measure severity of disturbance (Dekker, 1983). The degree to which the measures show similar results is an indication of the construct validity of the GAS.

The scale most frequently used along with the GAS in the 1970's and early 1980's was the Brief Psychiatric Rating Scale (BPRS) (Dekker, 1983). In most of the studies cited in this section the two scales were administered by the same rater.

Dekker's 1983 review summarized three studies (King & Goldstein, 1979; Bassuk & Gerson, 1980; Horowitz, Krupnick, Kaltrieder, Wilner, Leong, & Marman, 1981) which showed similar results in GAS and BPRS ratings. However, Larsen (1979) found that while the BPRS showed a significant positive effect in the outcome of therapy with a pre-therapy preparation interview, the GAS did not show a significant positive effect.

More recently, the BPRS and GAS were used in an outcome study to determine the effectiveness of rotating group psychotherapy leadership with chronic schizophrenic patients (Levin, Diamond, & Goldstein, 1985). The BPRS and GAS scores indicated that a group with rotating leadership was more improved than groups with 1 leader or 2 leaders.

Dekker's 1983 review cited three other articles (Rehm, 1981; Fink, Braden, & Qualls, 1982; Shenoy, Sadler, Goldberg, Hamer, & Ross, 1981) which found similar patterns

of scores between the GAS and other measures of severity of disturbance. However, three other studies (Baron, Gruen, & Asnis, 1982; Kanas, Rogers, Kreth, Patterson, & Campbell, 1980; Rounsaville, Weisman, Wilbur, & Kleber, 1982) found dissimilar patterns of scores on the GAS and concurrent measures of severity of disturbance.

Some additional studies have appeared in the last few years that relate to this type of construct validity. Fawcett, Edwards, Kravitz, & Jeffries (1987) investigated the relative effectiveness of alprazopam, desipramine, and an alprazopam-desipramine combination on depressed patients and found that the GAS, the Hamilton Depression Rating Scale, and the Hamilton Anxiety Scale all reflected a comparable degree of improvement at the end-point of treatment.

In a study of social network size and degree of psychopathology in substance abusers, Westermeyer and Nieder (1988) found that small network size was related to higher scores on a modified Michigan Alcoholism Screening Test, higher reported symptoms on the Symptoms Checklist-90 and the Beck Depression Inventory, more observed pathology on the BPRS, and lower scores on the GAS.

Finally, in a study on the effects of doxepin (Hameroff, Weiss, Lerman, Cork, Watts, Crago, Neuman, Womble, & Davis, 1984) the GAS and Hamilton Rating Scale for Depression showed similar improvements in depressed patients treated with doxepin as compared to those treated with a placebo.

In summary, looking at the validity evidence for the GAS as a whole, it must be concluded that the results are mixed at best. The content validity has been questioned because of the preponderance of behavioral descriptions in the lower part of the scale. With twice as many descriptions available to the rater at the anchor points from 0 to 50 compared to the range from 50 to 100, the scale appeared to be a more sensitive measure of severe psychiatric disturbance than mild to moderate disturbance. The concurrent validity has been specifically investigated very few times and in those instances the concurrent correlation coefficients are low to moderate. The lack of independent criteria in many studies casts doubt on their results. In terms of construct validity, there are numerous studies which demonstrate an agreement between the GAS and other instruments in measuring treatment effects, but exceptions to this trend could not be said to be rare. It should be noted that in most outcome studies, the GAS and other instruments have been administered by the same raters. This introduces the possibility of spuriously high *r*-values due to contamination.

The Global Assessment of Functioning Scale (GAF Scale)

Introduced in the DSM-III-R, the GAF Scale is a revision of the GAS (American Psychiatric Association, 1987). The range of the GAF Scale was reduced to 0-90, but there are still anchor points at ten point intervals. It is

clear that an effort was made to provide improved definitions for the anchor points in the upper half of the scale. Throughout the scale there is a better balance of succinct, more clearly worded clinical symptom descriptions and less formal behavioral examples, although there are still slightly more behavioral descriptions in the lower half of the scale.

There is no information provided in the DSM-III-R regarding the new scale's reliability or validity. In fact, no studies could be found in the research literature which used the GAF in any way. It appears that researchers are still using the GAS, and it remains to be seen whether or not the GAF Scale will supersede the GAS as a research instrument. Because of its inclusion in the DSM-III-R diagnosis system, however, it is safe to assume the GAF Scale will be used extensively in mental health program evaluation.

The Brief Psychiatric Rating Scale (BPRS)

The Brief Psychiatric Rating Scale (BPRS), first published in 1962 by Overall and Gorham, has been one of the most widely used general purpose psychiatric rating scales over the past 20 years. The scale as currently used, contains 18 items, each representing a separate symptom construct rated for severity on a 7-point scale ranging from 0 (not present) to 6 (extremely severe). The 18 items are reproduced in Appendix D. Five factor scores obtained by

summing ratings on related items provide composite measures of thinking disturbance, anxious depression, hostile suspiciousness, withdrawal retardation, and agitation excitement. A total pathology score is obtained by summing the ratings on all 18 BPRS symptom constructs.

Overall and Gorham developed the scale to evaluate treatment change in an efficient rapid manner while at the same time yielding a comprehensive description of patient characteristics (Overall & Gorham, 1962). The scale has been most frequently used to evaluate treatment response in controlled clinical drug trials. It has also been used in a wide range of other research.

Although several reviews or summaries of the BPRS have been published (Overall & Hollister, 1968; Overall & Klett, 1972; Lyerly, 1973; Overall, 1974; Guy, 1976) the best review of the psychometric properties of the BPRS, according to Overall (1988), is contained in an article by Hedland & Vieweg (1980). Out of over 300 articles surveyed, Hedland and Vieweg found 13 that reported BPRS reliability in terms of Pearson Product Moment Correlations as interrater reliability coefficients. Ten of the 13 reported reliability coefficients of .80 or more for the BPRS total pathology score. The authors found five studies which reported interrater reliability for the individual symptom scales, with a mean value of .75, and a range from .63 to .83. Interrater reliability for the higher-order factors ranged from .77 to .97 in the studies surveyed. Reliability estimates for the fifth factor, agitation excitement, have

not been reported due to its relatively recent addition to the factor scoring.

The validity of the BPRS was established by Hedland and Vieweg (1980) primarily by surveying over 150 published drug treatment studies that systematically used the BPRS with other measures to evaluate the treatment effect of different drug compounds. Over 35 standard rating measures were used one or more times each along with the BPRS in these studies. With only rare exceptions the BPRS reflected changes that were corroborated and supported by the other clinical ratings.

Concurrent validity of the BPRS had been specifically investigated in only 12 published studies at the time of the Hedlund and Vieweg (1980) review. These reported moderate (.58) to strong (.93) correlations between the BPRS and such scales as the Multidimensional Scale for Rating Psychiatric Patients ($\underline{r}$ = .93), the Clinical Global Impressions Scale ($\underline{r}$ = .86), the MMPI ($\underline{r}$ = .58). Additional studies which demonstrated relationships between the BPRS and other measures were cited by Hedland and Vieweg, but they did not report the obtained $\underline{r}$ values. These measures included the Multiple Affective Adjective Checklist, the Clinical Rating Scale, the Katz Adjustment Scale, the Psychotic Inpatient Profile, and the Psychotic Reaction Profile.

In 25 BPRS factor analyses reported in eight articles reviewed by Hedlund and Vieweg (1980), there is little doubt about the overall consistency of agreement about the BPRS higher order factors. This is in spite of the diversity of

patients, settings, cultures, and types of analyses conducted.

Interrater reliability of the BPRS is considered satisfactory by Hedlund and Vieweg (1980), and they conclude that it is a sensitive and effective measure of psychopathology and treatment related symptom changes.

Summary of Research Problem

In summary, the problem which is the basis of this study is that the psychometric properties of an important new measure of severity of psychiatric impairment, the GAF Scale, have not been adequately investigated. In fact, literature searches for articles using the GAF Scale have not uncovered a single study based on the GAF Scale. The scale from which the GAF Scale was derived, the GAS, has demonstrated only marginal concurrent validity in published studies, although it does exhibit adequate interrater reliability. Many validity studies of the GAS failed to use an independent criterion. Because the GAS and GAF Scale are so similar, questions arose regarding the reliability and validity of the GAF Scale. This investigation was conducted in order to evaluate those psychometric properties of the scale.

