2004

Correspondence of Self- and Observer-rated Depression Using the BDI-II and BDI-II-O

Linda A. Longan
Philadelphia College of Osteopathic Medicine, llongan@comcast.net

Philadelphia College of Osteopathic Medicine

Department of Psychology

CORRESPONDENCE OF SELF- AND OBSERVER-RATED DEPRESSION USING THE BDI-II AND BDI-II-O

By Linda A. Longan

Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Psychology

October 2004

PHILADELPHIA COLLEGE OF OSTEOPATHIC MEDICINE

DEPARTMENT OF PSYCHOLOGY

Dissertation Approval

This is to certify that the dissertation presented to us by Linda A. Longan on the 26th day of July 2004, in partial fulfillment of the requirements for the degree of Doctor of Psychology, has been examined and is accepted in both scholarship and literary quality.

Committee Members' Signatures:

Arthur Freeman, Ed.D., ABPP, Chairperson

Michael Ascher, Ph.D.

Donna Martin, Psy.D.

Arthur Freeman, Ed.D., ABPP, Chair, Department of Psychology

Acknowledgement

My gratitude extends widely and deeply to those who supported me in this effort.

I am indebted foremost to my husband, who selflessly, and always cheerfully, endured my many hours of absence from home, even when I was physically present. He relentlessly carried the ball and kept normal life afloat with enormous patience and tolerance.

I benefited greatly from the assistance of my committee members and mentors. I thank Art Freeman who provided excellent guidance and kept me focused, motivated, and moving forward. I thank Donna Martin for her calm and generous spirit, and her insight into the process which was required to reach this point. I thank Michael Ascher for sharing his unique perspective on my topic, and Robert DiTomasso for his valuable contributions during the analysis of the data.

I was fortunate to have the consistent cooperation and support of the staff and patients at The Keystone Center, whose willing participation made this study possible.

I am grateful to Aaron Beck for providing encouragement to pursue this research, and to The Psychological Corporation for granting permission to adapt the BDI-II for purposes of this study.

Abstract

This study describes the development, reliability, and validity of a new observer-rated version of the Beck Depression Inventory – Second Edition (BDI-II). This measure, called the Beck Depression Inventory – II – Observer (BDI-II-O), is identical in form, content, and scoring to the BDI-II, and allows informants to report on the depressive symptoms of others. The informants for this initial study were clinicians; however, the BDI-II-O is designed for use by a wider range of informants. A group of four clinicians completed the BDI-II-O on each of 36 adult psychiatric patients in an intensive outpatient/partial hospitalization program. The patients completed the BDI-II, and scores were compared between the patient and clinician groups, as well as between clinicians. The BDI-II-O demonstrated high internal consistency (r = .96) and test-retest reliability (r = .88). Results from an exploratory cluster analyses indicated that a two-cluster solution optimally summarized the data for the BDI-II-O, although a one cluster solution was salient for the BDI-II. Results showed a moderate level of correlation between patient and clinician ratings (r = .59) a high level of interrater reliability between the four clinicians (ICC = .88). Overall results supported the preliminary reliability and validity of the BDI-II-O as an observer rated measure of depressive severity, and suggested that further study of the BDI-II-O is warranted.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Depression is one of the most prevalent mental health problems in the United States (Kessler et al., 1994). Two major epidemiological studies, the Epidemiologic Catchment Area (ECA) study and the National Comorbidity Survey (NCS), have measured the incidence and prevalence of selected mental health disorders, including depression. The ECA study was sponsored by the National Institute of Mental Health in the early 1980's and utilized a geographically dispersed, community-based sample of 20,000 persons ages 18 and older. Results from this study showed overall 1-year and lifetime prevalence rates of 2.7% and 4.9%, respectively, for major depression within the total sample (Kaelber, Moul, & Farmer, 1995). Across age groups, people 30-44 had the highest 1-year and lifetime prevalence rates at 3.9% and 7.5%, respectively. People age 65 and older had the lowest reported 1-year (0.9%) and lifetime (1.4%) prevalence rates. Gender differences in prevalence rates were evident, with women showing over twice the rate of major depression as men. Among ethnic groups, Caucasians had higher prevalence rates than African Americans and Hispanics.

The National Comorbidity Survey, conducted by Kessler, McGonagle, Swartz, Blazer, and Nelson (1993), was a probability-based sample of adolescents and adults between the ages of 15 and 54. This study was designed to be more generalizable to the entire United States in comparison with the ECA study that had limited geographical coverage (i.e., five states). The overall results of the NCS study showed substantially higher rates of major depression than did the ECA study. Within the total sample, the NCS study found 1-year and lifetime prevalence rates of 8.6% and 14.9%, respectively,

for major depression. These rates were nearly three times greater than those found in the ECA study. Discrepancies between the two studies have been attributed to differences in methodological design, that is, variances associated with the assessment instruments and diagnostic criteria used, as well as the age range of the sample. Results by gender in the NCS study were consistent with ECA study and confirmed the skew of major depression toward women.

The ECA and the NCS studies also measured the prevalence of depressive episodes that were below the threshold for meeting diagnostic criteria for major depression. The ECA study showed 1-year and lifetime prevalence rates of 3.7% and 6.3%, respectively, for any major depressive episode, yet the NCS showed rates of 10.3% and 17.1% on the same measures. It is evident from these results that the disparities between the two studies on measures of major depression were also present, and of the same relative magnitude, for measures of depressive episodes.

Depression is a leading risk factor for suicide. Chiles and Strosahl (1995) reported that suicide ranks as the eighth leading cause of death in the general population. Among 18 to 24 year olds, it is the third leading cause of death. The suicide rate in the over 65 population is approximately double that of the 18- to 24-year-old population.

The public health burden of major depression is quite significant. The burden of various physical and mental illnesses on health and productivity (i.e., mortality and disability) throughout the world was estimated by the Global Burden of Disease study conducted by the World Health Organization, the World Bank, and Harvard University (1996). Researchers for this study developed a single measure, the Disability Adjusted Life Years (DALYs), to facilitate comparison of burden across many different illnesses.

The DALYs measure equivalized physical and mental disorders so that lost years of healthy life could be estimated regardless of whether or not the loss was due to death or disability. In established market economies such as the United States, major depression was found to be the leading source of disease burden within the mental illness category, accounting for 43.8% of the total burden attributable to mental disorders. Schizophrenia was a distant second source at 15.0%, followed by bipolar disorder, obsessive-compulsive disorder, panic disorder, and other mental illnesses. In addition, major depression ranked second only to heart disease as a leading cause of healthy life years lost in the combined category of physical and mental illnesses. Over 80% of healthy years lost to depression are reported to occur between the ages of 15 to 44, with women losing almost twice the number of healthy years as men. These findings support the opinion of Hays, Wells, Sherbourne, Rogers, and Spitzer (1995) that depression is a serious illness, associated with significant disability and decreased quality of life. As such, the psychological assessment of depression becomes an important area for continuing investigation.

In broad terms, psychological assessment refers to the evaluation of a patient's mental health status through the use of psychological tests. For purposes of this study, we have adopted the more complete definition of psychological assessment as provided by the National Council of Schools and Programs in Professional Psychology. The NCSPP states that "assessment is an ongoing, interactive, and inclusive process that serves to describe, conceptualize, characterize, and predict relevant aspects of a client". As such, psychological assessment serves the purposes of understanding the patient, identifying the most important problems and issues that need to be addressed, and developing

targeted treatment plans to address these problems. Accurate assessment is widely acknowledged as a necessary prerequisite for delivering treatment that is both appropriate and effective (Hesselbrock, Easton, Buchotz, Schuckit, & Hesselbrock, 1999).

Treatment planning has received increased attention and emphasis during recent years (Maruish, 1999). Reasons for this trend have included on-going efforts to make psychotherapy more efficient and cost effective, the growing influence of third parties such as insurance companies and the federal government, and movement away from forms of psychotherapy that are neither goal oriented nor time limited.