Hypotheses Examined

Hypothesis Related to Reliability

The hypothesis related to reliability was: There is a

significant interrater correlation when the GAF Scale is administered by independent raters.

Hypotheses Related to Concurrent Validity

The primary hypothesis with regard to concurrent validity was: There is a significant correlation between the GAF Scale and total pathology score taken from the BPRS when the scales are administered by independent raters.

The secondary hypotheses with regard to concurrent validity were:

(a) There is a significant correlation between the GAF Scale and each of the five higher order syndrome factors on the BPRS.

(b) There is a significant correlation between the GAF Scale and each of the 18 BPRS symptom constructs.

Construct validity was examined by comparing the mean GAF scores of severely disturbed clients (as defined by certain diagnostic categories) and the mild to moderately disturbed (certain other diagnoses).

Additionally, mean GAF scores of Chronically Mentally Ill (CMI) clients (See definition, Appendix E) were compared with those of non-CMI clients.

Hypotheses Related to Construct Validity

(a) The Mean GAF Score of clients with diagnoses considered severe is significantly lower than the mean GAF score of clients with diagnoses considered mild to moderate.

(b) The Mean GAF score of chronically mentally ill

clients is significantly lower than the mean GAF score of non-CMI clients.

CHAPTER III

METHOD

Design

The design of the study consisted of reliability and concurrent validity correlations between GAF Scale scores and several other variables which are shown in Figure 1. The statistical analyses were Pearson Product moment correlations (r). A two independent groups t-test was used to determine construct validity.

For testing the hypotheses, a $p < .05$ level of probability was determined as the level required for rejection of the null hypotheses.

Subjects

Subjects in the study were 62 adult (age 18 or older) outpatient clients at a mental health center serving a five county area in North Central Oklahoma. Only clients in active treatment who had been seen two times or more in face-to-face therapy sessions by both a staff psychiatrist and a staff therapist were included in the study.

This sample of 62 clients included 41 chronically mentally ill clients (as defined by the Oklahoma Department of Mental Health, Appendix E) and 21 non-chronically

mentally ill clients. The sample included 23 males and 39 females, and ranged in age from 22 to 75 years old, with a mean age of 36.2.

Instrumentation

Clients were rated by staff members on two scales, the GAF Scale and the BPRS. A complete discussion of the development and information known about the psychometric properties of these scales is presented in the review of the literature (Chapter II). For the purpose of this study several versions of the BPRS were integrated, resulting in somewhat amplified symptom construct descriptions. The BPRS as used in this study is presented in Appendix D.

Raters

Raters included in this study included one staff psychiatrist and nine therapists at the mental health center. One of the therapists was a licensed clinical social worker (LCSW), one possessed an MSW degree, four possessed a Masters (M.S.) degree, and three had B.S. degrees. While two of the raters had extensive professional experience (the psychiatrist and the LCSW) the remaining eight were considerably less experienced with between one and three years in the mental health field. Four of the therapist raters were graduate students working at the center in a part-time practicum status. Three of the raters were case managers who worked exclusively with chronically mentally ill clients and tended to carry the largest case

loads in the center. These three accounted for 58% of the therapist ratings while the other six therapists accounted for the remaining 42%.

While none of the raters had used the BPRS prior to the study, all had used the GAF in their work in the center. None of the raters had received inservice GAF Scale training while working at the mental health center. All raters were given a brief refresher training session on the use of the GAF and a more extensive and detailed training session in the use of the BPRS.

It should be noted that, in their clinical use of the GAF prior to this investigation, the raters had been subject to directions from the Oklahoma Department of Mental Health (DMH) to not exceed a GAF score of 50 for any chronically mentally ill client. In their role as raters for this study the staff members were specifically instructed to disregard this DMH directive and freely assign GAF ratings based on behavioral and symptom descriptions at the anchor points regardless of CMI status or diagnosis.

Procedure

Clients were rated independently by the psychiatrist and therapist who were currently treating them. The psychiatrist rated all 62 clients only on the GAF Scale, while the nine therapists rated only their own clients on both GAF Scale and BPRS. Ratings of any particular client were made independently by the psychiatrist and therapist during the same five-day work week, during which each had

THERAPISTS' GAF SCORES (INTERRATER RELIABILITY)

THERAPISTS' BPRS SCORES (CONCURRENT VALIDITY)

- Total Pathology Score
- Factor Scores
  - Thinking Disturbance
  - Withdrawal Retardation
  - Anxious Depression
  - Hostile Suspiciousness
  - Agitation Excitement
- Symptom Constructs
  - Somatic Concern
  - Anxiety
  - Emotional Withdrawal
  - Conceptual Disorganization
  - Guilt Feelings
  - Tension
  - Mannerisms and Posturing
  - Grandiosity
  - Hostility
  - Suspiciousness
  - Hallucinatory
  - Motor Retardation
  - Uncooperativeness
  - Unusual Thought Content
  - Blunted Affect
  - Excitement
  - Disorientation

Figure 1. Variables correlated with the psychiatrist's GAF scores

independently seen the client face-to-face. In most cases, the psychiatrist had seen the clients much less than the therapists in actual face-to-face contact over the course of their treatment, due to shorter sessions. The psychiatrist saw the clients an average of 15 minutes per session over the course of treatment, while the therapists averaged 45 minutes per session. Information regarding the current DSM-III-R diagnosis for each client and whether or not the client was chronically mentally ill (CMI) was obtained from the client chart. The therapists were aware of the CMI status of their clients but the psychiatrist was not.

The names of clients who were rated in the study were kept confidential and ratings were placed in client files only with written client consent. Rights of human subjects were safeguarded throughout the entire procedure.

GAF Scale ratings by the psychiatrist and therapists were correlated to provide an interrater correlation coefficient. GAF scale ratings by the psychiatrist were correlated with BPRS ratings from the staff therapists to provide concurrent validity coefficients. The statistic used in analyzing the data was the Pearson product moment correlation.

In order to study the construct validity of the GAF, clients were divided by the present author into two groups according to diagnosis. The decision regarding which diagnoses would be designated severe and which mild to moderate was made by the author, who had no knowledge of the subject's GAF scores. The two diagnostic groups were

defined as follows: (1) the severely disturbed; any client with a diagnosis of Schizophrenia, Delusional Disorder, Psychotic Disorders not classified elsewhere, Bipolar Disorder, Major Depression with Psychotic Features, Major Depressive Episode, Severe Borderline, Schizoid, or Schizotypal Personality Disorders, Organic Mental Syndromes, or Mental Retardation, and (2) the mild to moderately disturbed: Psychoactive Substance Use Disorders, Mild to Moderate Mood Disorders, Somatoform Disorders, Dissociative Disorders, Sexual Disorders, Sleep Disorders, Factitious Disorders, Impulse Control Disorders not classified elsewhere, Adjustment Disorders, Eating Disorders, and other Personality Disorders. A composite GAF Score for each individual client was calculated by averaging the ratings of the psychiatrist and therapist. A two independent groups t-test was utilized to determine if the composite GAF scores of these two groups were significantly different from one another. A similar analysis was conducted to determine if the composite GAF scores of the CMI and non-CMI clients were significantly different.

CHAPTER IV

RESULTS

The mean of GAF scores assigned by the psychiatrist was 52.76 (SD = 11.73) as compared with the mean of GAF scores assigned by the nine therapists of 48.74 (SD = 17.04). The difference between these means was not significant, t (122) = 1.53, $p < .05$. The GAF scores assigned by the psychiatrist and therapist for each subject were averaged to yield a composite GAF score. The mean composite GAF score was 51.13 (SD = 12.22). These scores are presented in Table 1.