Treatment planning is defined as the process of setting treatment goals and developing appropriate strategies to achieve those goals. According to Maruish, the assumptions underlying the treatment planning process include: (1) patients are experiencing problems that have been identified by themselves or others; (2) patients have some degree of internal or external motivation to eliminate or reduce the identified problems; (3) the goals of treatment are tied to the identified problem, either directly or indirectly; (4) the goals are measurable, achievable, and developed in collaboration with the patient; (5) the goals are prioritized; (6) progress toward goals can be tracked against an expected path of improvement based on the clinician's experience or, preferably, through objective data; and (7) departures from the expected path of improvement necessitate a modification of the treatment plan, followed by subsequent monitoring to determine the effectiveness of the modifications. We would also add to this list the assumption that diagnosis is a necessary but not sufficient condition for treatment and, in order for treatment to be effective there must be substantial agreement between the clinician and patient on the goals and tasks of treatment.

The use of psychological assessment instruments to develop and monitor treatment can provide useful feedback to both the clinician and the patient. The sharing of psychological assessment results with the patient is not only mandated (American Psychological Association, 1972), but is also viewed to be a therapeutic intervention in and of itself. Finn and Tonsager (1992) summarized the potential benefits of providing clients with feedback on assessment results as: increased feelings of self-esteem and hope, reduced symptomatology and feelings of isolation, increased self-understanding, and increased motivation to be more actively involved in treatment. Finn and Tonsager also noted that the feedback process provides a model for relationships that can result in increased mutual respect.

Psychological assessment as an outcome measure assists the clinician in making decisions about treatment termination. From the patient's standpoint, outcome measurement can provide an objective view of progress, and point out the need for further treatment or the readiness for termination.

In summary, psychological assessment aids in case conceptualization, in formulating initial and on-going goals for therapy, in understanding the particular factors that contribute to the onset or maintenance of depressive symptoms, in identifying problems that increase the risk of relapse or recurrence, and in monitoring progress.

Given the prevalence and disabling nature of depression, there has been much interest in developing reliable and valid assessment instruments for depression. Accurate assessment of depression is especially important because, for many patients, depression is a recurrent and chronic disorder. Rating scales for depression have played a critical role in both the clinical assessment of depressive symptoms and in depression research (Enns,

Larson, & Cox, 2000). The two most widely used depression rating scales are the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961; Beck, Steer, and Brown, 1996) and the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960, 1967), both of which have well-demonstrated reliability and validity. The BDI is a self-report instrument that is completed by patients, and the HRSD is a clinician rated instrument that is completed by utilizing a semi-structured interview process. Although both instruments seek to quantify depressive severity, they are different in form and content, making comparisons between the two difficult at best. It is this lack of comparability between the BDI and existing clinician rated instruments that provided the impetus for the present study. There is currently no parallel version of the BDI with which observers can rate the depression of others, and against which the ratings of patients can be compared.

An observer rated version of the BDI was developed to address the inconsistencies between the BDI and the existing other-rated measures of depression. Although this new instrument, the BDI-II-O, could eventually be used by a variety of informants, the purpose of this initial research is to examine the BDI-II-O's viability as a clinician rated instrument. The underlying assumption of this study is that the use of parallel instruments for self-reporters and clinicians will lead to a more consensual validation of depression based on common constructs and criteria and, in turn, to better treatment planning and outcome. It is under this premise that the BDI-II-O was developed, with the expectation that it would exhibit excellent psychometric properties and the same strong construct validity as that evidenced by the BDI-II.

Chapter 2
Construct of Depression

A major distinction that separates psychological assessment from other forms of scientific assessment, such as that found in the physical sciences, resides in the object of measurement: it is usually a hypothetical construct rather than something tangible. (Derogatis & Lynn, 1999). Depression as a construct is generally accepted to be a multi-dimensional rather than a unitary concept, comprising several different domains of depressive symptomatology. As Katz, Shaw, Vallis, and Kaiser (1995) pointed out "depression has always been identified as a varied, diverse problem with affective, physiological, behavioral, and cognitive symptoms associated with it" (p. 71).

Many depression-related constructs have been identified in the literature. Various authors have proposed different conceptualizations and models of depression, including Lewinsohn's behavioral model, Becker's interpersonal skills model, Rehm's self-control model, Nezu's problem solving model, and Beck's cognitive model. The cognitive model of depression is the appropriate theoretical backdrop for this study due to its relevance to the topic under investigation. In addition, it is an empirically validated model based on a well-defined construct with an extensive body of literature.

Beck's theory of depression is based on the notion of a cognitive triad – a concept that defines and helps to explain how people view the self, the world, and the future (Beck, Rush, Shaw, & Emery, 1979). Originally an outgrowth of Beck's research on depression, the triad concept has subsequently been extended to apply to other disorders such as anxiety, panic, and obsessive compulsive disorders. This "transportability" of the concept is known as the cognitive-content specificity hypothesis of the model. This

hypothesis proposes that every psychological disorder has a distinctive cognitive profile which is reflected at all levels of cognitive functioning, allowing the triad concept to be applied across a range of disorders (Beck, 1987).

The depressive cognitive triad consists of a negative view of the self, the world, and the future. According to Beck (1979), depression is caused by the belief in one's own inability to be effective, a view of the world as unresponsive or incapable of helping, and a prediction of the future as inevitably negative and hopeless. The negative view of the self encompasses guilt, self-blame, low self-esteem, and feelings of incompetence. The negative view of the world encompasses a sense that problems are overwhelming and the universe is a hostile and rejecting place. The negative view of the future encompasses a belief that nothing will ever change and the future is doomed to be as dreadful as the present. Taken together, these views compel the depressed person to expect failure, dissatisfaction, and indefinite continuation of the negative situation. According to Beck, depressed patients consistently present with a pattern of thoughts in which they view themselves as incompetent or unworthy, other people as hostile and rejecting, and the future as negative and painful. When active, these collective viewpoints contribute to depressive symptoms and impairments in functioning.

Examining the differences between the depressive cognitive triad and the anxiety cognitive triad may help to clarify further an understanding of the triad model. Although the depressive triad revolves around the cognitive themes of personal loss and failure in the interpersonal and achievement domains, the anxiety triad speaks to a theme of physical or psychological threat of danger to self or significant other, with an increased sense of personal vulnerability. In depression, thoughts take the form of negative self-

statements; however, in anxiety, thoughts take the form of "what if" questions involving possible harm and danger. In depression there is elevated cognitive processing of negative (and exclusion of positive) self-referent information; appraisals are pervasive, global, absolute, and exclusive. Responsiveness to the outside world may be limited due to an increase in self-focus. In anxiety, there is selective processing of threat cues with an overestimation of vulnerability; appraisals are selective, tentative, and specific to a feared situation. Increased self-focus is present in anxiety as well, but this reflects attempts to control, rather than withdraw, from the outside world. Depression and anxiety, although sharing the dimension of negative affectivity, are viewed as separate and distinct constructs in the cognitive model – a notion that is broadly supported in the literature.

Cognitive therapy for depression is effective for short-term symptom reduction and is superior to drug therapies in relapse prevention (Hollon, DeRubeis, & Seligman, 1992). It is active, directive, structured, and psychoeducational. The assumptions underlying cognitive therapy for depression are that: (1) negative processing of information is responsible for the immediate onset, maintenance, and exacerbation of depression; (2) cognitions can be self-monitored by the client and communicated (i.e., they are not unconscious); and (3) changes in cognitions lead to changes in behavior.

There is a plethora of depression rating instruments, each based on the author's particular construct of depression. All are designed to assess the variety of symptoms observed to be present in depressive illness. However, each instrument tends to emphasize some categories of symptoms over others. Although most depression rating instruments show positive correlation with one another, they do not begin to approach unity. The following section will provide a comprehensive review of the most widely

used depression rating instruments.

Measurement and Assessment of Depression

Structured Diagnostic Interviews

Structured diagnostic interviews use standardized content, format, and ordering of questions to be asked, along with algorithms for arriving at diagnostic conclusions. There are five such instruments currently in use; the two most frequently used instruments will be reviewed: the Schedule of Affective Disorders and Schizophrenia, and the Structured Clinical Interview for DSM-IV Axis I Disorders.