TABLE 1

GAF SCORES

	Mean	SD
Psychiatrist GAF	52.76	11.73
Therapist GAF	48.74	17.04
Composite GAF	51.13	12.22

On the BPRS, scores were obtained on 18 symptom constructs, on 5 factors, and on a total pathology measure. Scores on each of the 18 symptom constructs had a possible range from 0 to 6. Symptom constructs with the highest mean scores were Anxiety (3.61), Depressive Mood (2.42),

Suspiciousness (2.08), and Somatic Concern (2.08). Those with the lowest mean scores were Disorientation (.23), Uncooperativeness (.61), Mannerisms and Posturing (.74), and Grandiosity (.84). A summary of all BPRS mean scores is presented in Table 2. The five factor scores, obtained by adding three associated symptom construct scores (See Table 2) each had a possible range from 0 to 18. The factor score means were as follows: Factor I Thinking Disturbance (4.44), Factor II Withdrawal-Retardation (4.45), Factor III Anxious Depression (7.24), Factor IV Hostile-Suspiciousness (4.06) and Factor V Agitation Excitement (4.11). The Total Pathology Score for a particular subject was obtained by adding all 18 symptom construct scores. The mean BPRS Total Pathology Score was 27.53 (SD = 16.47) with an obtained sample range of 0 to 60. This compares with a total possible range of 0 to 108.

Testing of Reliability Hypotheses

An interrater correlation coefficient of .47 was obtained when the 62 GAF ratings assigned by the psychiatrist were correlated with those assigned by the nine therapists, who each rated only their own clients. While this value is significant at the $p < .001$ level, it is below the generally acceptable range for reliability coefficients of .70 or above.

Testing of Concurrent Validity Hypotheses

Results indicate that the correlation between GAF

TABLE 2

BPRS MEAN SCORES

	Mean	SD
BPRS Total Pathology Score	27.53	16.47
BPRS SYMPTOM CONSTRUCTS		
1. Somatic Concern	2.08	1.87
2. Anxiety	3.61	5.47
3. Emotional Withdrawal	1.76	1.56
4. Conceptual Disorganization	1.60	1.59
5. Guilt Feelings	2.00	1.77
6. Tension	2.16	1.71
7. Mannerism & Posturing	0.74	1.31
8. Grandiosity	0.84	1.42
9. Depressive Mood	2.42	1.66
10. Hostility	1.32	1.45
11. Suspiciousness	2.08	1.80
12. Hallucinatory Behavior	1.44	1.97
13. Motor Retardation	1.05	1.44
14. Uncooperativeness	0.61	1.19
15. Unusual Thought Content	1.40	1.58
16. Blunted Affect	1.68	1.54
17. Excitement	1.21	1.47
18. Disorientation	0.23	0.83

TABLE 2 (Continued)

BPRS FACTORS (with associated symptom constructs)		Mean	SD
I. Thinking Disorder	Conceptual Disorganization Hallucinatory Behavior Unusual Thought Content	4.44	4.59
II. Withdrawal-Retardation	Emotional Withdrawal Motor Retardation Blunted Affect	4.45	3.87
III. Anxious Depression	Anxiety Guilt Feelings Depressive Mood	7.24	4.12
IV. Hostile-Suspiciousness	Hostility Suspiciousness Uncooperativeness	4.06	3.56
V. Activation	Tension Mannerisms & Posturing Excitement	4.11	3.57

ratings assigned by the psychiatrist and BPRS Total Pathology Scores as assigned by the therapists was -.42, which is in the expected direction and is significant at the $p < .001$ level. Correlations between the GAF and the 5 BPRS factors varied widely. A significant correlation ($p < .001$) of -.52 was obtained between the GAF and the BPRS Thinking Disturbance factor. The only other BPRS factor to correlate significantly with the GAF was Withdrawal-Retardation which correlated -.31, significant at the $p < .05$ level. The other factors were correlated at non-significant levels as follows: Anxious Depression -.22, Hostile Suspiciousness -.13, and Agitation Excitement -.26. Of the 18 BPRS symptom constructs, 6 correlated significantly with the GAF. These included Conceptual Disorganization -.52 ($p < .001$), Hallucinatory Behavior -.45 ($p < .001$), Motor Retardation -.50 ($p < .001$), Unusual Thought Content -.41 ($p < .001$), Depressive Mood -.39 ($p < .01$), and Grandiosity -.38 ($p < .01$). The complete listing of concurrent validity coefficients is found in Table 3.

Testing of Construct Validity Hypotheses

Severely disturbed clients as defined by their diagnosis (See Chapter II) were expected to have significantly lower composite GAF ratings than less disturbed clients. The difference found in the group means was as expected. The severe diagnosis group had a mean composite GAF score of 49.6 (SD = 12.3) while the mild to moderate diagnosis group had a mean composite GAF score of

TABLE 3

CORRELATION OF PSYCHIATRIST GAF RATINGS AND THERAPIST GAF AND BPRS RATINGS

Variable	Correlation Coefficient (r)	Common Variance (r^2)
Therapist GAF (interrater reliability)	.47***	.22
Therapist BPRS Scores (concurrent validity)		
Total Pathology Score	-.42***	.18
Factor Scores		
Thinking Disturbance	-.52***	.27
Withdrawal-Retardation	-.31*	.10
Anxious Depression	-.22	.05
Hostile Suspiciousness	-.13	.02
Agitation Excitement	-.26	.07
Symptom Constructs		
Somatic Concern	-.20	.04
Anxiety	+.09	.01
Emotional Withdrawal	-.10	.01
Conceptual Disorganization	-.52***	.27
Guilt Feelings	+.04	.002
Tension	+.16	.03
Mannerisms and Posturing	-.22	.05
Grandiosity	-.38**	.14

*** $p < .001$
** $p < .01$
* $p < .05$

TABLE 3 (Continued)

Variable	Correlation Coefficient (r)	Common Variance (r^2)
Depressive Mood	-.39**	.15
Hostility	-.10	.01
Suspiciousness	-.24	.06
Hallucinatory Behavior	-.45***	.21
Motor Retardation	-.50***	.25
Uncooperativeness	+.09	.01
Unusual Thought Content	-.41***	.17
Blunted Affect	-.24	.06
Excitement	-.23	.05
Disorientation	-.23	.05

*** $p < .001$
** $p < .01$
* $p < .05$

57.2 (SD = 8.9). The difference was significant, t (62) = -2.17, $p < .05$. Of the 62 subjects, 48 had been assigned a severe diagnosis. Clients classified as chronically mentally ill (CMI) were expected to have lower composite GAF scores than non-CMI clients. These group means also differed in the expected direction with the CMI group mean of 48.8 (SD = 13.1) compared to the non-CMI group mean of 56.0 (SD = 8.5). The difference is significant, t (62) = -2.24, $p < .05$. Of the 62 subjects, 40 were classified as CMI and 22 were non-CMI.

Supplemental Analysis of Data

A possible source of error in obtaining the interrater reliability coefficient in this study stemmed from the fact that not all the therapist GAF ratings were obtained from the same therapist. This introduced a problem in interpreting the reliability coefficient because the individual differences among the nine raters undoubtedly contributed to some increased error variance in their scores compared to the psychiatrist's ratings, which were all made by the same person.

In order to investigate whether or not the therapists' ratings contained excessive error compared to the psychiatrist's ratings, both sets of GAF ratings (psychiatrist's and therapists'), were separately used to compute t-values with two outside criteria - CMI vs. non-CMI and severe diagnoses vs. mild to moderate diagnoses. If excessive error variability existed in the therapists' GAF

ratings, it would be expected to show up in a lower *t*-value, compared to the psychiatrist's ratings, with each outside criterion.

With regard to the CMI vs. non-CMI comparison, the psychiatrist's mean GAF rating for CMI clients was 50.49 (SD = 12.52), while the psychiatrist's mean for non-CMI clients was 57.19 (SD = 8.64). The difference was significant, t (62) = -2.2, $p < .05$. The therapists' mean GAF rating for CMI clients was 45.65 (SD = 18.08) and for non-CMI clients their mean was 54.76 (SD = 13.19). The difference was significant t (62) = -2.04, $p < .05$. With regard to the severe diagnosis vs. mild to moderate diagnosis comparison, the psychiatrist's mean GAF rating for severely diagnosed clients was 51.44 (SD = 11.81). The psychiatrist's mean rating for the mild to moderate diagnosed clients was 57.27 (SD = 10.61). This difference was significant, t = -1.67, $p < .10$. The therapist's mean rating for severely diagnosed clients was 46.37 (SD = 17.43) while their mean rating for the mild to moderately diagnosed clients was 56.85 (SD = 13.11). This difference was significant, t = -2.08, $p < .05$. To summarize these results, the therapists' ratings did not result in lower t values compared to the psychiatrist's ratings.