Schedule of Affective Disorders and Schizophrenia (SADS). The Schedule of Affective Disorders and Schizophrenia (Endicott & Spitzer, 1978) was an outgrowth of the multi-centered National Institute of Health Collaborative Study of the Psychobiology of Depression. It represented the first attempt to develop a structured interview for clinical research that would enhance the consistency of diagnostic evaluation across multiple geographic locations and raters, thus helping to provide homogeneous groups of subjects for research. In conjunction with the development of the SADS, Spitzer and Endicott, along with their associate, Robins, created the companion Research Diagnostic Criteria (Spitzer, Endicott, & Robins, 1978) which was the forerunner of the DSM-III. The RDC is a manual consisting of explicit operational definitions for 25 psychiatric disorders. Prior to the development of the SADS and the RDC, research was conducted

largely through unstructured clinical interviews without the benefit of reliable diagnostic categories. Through the introduction of an instrument with an organized progression of questions based on established criteria, the SADS was a groundbreaking, systematic approach to reducing two major sources of clinician unreliability in research (i.e., information variance and criterion variance). Therefore, the SADS represented an important evolution in the reliability of psychiatric assessment.

The SADS assesses a broad range of psychiatric illnesses including major depressive disorder, dysthymia, schizophrenia, anxiety disorders, and to a lesser extent, personality disorders. There are multiple versions of the SADS that differ primarily in the time period being assessed – one measures current status (SADS), another measures lifetime status (SAD-L), and a third measures change from a former state (SADS-C).

Ratings are made on a 6-point Likert scale and the instrument produces the following eight summary scales: (1) Depressive Mood and Ideation, (2) Endogenous Features, (3) Depressive-associated Features, (4) Suicidal Ideation and Behavior, (5) Anxiety, (6) Manic Syndrome, (7) Delusions-hallucinations, and (8) Formal Thought Disorder. The first four scales are considered, collectively, to measure depression.

Interrater reliability in the case of diagnostic instruments refers to the extent of agreement between multiple raters relative to the presence or absence of a symptom or disorder. It is measured by either intraclass correlation coefficients (ICC's) or by Kappa coefficients, depending on the size of the sample and nature of the measurements. Endicott and Spitzer (1978) reported excellent interrater reliability for the SADS depression scales as follows: Depressive Mood and Ideation, .95; Endogenous Features, .96; Depressive-associated Features, .96; and Suicidal Ideation and Behavior, .97. The

internal consistency (Cronbach, 1951) of the summary scales was also high, ranging from .79 on Suicidal Ideation and Behavior to .87 on Depressive Mood and Ideation. Test-retest reliability ranged from .78 for Depressive Mood and Ideation to .88 for Depressive-associated Features. The depression scales were reported to be moderately intercorrelated to highly intercorrelated, in the range of .40 to .90.

Diagnostic reliability of the RDC was high, with a kappa coefficient of .90 for the diagnosis of major depressive disorder, and a kappa coefficient of .81 for minor depressive disorder (Spitzer et al., 1978). Simon, Endicott, and Nee (1987) reported an even higher kappa coefficient of .99 for major depressive disorder. The Kappa statistic is equivalent to the intraclass correlation coefficient when large samples are used (i.e., more than 25-30 observations). The Kappa coefficient is interpreted as follows: a value of .85 or greater suggests excellent agreement between raters, .84-.70 suggests good agreement, .69-.40 indicates fair agreement, and below .40 indicates poor agreement (Segal, Hersen, and Van Hasselt, 1994).

The SADS has a number of weaknesses, particularly for clinical practice. It is time-consuming, and hence costly to administer, taking about 1½ to 2 hours to complete. In addition, effective use of the instrument requires a substantial amount training and experience on the part of the clinician (Beckham & Leber, 1995). Furthermore, it uses terminology and a diagnostic classification system that is now dated.

Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I). The Structured Clinical Interview for DSM-IV Axis I Disorders (First, Gibbon, Spitzer, & Williams, 1996; First, Spitzer, Gibbon, & Williams, 1997) differs from the SADS

because the diagnoses are based on DSM-IV (APA, 1994), rather than RDC, criteria. Although both instruments cover a wide spectrum of clinical disorders, the SCID and its versions are superior in this regard, having more breadth and depth of coverage. There are research and clinician versions of the SCID. Research versions include: (1) the SCID-I/P which is the most comprehensive version and is designed for subjects already identified as psychiatric patients; (2) the SCID-I/P with Psychotic Screen, an abbreviated version of the SCID-I/P, is used when psychotic disorders are expected to be rare or when a screen for psychotic disorders would suffice, and (3) the SCID-I/NP is a non-patient version for research in community settings (e.g., primary care, general medical). The clinician version (SCID-CV) is the briefest of the versions, covering only those DSM-IV disorders commonly seen in clinical practice. Unlike the research versions, the SCID-CV does not cover disorder subtypes and severity/course specifiers. The full SCID/IP contains nine diagnostic modules: mood episodes, psychotic symptoms, psychotic disorder differential, mood disorders differential, substance abuse disorders, anxiety disorders, somatoform disorders, eating disorders, and adjustment disorders. An optional module contains some appendixed disorders of the DSM-IV such as minor depressive disorder. All versions permit the assessment of both current and lifetime mood disorders. In addition to the SCID-I instruments for Axis I disorders, there is a SCID-II for Axis II personality disorders. Finally, there are computerized versions of the SCID-I and SCID-II for both clinicians and self-reporters, and a computerized self-report screener for Axis I disorders. A detailed history, rationale, and description of the development of the SCID can be found in Spitzer, Williams, Gibbon, and First (1992).

All versions of the SCID include open-ended questions, as well as a skip structure that guides the interview through alternate decision trees depending on responses to the last question. For each item, a judgement is made about whether or not the symptom or criterion is present, absent, or subthreshold, and whether or not adequate information is available with which to rate the item.

A number of studies have evaluated the interrater reliability of the original versions of the SCID which are based on DSM-III-R criteria (Riskind, Beck, Berchick, Brown, & Steer, 1987; Segal, Williams, Gibbon, & First, 1994; Skre, Onstand, Torgersen, & Kinglen, 1991; Williams et al., 1992). For major depressive disorder, Riskind et al. (1987) found a Kappa coefficient of .72 and a simple agreement rate of 82% among pairs of raters consisting of an initial interviewer and a post hoc videotape rater for outpatients with a mean age of 34 years. Segal et al. (1994), using a similar method in a study of interrater reliability in a sample of older adults, found a Kappa coefficient of .70 and a simple agreement rate of 85% for major depression. Skre et al. (1991) reported a Kappa of .93 for major depression in a study of Axis I disorders among adult Norwegian twins. The original developers of the SCID, Williams et al. (1992) performed the most comprehensive study of interrater reliability. Their study used a large sample consisting of 592 subjects at four patient sites and two non-patient sites in the United States and one patient site in Germany. Pairs of mental health professionals, consisting of an initial interviewer and an audiotape rater, independently evaluated the same individual. Overall reliability for current disorders was fair to good in the patient sample with a Kappa of .61, but poor in the non-patient sample with a Kappa of .37. For

major depressive disorder, the Kappas were .64 and .42 for the patient and non-patient samples, respectively, mirroring the overall results.

The above studies compared ratings between two separate evaluations using the SCID; the overall results suggested fair to good interrater reliability. There have been few studies comparing the SCID to unstructured clinical diagnoses. Steiner, Tebes, Sledge, and Walker (1995) addressed this issue by comparing diagnoses derived from the SCID-I/P with Psychotic Screen and an unstructured clinical interview on the same patient. These authors found poor interrater reliability for most diagnostic categories including major depression, which showed a Kappa of .34 and a simple agreement rate of 74%.

Although the SCID is unmatched in breadth of diagnostic coverage and adherence to the DSM-IV framework, it has a few drawbacks: (1) extensive training and diagnostic experience is required for effective administration, (2) administration time is lengthy, taking a proficient interviewer about 60 minutes, and (3) the patient may find it tedious.