Another method of uncovering the suspected excessive variability of the therapist's ratings would involve simply observing the extent of differences in their mean GAF ratings. However, in this study such a technique is not appropriate because the means of GAF ratings made by the

individual therapists were obviously affected by the types of clients they typically were seeing in the mental health center. The three case managers worked only with CMI clients and thus could be expected to have assigned lower mean GAF ratings. The practicum student therapists saw relatively few CMI clients and thus would be expected to make higher mean GAF ratings. For instance, the four practicum student raters had a mean GAF rating of 58.1 (N =10). The least experienced case manager had a mean GAF rating of 66.25 (N = 16), but the other two case managers had a combined mean GAF score of 36.1 (N = 20). The overall case manager mean GAF rating was 50.91 (N = 35). The most experienced rater, the LCSW, had a wide variety of clients and produced a mean GAF rating of 46.47 (N = 15).

Interrater correlation coefficients were obtained comparing the psychiatrist's GAF ratings and each of the two therapist raters who made the most GAF ratings. The therapist who made the most GAF ratings (N = 16) was a case manager with the least clinical experience, the least experience using the GAF Scale, and the least mental health related academic training among the therapist raters. This therapist's GAF ratings correlated -.07 with the psychiatrist's ratings of the same clients. The therapist rater (the LCSW) who made the next highest number of GAF ratings (N = 15) was the most experienced clinician in the study, as well as most experienced in the use of the GAF. This therapist was also the most academically trained of the therapist raters. This therapist's GAF ratings correlated

.86 with the psychiatrist's ratings ($p < .001$).

Finally, evidence was obtained demonstrating that concurrent validity coefficients based on non-independent ratings can be artificially inflated. Correlation of GAF ratings assigned by therapists with BPRS Total Pathology Scores also assigned by therapists yielded a coefficient of -.82. A correlation coefficient of -.75 was obtained when comparing composite GAF Scores with BPRS Total Pathology Scores.

CHAPTER V

DISCUSSION

The purpose of this study was to investigate the reliability and validity of the Global Assessment of Functioning Scale (GAF Scale). Possibly because the GAF Scale is a revision of the widely accepted Global Assessment Scale (GAS), there has been little or no effort to investigate the psychometric properties of the revised version. A comprehensive review of the literature regarding the reliability data available on the GAS (Dekker, 1983) showed that the scale has been judged to be reliable. However, Dekker (1983) noted that reliability studies on the GAS have been done in tightly controlled research or clinical settings. Furthermore, an analysis of data presented in Dekker's review revealed that the reliability of the scale reached or exceeded an acceptable level (.70 or above) only when raters were trained in the use of the GAS. Studies which simulated the spotty or nonexistent inservice training on the use of the GAF which characterizes many mental health treatment settings found that the reliability of the GAS was not at an acceptable level. The literature review also revealed a striking difference in the reliability of the GAS in favor of highly educated professional raters as compared with poorly educated raters.

From these findings it would not be surprising to find that, in actual practice, mental health programs might have trouble getting reliable GAF ratings. Spotty training procedures for the use of the newer GAF Scale, centers staffed with increasingly more poorly educated therapists due to budget cuts, and centers requiring those therapists to carry increasingly large case loads, again due to budget problems resulting in understaffing, are all conditions which exist in the real world as opposed to tightly controlled research settings.

Despite these problems which may adversely affect the use of the GAF Scale in clinical settings, the scale is being increasingly used in program evaluation and quality assurance programs in mental health delivery systems. A standard measure of client functioning such as the GAF Scale will be used in many important applications such as outcome research, treatment evaluation, cost effectiveness studies, client placement decisions, etc. (Ciarlo, 1982). A problem with the increasing use of the GAF Scale is that there has been no research to determine how much staff GAF Scale training and level of staff academic training is necessary to ensure reliable use of the scale in actual clinical settings.

The results of the present study demonstrate that the GAF Scale is not a particularly reliable instrument as used by the mental health clinic staff who served as raters. Although the interrater reliability coefficient (.47) was significant at the $p < .001$ level of confidence, there was

only 22% common variance in the two distributions of scores (r^2 = .2209). An interrater correlation coefficient of .70 or higher is usually considered adequate as a reliability coefficient. This unacceptably low level of interrater reliability is the most striking finding of the present study, because of the pattern of evidence in prior studies that suggested the GAF Scale's predecessor, the GAS, was a reliable instrument.

Any attempt to analyze this unexpected result must focus at least partially on the raters who used the GAF Scale in the study. To summarize, the ratings of a licensed psychiatrist were correlated with those of nine staff therapists (three B.S. degrees, one L.C.S.W., one M.S.W., and four M.S.). Four of the raters (three M.S. and one B.S.) were graduate students working at the center half-time time in a practicum-therapist role. Three of the raters (one M.S. and two B.S.) were case managers who worked exclusively with chronically mentally ill clients. Only the psychiatrist and LCSW, a full time outpatient therapist, could be considered to be experienced clinicians. The rest had three years experience or less in mental health therapist positions. The three case managers, who tended to carry the largest case loads in the center, accounted for 36 of the 62 therapist GAF Scale ratings (58%). The other six therapists accounted for the remaining 26 ratings (42%).

All of the raters were familiar with the GAF Scale and used it in their work. None of the raters, including the psychiatrist, had received any formal inservice training in

the use of the scale prior to the study. As was the custom at this particular center, staff members were expected to get training in the use of the GAF Scale from their supervisor if they needed it, but there was no written policy in this regard. All raters were given a 20 minute briefing on the GAF Scale immediately preceding the implementation of the study.

It is likely, based on past research, that the characteristics of this mental health center staff - spotty training in use of GAF Scale, and relatively low experience and education - contributed to the low obtained reliability coefficient. The use of multiple therapist raters undoubtedly introduced some extra error variability due to individual differences in the therapist's rating techniques. However, it was found that the therapist's GAF ratings were as powerful as the psychiatrist's ratings in differentiating between severely disturbed and non-severely disturbed groups of clients. Also, when the least experienced therapist rater's GAF scores were correlated with the psychiatrist's GAF scores, a very low interrater correlation coefficient of -.07 resulted. This compared with a coefficient of .87 ($p < .001$) which was obtained when the most experienced therapist rater's scores were compared with the psychiatrist's. Although these correlations were based on low N's, they suggest that clinical experience and training are important factors in the reliable use of the GAF Scale.

Also, it should be noted that the psychiatrist had spent much less actual face-to-face time with the clients

than the therapists in this study. However, there had been frequent consultations between the therapists and psychiatrist, and the psychiatrist was well informed about the clients in terms of history and current treatment issues. Nevertheless, it is possible that the scores of the psychiatrist contained more error because of the lack of face-to-face time with each client relative to the therapists.

Even though the psychiatrist was aware of the history, life circumstances, and current condition of the clients, the fact remains that the psychiatrist spent much less face-to-face time with the clients over the course of treatment. The psychiatrist/client average face-to-face session time was 15 minutes. The therapist/client average face-to-face session time was 45 minutes. This difference could have contributed to the lower obtained interrater reliability coefficient.

Despite the importance of the GAF Scale, there is no standardized training procedure available to mental health programs. Unlike the authors of the HSRS who produced 32 case vignettes tied to ratings every 5 points up the scale, there are no such training aids available for the GAF Scale. Each mental health treatment program is left to devise its own training procedure and these obviously may vary widely from program to program if they are implemented at all. According to Pokorny (in press), the Oklahoma Mental Health Research Institute has found that training in the use of the GAF Scale is simply not occurring.

With regard to the concurrent validity of the GAF Scale, it is difficult to interpret validity data when the reliability is not at an adequate level. If a scale is not reliable it cannot be valid. However, despite the fact that the interrater reliability coefficient (.47) did not reach the generally acceptable level of .70, it was a statistically significant correlation ($p < .001$ level) so it may be prudent to attempt to analyze the concurrent validity data produced in this study.

Based on past research on the GAS, it was in the area of concurrent validity that the GAF Scale was expected to show the greatest weakness. According to Dekker (1983) the GAS never demonstrated good concurrent validity with another psychometric measure of severity of disturbance in any well controlled study with psychiatric patients.

The difficulty in interpreting the concurrent validity data in this study relates to the fact that with an interrater reliability coefficient of .47 there is only 22% common variance between the two distributions of ratings. The remaining 78% was due to some type of error. It is entirely possible that the psychiatrist was using the scale improperly and this accounted for most of the error. Actually, there is no way to be sure where the error arose. It is not possible to know whether or not the therapists' BPRS scores were correlated with error-ridden GAF Scores produced by the psychiatrist, or vice versa. Also, the use of multiple therapist raters must have contributed to the error variance of the therapists' GAF scores due to the

individual differences among the raters. The problems of interpretation notwithstanding, there is a pattern which emerges in the concurrent validity data which warrants discussion. The concurrent validity coefficient of -.42 (between the GAF Scale and BPRS Total Pathology Score) is significant ($p < .001$) and would ordinarily be considered to be in the moderate and acceptable range of validity coefficients. While that result may be suspect, because a scale cannot be considered valid if it is not reliable, it is interesting that there is a preponderance of significant correlations between the GAF Scale and the BPRS measures of psychoticism (Thinking Disturbance factor and the symptom constructs Conceptual Disorganization, Hallucinatory Behavior, Unusual Thought Content, Grandiosity), as opposed to correlations with non-psychotic illness. This suggests that the GAF Scale, like the GAS before it, is more sensitive to the presence of psychotic illness than it is to other, perhaps less severe types of disturbance.

The effect of rater error in using the GAF Scale was reduced somewhat in the portion of the study which investigated construct validity. Each client was given a composite GAF Score based on the average of the psychiatrist's and the therapist's rating. According to Green, Nguyen, and Attkisson (1979) this method produces significantly increased reliability of ratings. Utilizing these composite GAF scores, it was found that groups of clients with severe diagnoses as defined by the experimenter were significantly lower in mean GAF scores than those with

mild to moderate diagnoses ($p < .05$). Also, CMI clients obtained significantly lower GAF scores than non-CMI clients ($p < .05$). This corresponds with previously published data suggesting that the GAS was capable of adequately distinguishing between groups of clients based on group means (Dekker, 1983).

Recommendations

Based on the findings of this study it is recommended that the reliability of the GAF Scale be further investigated to determine what level of GAF rater training is necessary to make the scale reliable as a standard measure of client pathology in clinical settings. Research is also needed to determine what particular training procedure is most effective in producing reliable scores. Once identified, this training procedure should be standardized by the American Psychiatric Association, which published the GAF Scale, and presented at mental health treatment settings nationwide on a recurring basis. A standardized collection of case vignettes tied to ratings at various points on the scale, such as that provided with the HSRS, would be a helpful addition by the authors of the GAF Scale. Further research is also needed to clarify how much clinical experience and academic training are necessary to use the scale reliably. Regardless of whether or not such research is done, there is enough evidence available to recommend that mental health programs using the GAF in their data collection systems take the following actions: (1)

devise a standardized training program of their own for the use of the GAF Scale (2) require that all clinical staff receive training periodically (3) institute clinical privileging policies for the use of the GAF Scale to include requirements that raters be trained and that they have attained a specified level of education. Finally, it is recommended that further research be done regarding the concurrent validity of the GAF Scale.

Specific follow up studies might involve comparisons between the GAF ratings of an experienced psychiatrist and individual therapists rather than multiple raters. Therapist raters of differing levels of clinical experience, academic training, and GAF Scale training could be individually compared to the psychiatrist in terms of GAF scores (interrater reliability) and therapist BPRS scores vs. psychiatrist GAF scores (concurrent validity). Such a technique would eliminate the extra error variance inherent in the use of multiple therapist raters, but at the same time retain the advantage of conducting the study in an actual clinical setting. However, the problem of obtaining sufficient N's for each correlation procedure might make it difficult to implement these procedures in clinical settings.

Conclusions

There is evidence presented in this study that suggests that the GAF Scale is not a reliable instrument as used in some mental health treatment settings. The lack of

reliability appears to stem from the inadequate GAF Scale training, limited clinical experience, and low level of academic training that characterizes many staff members who are called upon to use the scale in clinical settings. It must be acknowledged, however, that excessive error variance in the therapist's ratings may have been produced by the use of multiple raters.

Conclusions regarding the validity of the GAF Scale based on these results are not possible due to the lack of adequate reliability. However, it appears the scale may be more sensitive in assessing psychotic illness than it is in assessing non-psychotic illness. Questions remain about whether or not the GAF Scale is a valid measure of the construct of severity of psychological disturbance.

The GAF Scale does appear to be valid as a measure for distinguishing between contrasting groups, which are known to be different. The validity of individual GAF scores remains in doubt.

BIBLIOGRAPHY

American Psychiatric Association (1987). Diagnostic and Statistical Manual of Mental Disorders, Third Edition, Revised. Washington, D.C., American Psychiatric Association.

Baron, M., Gruen, R., & Asnis, L. (1982). Schizophrenia: A comparative study of patients with and without family history. British Journal of Psychiatry, 140, 516-517.

Bassuk, E., & Gerson S. (1980). Chronic crisis patients: A discrete clinical group. American Journal of Psychiatry, 137, 1513-1517.

Battista, J. R. (1982). Empirical test of Vaillant's hierarchy of ego functions. American Journal of Psychiatry, 139, 356-357.

Ciarlo, J. A. (1982). Accountability revisited: The arrival of client outcome evaluation. Evaluation and Program Planning, 5, 31-36.

Ciarlo, J. A., Edwards, D. W., Kiresuk, T. J., Newman, F. L., & Brown, T. R. The assessment of client/patient outcome techniques for use in mental health programs (Final Report, Contract No. 278-80-0005-DB). Rockville, MD: National Institue of Mental Health, 1981.

Curran, J. P., Miller, I. W., Monti, P. M., Zwick, W. R., & Stout, R. L. (1980). The socially inadequate patient: Incidence rate, demographic and clinical features, and hospital and posthospital function. Journal of Consulting and Clinical Psychology, 48, 375-382.

Dahl, A. A. & Waal, H. (1983). An outcome study of primal therapy. Psychotherapy and Psychosomatics, 39, 154-164.

Dekker, D. J. (1983). A Study of the Validity of the Global Assessment Scale. Unpublished Dissertation. Western Michigan University, Kalamazoo, MI.

Endicott, J., Spitzer, R. L., Fleiss, J. L., & Cohen, J. (1976). The Global Assessment Scale: A procedure for measuring overall severity of psychiatric disturbance. Archives of General Psychiatry, 1976, 33, 766-771.

Fawcett, J. Clark, D. C., Schneftner, W. A., & Gibbons, R. D. Assessing anhedonia in psychiatric patients. Archives of General Psychiatry, 1983, 40, 79-84.

Fawcett, J., Edwards, J. H., Kravitz, H. M., & Jeffries, H. (1987) Alprazopam: An antidepressant? Alprazopam, desipramine, and an alprazopam-desipramine combination in the treatment of adult depressed outpatients. Journal of Clinical Psychopharmacology, 7, 295-310.

Fink, E.B., Braden, W., and Qualls, C.B. (1982). Predicting pharmacotherapy outcome by subjective response. Journal of Clinical Psychiatry, 40, 79-84.

Goldberg, S.C. (1986). Borderline and Schizotypal personality disorders treated with low-dose thiothixene vs. placebo. Archives of General Psychiatry, 43, 680-686.

Goldstein, M. J. Family therapy during the aftercare treatment of acute schizophrenia. In J. S. Strauss, M. Bowers, T W. Downey, S. Fleck, S. Jackson & I. Levine (Eds.), Psychotherapy of Schizophrenia. New York: Pleanum, 1980.

Green, R., Nguyen, T., & Attkisson, C. Harnessing the reliability of outcome measures. Evaluation and Program Planning, 1979, 2, 137-142.

Gudeman, J. E., Dickey, B., Rood, L., Hellman, S., & Grinspoon, L. (1981). Alternative to the backward: The quarterway house. Hospital and Community Psychiatry, 32, 330-334.

Guy, W. (1976). ECDEU Assessment Manual for Psychopharmacology, DHEW No. (ADM) 76-338. Rockville, MD, NIMH.

Hagedorn, H. J., Beck, K. J., Neubert, S. F., & Werlin, S. H. A working manual of simple program evaluation techniques for community mental health centers (DHEW Publication No. ADM 76-404). Washington, D. C.: U.S. Government Printing Office, 1976.

Hameroff, S. R., Weiss, J. L., Lerman, J. C., Cork, R. C., Watts, K. C., Crago, B. R., Neuman, C. P., Womble, J. R., & Davis (1984). Doxepin's effects on chronic pain and depression: A controlled study. Journal of Clinical Psychiatry, 45, 47-53.

Hargreaves, W. A., McIntyre, M. H., Attkisson, C. C., & Siegel, L. M. Outcome measurement instruments for use in community mental health program evaluation. In W. A. Hargreaves, C. C. Attkisson, & J. E. Sorenson (Eds.), Resource materials for community mental health program evaluation (2nd ed.) (DHEW Publication No. ADM 77-328). Washington, D. C.: U.S. Government Printing Office, 1977.

Hedlund, J. L. & Vieweg, B. W. (1980). The brief psychiatric rating scale (BPRS): A comprehensive review. Journal of Operational Psychiatry, 11, 48-65.

Herz, M. I., Endicott, J., & Spitzer, R. L. (1975). Brief hospitalization of patients with families: Initial results. American Journal of Psychiatry, 132, 413-418.

Holcomb, W. R., & Otto, R. L. (1988). Concurrent validity of the Global Assessment Scale: What's in a number? Psychological Reports, 64, 455-461.

Horowitiz, M. J., Krupnick, J., Kaltreider, N., Wilner, N., Leong, A., & Marman, C. (1981). Initial psychological response to parental death. Archives of General Psychiatry, 38, 316-323.

Hunt, D. D., Carr, J. E., Dagadakis, C. S., & Walker, E. A. (1985). Cognitive match as predictor of psychotherapy outcome. Psychotherapy, 22, 718-721.

Husby, R. (1985). Short-term dynamic psychotherapy: V. Global assessment scale as an instrument for description and measurement of changes in 33 neurotic patients. Psychotherapy and Psychosomatics, 43, 28-31.

Kansas, N., Rogers, M., Kreth, E., Patterson, L., & Campbell, R. (1980). The effectiveness of group psychotherapy during the first three weeks of hospitalization. Journal of Nervous and Mental Disease, 168, 487-492.

King, C. E., & Goldstein, M. J. (1979). Therapist ratings of achievement of objectives in psychotherapy with acute schizophrenics. Schizophrenia Bulletin, 168, 487-492.

Kirshner, L. A. & Johnston, L. (1983). Effects of gender on inpatient psychiatric hospitalization. Journal of Nervous and Mental Disease, 17, 651-657.

Larkin, A. R. (1979). The form and content of schizophrenic hallucinations. American Journal of Psychiatry, 136, 940-943.

Larsen, D. L (1979). Enhancing client utilization of outpatient services. Dissertation of Abstracts International, 39, 4041B.

Lefkovit, P. M., Morrison, D. P., & Davis, H. J. (1982). The assessment of current functioning scale (ACFS). Journal of Psychiatric Treatment and Evaluation, 4, 297-305.

Levin, D., Diamond, R., & Goldstein, S. (1985). A study of alternating leadership in group psychotherapy in an aftercare clinic. Perspectives in Psychiatric Care, 23, 33-38.

Luborsky, L. (1962). Clinicians' judgement of mental health: A proposed scale. Archives of General Psychiatry, 7, 407-417.

Luborsky, L., & Bachrach, H. (1974). Factors influencing clinicians' judgements of mental health: Eighteen experiences with the Health-Sickness Rating Scale. Archives of General Psychiatry, 31, 292-299.

Lyerly, S. B. (1973). Handbook of Psychiatric Rating Scales, (DHEW No. 73-9061). Rockville, MD, NIMH.

Newman, F. L. (1980). Global scales: Strengths, uses, and problems of global scales as an evaluation instrument. Evaluation and Program Planning, 3, 257-268.

Overall, J. E. (1988). Brief psychiatric rating scale. In M. Hersen and A. S. Bellack (Eds.) Dictionary of Behavioral Assessment Techniques, Elmsford, New York: Pergamon Press.

Overall, J. E. (1974). The brief psychiatric rating scale in psychopharmacology research. In P. Pickot (Ed.), Modern Problems of Pharmacopsychiatry, Vol. 7, Basel, Switzerland: Karger.

Overall, J. E. & Gorham, D. R. (1962). The brief psychiatric rating scale. Psychological Reports, 10, 799-812.

Overall, J. E., & Hollister, L. E. (1968). Studies of quantitative approaches to psychiatric classification. In M. M. Katz, J. O. Cole, and W. E. Barton (Eds.), The Role and Methodology of Classification in Psychiatry and Psychopathology, (PHS Pub. #1584). Washington, D.C.: U.S. Dept. of Health, Education, and Welfare.

Overall, J. E. & Klett, C. J., (1972). Applied Multivariate Analysis. New York, McGraw-Hill.

Pokorny, L. J. (in press). A summary measure of clients' levels of functioning: Progress and challenges for use within mental health agencies. Journal of Mental Health Administration.

Rhem, L. P. (1981). A self-control therapy program for treatment of depression. In J.F. Clarkin and H.I. Glazer (Eds.), Depression: Behavioral and Directive Intervention Strategies. New York: Garland Publishing.

Rounsaville, B. J., Weisman, W. W., Wilber, C. H., and Kleber, H. D. (1982). Pathways to opiate addiction: and evaluation of differing antecedents. British Journal of Psychiatry, 141, 437-446.

Shenoy, R. S., Sadler, A. G., Goldberg, S. C., Hamer, R. M., & Ross, B. (1981). Effects of a six-week drug holiday on symptom status, relapse, and tardive dyskinesia in chronic schizophrenics. Journal of Clinical Psychopharmacology, 1, 324-326.

Sohlberg, S. (1989). There's more in a number than you think: New validity data for the global assessment scale. Psychological Reports, 64, 455-461.

Sorenson, J. L., Hargreaves, W. S. and Friedlander, S. (1982). Child global rating scales: Selecting a measure of client functioning in a large mental health system. Evaluation and Program Planning, 5, 337-347.

Stone, M. H. (1979) A psychoanalytic approach to abnormalities of temperament. American Journal of Psychotherapy, 33, 263-280.

Stone, M. H. (1987). Psychotherapy of borderline patients in light of long-term follow-up. Bulletin of the Menninger Clinic, 51, 231-247.

Stone, M. H. (1983). Psychotherapy with schizotypal borderline patients. Journal of the American Academy of Psychoanalysis, 11, 87-111.

Vaillant, G. E. & Schnurr, P. (1988). What is a case? A 45-year study of psychiatric impairment within a college sample selected for mental health. Archives of General Psychiatry, 45, 313-319.

Westermeyer, J., & Neider, J. (1988). Social networks and psychopathology among substance abusers. American Journal of Psychiatry, 145, 1265-1269.

Youssef, F. A. (1987). Discharge planning for psychiatric patients: The effects of a family-patient teaching program. Journal of Advanced Nursing, 12, 611-616.

Zheng, Y., Zhao, J., Phillips, M., Liu, J., Cai, M., Sun, S., & Huang, M. (1988). Validity and reliability of the Chinese Hamilton Depression Rating Scale. British Journal of Psychiatry, 152, 660-664.

APPENDIX A

GLOBAL ASSESSMENT OF FUNCTIONING SCALE

(GAF SCALE)

Global Assessment of Functioning Scale (GAF Scale)

Consider psychological, social, and occupational functioning on a hypothetical continuum of mental health-illness. Do not include impairment in functioning due to physical (or environmental) limitations. See p. 20 for instructions on how to use this scale.

Note: Use intermediate codes when appropriate, e.g., 45, 68, 72

Code

90 Absent or minimal symptoms (e.g., mild anxiety before an exam), good functioning in all areas, interested and involved in a wide range of activities, socially effective, generally satisfied with life, no more than everyday problems or concerns (e.g., an occasional
81 argument with family members).

80 If symptoms are present, they are transient and expectable reactions to psychosocial stressors (e.g., difficulty concentrating after family argument); no more than slight impairment in social, occupational, or school functioning (e.g., temporarily falling behind in
71 school work).

70 Some mild symptoms (e.g., depressed mood and mild insomnia) OR some difficulty in social, occupational, or school functioning (e.g., temporarily falling behind
61 in school work).

60 Moderate symptoms (e.g., flat affect and circumstantial speech, occasional panic attacks) OR moderate difficulty in social, occupational, or school functioning (e.g., few friends, conflicts with co-
51 workers).

50 Serious symptoms (e.g., suicidal ideation, severe obsessional rituals, frequent shoplifting) OR any serious impairment social, occupational, or school
41 functioning (e.g., no friends, unable to keep a job).

40 Some impairment in reality testing or communication (e.g., speech is at times illogical, obscure, or irrelevant) OR major impairment in several areas, such as work or school, family relations, judgement, thinking, or mood (e.g., depressed man avoids friends, neglects family, and is unable to work; child frequently beats up younger children, is defiant at
31 home, and is failing at school).

30 Behavior is considerably influenced by delusions or hallucinations OR serious impairment in communication or judgment (e.g., sometimes incoherent, acts grossly inappropriately, suicidal preoccupation) OR inability to function in almost all areas (e.g., stays in bed all
21 day; no job, home, or friends).

20 Some danger of hurting self or others (e.g., suicide attempts without clear expectation of death, frequently violent, manic excitement) OR gross impairment in
11 communication (e.g., largely incoherent or mute.

10 Persistent danger of severely hurting self or others (e.g., recurrent violence) OR persistent inability to maintain minimal personal hygiene OR serious suicidal
1 act with clear expectation of death.

0 Inadequate information.

APPENDIX B

GLOBAL ASSESSMENT SCALE

(GAS)

Global Assessment Scale (GAS)

Robert L. Spitzer, M.D., Miriam Gibbon, M.S.W.,
Jean Endicott, Ph.D.

Rate the subject's lowest level of functioning in the last week by selecting the lowest range which describes his functioning on a hypothetical continuum of mental health-illness. For example, a subject whose "behavior is considerably influenced by delusions" (range 21-30), should be given a rating in that range even though he has "major impairment in several areas" (range 31-40). Use intermediary levels when appropriate (e.g., 35, 58, 62). Rate actual functioning independent of whether or not subject is receiving and may be helped by medication or some other form of treatment.

Name of Patient _____ ID No. _____

Group Code _____

Admission Date _____ Date of Rating _____

Rater _____

GAS Rating: _____

100 Superior functioning in a wide range of activities, life's problems never seem to get out of hand, is sought out by others because of his warmth and integrity. No
91 Symptoms.

90 Good functioning in all areas, many interests, socially effective, generally satisfied with life. There may or may not be transient symptoms and "everyday" worries that
81 only occasionally get out of hand.

80 No more than slight impairment in functioning, varying degrees of everyday worries and problems that sometimes get out of hand. Minimal symptoms may or may not be
71 present.

70 Some mild symptoms (e.g., depressive mood and mild insomnia) OR some difficulty in several areas of functioning, but generally functioning pretty well, has some meaningful interpersonal relationships and most
61 untrained people would not consider him "sick".

60 Moderate symptoms OR generally functioning with some difficulty (e.g., few friends and flat affect, depressed mood and pathological self-doubt, euphoric mood and pressure of speech, moderately severe antisocial
59 behavior).

50 Any serious symptomatology or impairment in functioning that most clinicians would think obviously requires treatment or attention (e.g., suicidal preoccupation or gesture, severe obsessional rituals, frequent anxiety attacks, serious antisocial behavior, compulsive
41 drinking, mild but definite manic syndrome).

40 Major impairment in several areas, such as work, family relations, judgement, thinking or mood (e.g., depressed woman avoids friends, neglects family, unable to do housework), OR some impairment in reality testing or communication (e.g., speech is at times obscure,
31 illogical or irrelevant), OR single suicide attempt.

30 Unable to function in almost all areas (e.g., stays in bed all day) OR behavior is considerably influenced by either delusions or hallucinations OR serious impairment in communication (e.g., sometimes incoherent or unresponsive) or judgement (e.g., acts grossly
21 inappropriately).

20 Needs some supervision to prevent hurting self or others, or to maintain minimal personal hygiene (e.g., repeated suicide attempts, frequently violent, manic excitement, smears feces), OR gross impairment in communication
11 (e.g., largely incoherent or mute).

10 Needs constant supervision for several days to prevent hurting self or others (e.g., requires an intensive care unit with special observation by staff), makes no attempt to maintain minimal personal hygiene, or serious suicide
1 act with clear intent and expectation of death.

APPENDIX C

HEALTH-SICKNESS RATING SCALE

(HSRS)

Health Sickness Rating Scale

Definition of Scale Points	Example of Scale Points (See also the 34 ranked sample cases.)
100 At 100: An ideal state of complete functioning integration, resiliency in the face of stress, happiness and social effectiveness. (99 to 76: Degrees of "everyday" adjustment. Few of these people seek treatment.)	(Some patients who complete treatment will fall within this range, and some patients who come for and need only "situational" counseling.)
75 At 75, inhibitions, symptoms, character problems become severe enough to cause more than "everyday" discomfort. May occasionally seek treatment.	Very mild neurosis or mild addictions and behavior disorders begin here and go on down, depending on severity.
65 At 65, generally functioning pretty well but have focalized problem or more generalized lack of effectiveness without specific symptoms.	Clearly neurotic conditions (most phobias, anxiety neuroses, neurotic characters).
50 At 50, definitely needs treatment to continue work satisfactorily and has increasing difficulty in maintaining himself autonomously (even without expressed or recognized need for formal treatment). Patient may neither be in a stable unsatisfactory adjustment (where most energy is bound in the conflicts) or an unstable adjustment form which he will very likely regress.	Severe neuroses such as severe obsessive-compulsive, must be rated at 50 or lower, rarely below 35. Some compensated psychoses. Many character disorders, neurotic depressions. Most borderline schizophrenics; severe character problems. Psychotic depressions may be this high, or go all the way to 0.

25 At 25, person obviously unable to function autonomously. Needs hospital protection (or would need it if it were not for the support of the therapist). (The fact that the patient is in the hospital does not mean he must be rated at this point - he may have changed since admission or be in for a variety of reasons.) (24 to 1: Increased loss of contact with reality: need for protection of patient or others from the patient; high degree of regression.)	Most clear-cut, overt psychoses, psychotic characters, severe additions (which require care).
10 At 10, extremely difficult to make any contact with patient. Needs closed ward care. Not much chance of continued existence without care.	"Closed ward" patients such as chronic schizophrenics, excited manics, profound suicidal depressions.
0 At 0, any condition which, if unattended would quickly result in the patient's death, but not necessarily by his own hand.	Completely regressed schizophrenics, incontinent, out of contact, who require complete nursing care, tube feedings.

APPENDIX D

BRIEF PSYCHIATRIC RATING SCALE

(BPRS)

BRIEF PSYCHIATRIC RATING SCALE

	Not Present	Very Mild	Mild	Moderate	Mod. Severe	Severe	Extremely Severe
1. SOMATIC CONCERN - Degree of concern over present bodily health. Pre-occupation with physical health, fear of illness, hypochondriasis. Rate the degree to which health is perceived as a problem by the patient, whether complaints have a realistic basis or not.							
2. ANXIETY - Worry, fear, over-concern for present or future, uneasiness. Rate solely on the basis of verbal report of the patient's own subjective experiences. Do not infer anxiety from physical signs or from neurotic defense mechanisms.							
3. EMOTIONAL WITHDRAWAL - Lack of spontaneous interaction, isolation deficiency in relating to others. Deficiency in relating to others. Deficiency in relating to the interviewer and to the interviwer situation. Rate only the degree to which the patient gives the impression of failing to be in emotional contact with other people in the interview situation.							

		Not Present	Very Mild	Mild	Moderate	Mod. Severe	Severe	Extremely Severe
4.	CONCEPTUAL DISORGANIZATION - Thought processes confused, disconnected, disorganized, disrupted. Rate on the basis of integration of the verbal products of the patient; do not rate on the basis of the patient's subjective impression of his own level of functioning.	—	—	—	—	—	—	—
5.	GUILT FEELINGS - Self blame, over-concern, shame, remorse for past behavior. Rate on the basis of the patient's subjective experiences of guilt as evidenced by verbal report with appropriate affect; do not infer guilt feelings from depression, anxiety, or neurotic defenses.	—	—	—	—	—	—	—
6.	TENSION - Physical and motor manifestations of nervousness, tension, and heightened activation level. Tension should be rated solely on the basis of physical signs and motor behavior and not on the basis of subjective experiences of tension reported by the patient.	—	—	—	—	—	—	—

	Not Present	Very Mild	Mild	Moderate	Mod. Severe	Severe	Extremely Severe
7. MANNERISMS AND POSTURING - Peculiar, odd, bizarre, unnatural motor behavior (not including tic). The type of behavior which causes certain mental patients to stand out in a crowd of normal people. Rate only abnormality of movements; do not rate simple heightened motor activity here.	—	—	—	—	—	—	—
8. GRANDIOSITY - Exaggerated self-opinion, arrogance, conviction of unusual power or abilities. Rate only on the basis of patient's statements about himself or self-in-relation-to-others, not on the basis of his demeanor in the interview.	—	—	—	—	—	—	—
9. DEPRESSIVE MOOD - Despondency in mood, sadness, sorrow, pessimism. Rate only degree of despondency; do not rate on the basis of inferences concerning depression based upon general retardation and somatic complaints.	—	—	—	—	—	—	—

		Not Present	Very Mild	Mild	Moderate	Mod. Severe	Severe	Extremely Severe
10.	HOSTILITY - Animosity, contempt, belligerence, disdain for others outside the interview situation. Rate solely on the basis of the verbal report of feelings and actions of the patient toward others; do not infer hostility from neurotic defenses, anxiety, nor somatic complaints. (Rate attitude toward interviewer under "uncooperativeness.")	—	—	—	—	—	—	—
11.	SUSPICIOUSNESS - Mistrust, belief (delusional or otherwise) that others harbor malicious or discriminatory intent toward the patient. Rate on the basis of verbal report, rate only on those suspicions which are currently held whether they concern past or present circumstances.	—	—	—	—	—	—	—
12.	HALLUCINATORY BEHAVIOR - Perceptions without normal external stimulus correspondence. Rate only those experiences which are reported to have occurred within the last week and which are described as distinctly different from the thought and imagery process of normal people.	—	—	—	—	—	—	—

	Not Present	Very Mild	Mild	Moderate	Mod. Severe	Severe	Extremely Severe
13. MOTOR RETARDATION - Slowed weakened movements or speech, reduction in energy level evidenced in slowed movements. Do not rate on the basis of patient's subjective impression of own energy level.	—	—	—	—	—	—	—
14. UNCOOPERATIVENESS - Evidence of resistance, unfriendliness, guardedness, rejection of, and lack of readiness to cooperate with the interviewer. Rate only on the basis of patient's attitude and responses to the interviewer and the interview situation; do not rate on basis of reported resentment or uncooperativeness outside the interview situation.	—	—	—	—	—	—	—
15. UNUSUAL THOUGHT CONTENT - Unusual, odd, strange, or bizarre thought content. Rate here the degree to which the patient's verbalizations differ from the usual or ordinary or accepted.	—	—	—	—	—	—	—
16. BLUNTED AFFECT - Reduced emotional tone, reduction in formal intensity of feelings, apparent lack of normal feeling or involvement, flatness.	—	—	—	—	—	—	—

	Not Present	Very Mild	Mild	Moderate	Mod. Severe	Severe	Extremely Severe
17. EXCITEMENT - Heightened emotional tone, agitation, increased reaction.							
	—	—	—	—	—	—	—
18. DISORIENTATION - Confusion or lack of proper association for person, place, or time. Degree to which the patient's sense of identity is mixed up or blurred, the degree to which he is unable to understand where he is, and the degree to which he lacks spatial time dimensions such as knowing the current hour, day, month, year, season, etc.							
	—	—	—	—	—	—	—

APPENDIX E

DEFINITION OF CHRONICALLY MENTALLY ILL

DEFINITION OF CHRONICALLY MENTALLY ILL (CMI)

(Client Data Core)

AN INDIVIDUAL WHO MEETS ALL OF THE FOLLOWING CRITERIA:

I. Has clinical evidence of a psychotic disorder, severe depression, borderline personality disorder, or other serious psychiatric disorder including but not limited to Alzheimer's disease, organic brain syndrome. (This excludes adjustment, psychosexual, personality, and other disorders that are situational or characterological in nature.)

AND

II. Has had a severe mental disorder for at least two years in duration which can be substantiated by clear and convincing clinical evidence.

AND

III. Has at least substantial or serious impairment in personal maintenance, social relations or occupational functioning, as measured by the following:

A DSM III rating of five or more or a Spitzer, Gibbon, Endicott Scale rating (OMHIS) of 50 or less.

At least three (3) of the following functional criteria:

Is unemployed, is employed in a sheltered setting, or has no marketable work skills;

Requires public financial assistance for out-of hospital maintenance and/or is unable to procure such assistance without help;

Shows severe inability to establish or maintain a personal social support system (has no family, close friends or group affiliations, is isolative or is highly transient);

Has a documented history of failure to maintain medication regimen;

Lacks daily living skills (clothing care, care of immediate living environment, transportation, cooking, personal health and hygiene, and personal finances).

APPENDIX F

SUBJECT INFORMATION SHEET AND CONSENT FORM

INFORMATION AND CONSENT FORM
FOR RESEARCH PROJECT

The Edwin Fair Community Mental Health Center is encouraging your participation in a research project designed to determine the reliability and validity of a rating scale our therapists use to rate how well a client is functioning. Over the next few weeks, your psychiatrist and regular therapist will be helping us compare two different rating scales by rating how well their clients are doing on these scales. The ratings made on the rating scales will be kept confidential, and the study will be conducted in accordance with the Client's Bill of Rights, which you received when you became a client. Your participation will be limited to your next visit with your therapist, _____ _____, and your psychiatrist, _____ _____.

If you specifically request it below, information about your individual scores or about the results of the overall research study will be provided to you by your therapist.

If, at any time, you decide not to participate in the research project, you are free to withdraw by telling your therapist, psychiatrist, or the project director whose name, address and phone number are listed below. There is no penalty for refusal to participate. If, as a result of your participation you have any adverse reaction or psychological problem, you may contact your psychiatrist, therapist or the project director immediately.

For any further information regarding this project, please contact the project director, Jack P. Schaefer, M.S., Edwin Fair Community Health Center, 712 Devon Road, Stillwater, OK 74074, (405) 372-1250; or his authorized representative, Mrs. Lela Holzer, Secretary, at the same address and phone number. You may also contact Terry Maciula, University Research Services, 001 Life Sciences East, Oklahoma State University, Stillwater, OK 74078 (405) 744-5700.

STATEMENT OF CONSENT TO PARTICIPATE

I, _____, have read the above information and understand it. I affirm that I am 18 years of age or older and I understand that participation in the project is voluntary. I understand that there is no penalty for refusal to participate. I have been given the opportunity to ask questions and they have been answered to my satisfaction.

I have circled the correct response (do) or (do not) on each of the following statements as they apply to me:

I (do) (do not) agree to let my therapist and psychiatrist rate how well I am functioning as part of this research project.

I (do) (do not) want feedback about my individual scores.

I (do) (do not) agree to have my individual scores placed in my client file.

I (do) (do not) want to know the overall results of the research project.

I agree to participate in this research project entitled "A Study of the Reliability and Validity of the Global Assessment of Functioning Scale." I sign this consent form freely and voluntarily. A copy of this form has been given to me.

Signature of Client:_____, Date:_____,

Time:_____ (a.m./p.m.) Witness:_____,

Date:_____

I certify that I have personally explained all elements of this form to the subject or his/her representative before requesting the subject or his/her representative to sign it.

Signed:_____
(Project Director or his/her authorized representative)

VITA

Jack P. Schaefer

Candidate for the Degree of

Doctor of Philosophy

Thesis: A STUDY OF THE RELIABILITY AND VALIDITY OF THE GLOBAL ASSESSMENT OF FUNCTIONING SCALE

Major Field: Psychology

Biographical:

Personal Data: Born in Vinita, Oklahoma, November 24, 1946, the son of Harry P. and Jane Schaefer.

Education: Graduated from Pryor High School, Pryor, Oklahoma, in May, 1965; received Bachelor of Science Degree in Psychology from Oklahoma State University in May, 1969; received the Master of Science degree at Oklahoma State University in May, 1980; received the Doctor of Philosophy from Oklahoma State University in July, 1991.

Professional Experience: Director, Edwin Fair Community Mental Health Center, Payne County Center, March, 1980, to present.