Self-Administered Diagnosis Instruments

Inventory to Diagnose Depression (IDD). The Inventory to Diagnose Depression (Zimmerman, Coryell, Corenthal, & Wilson, 1986) is a 22-item self-administered instrument designed to diagnose major depressive disorder, assessing the severity of depressive symptoms. The instrument was developed based on the criteria for major depressive disorder as defined in the DSM-III (APA, 1980), which distinguished it from earlier instruments. It has not been updated for DSM-IV criteria. Each item is rated on a

5-point scale, ranging from 0 (no disturbance) to 4 (severe disturbance). A score of 2 or greater signifies the existence of a symptom. Ratings are based on the presence of symptoms in the current week. A unique feature of the IDD is that it asks the respondent to indicate whether or not a symptom has been present for more than two weeks or less than two weeks; with this feature the symptom duration is captured. Symptoms assessed by the IDD include: low mood, decreased energy, psychomotor agitation or retardation, decreased interest in usual activities, decreased pleasure in usual activities, decreased libido, guilt, worthlessness, suicidal/death thoughts, decreased concentration, indecisiveness, decreased appetite, weight loss, increased appetite, weight gain, insomnia, hypersomnia, anxiety, hopelessness, irritability, and somatic complaints. Only 20 of the 22 IDD items are needed to diagnose major depressive disorder. The two excluded items, anxiety and somatic complaints, are used only when computing the severity score. The instrument takes 15 minutes to complete.

The psychometric properties of the IDD are good, based on the developers' study of the instrument with psychiatric patients. Test-retest reliability was reported to be .98 for 16 randomly selected patients over a two-day period. Internal consistency was reported to be .92. The Kappa coefficient for agreement between self-report and clinical diagnosis according the DSM-III criteria was .58, indicating moderate agreement.

Clinician Rated Depression Instruments

*A*t least nine clinician rated depression scales exist (Nezu, Ronan, Meadows, & McClure, 2000). This study will review the most well known and most extensively used

instruments, the Hamilton Rating Scale for Depression and the Brief Psychiatric Rating Scale.

Hamilton Rating Scale for Depression (HRSD). The Hamilton Rating Scale for Depression (Hamilton, 1960) was one of the first semi-structured interview measures developed for the clinical evaluation of depression severity in adults. It was used in the National Institute of Mental Health's Early Clinical Drug Evaluation program (ECDEU) as part of an effort to the ensure uniformity of assessment measures in psychotropic drug evaluations. Subsequently, it has become the standard outcome measure used by pharmaceutical companies in clinical trials of new antidepressant medications. (Kobak & Reynolds, 2000). The HRSD was also the primary outcome measure in the National Institute of Mental Health's Treatment of Depression Collaborative Research Program which evaluated the relative efficacy of antidepressant medication and psychotherapy for the treatment of depression (Elkin et al., 1989). Given its long history and strong psychometric properties, the HRSD is often used as the standard or criterion measure against which other depression rating scales are validated.

The original version of the HRSD consists of 17 items which are each rated on either a 5-point (0-4) or 3-point (0-2) scale. The 5-point scale ratings are: 0 = absent, 1 = doubtful to mild, 2 = mild to moderate, 3 = moderate to severe, 4 = very severe. The 3-point scale ratings are: 0 = absent, 1 = probable or mild, and 2 = definite (Kobak & Reynolds, 2000). The 3-point scale is generally used for items deemed difficult or impossible to quantify with precision, for example insight, insomnia, and somatic symptoms. Several of the items, such as Depressed Mood and Retardation contain

multiple symptoms that are considered in determining the final rating for that item. Total scores may range from 0-52. A score of 6 or below indicates normal, nondepressed functioning; scores of 7-17 indicate mild depression; scores of 18-24 suggest moderate depression; and scores of 25 or above reflect severe depression. Ratings are based on symptoms during the previous week, which is now inconsistent with DSM-IV criteria that specify a 2-week time frame.

Although most researchers use the original 17-item version of the HRSD (Katz, Shaw, Vallis, & Kaiser, 1995), another version encompassing 23 items is also available. This modification from the original version added cognitive items that addressed symptoms of hopelessness, helplessness-pessimism, and worthlessness, in addition to detachment, difficulty making decisions, and hypersomnia (Reynolds & Kobak, 1995). Unfortunately, many research studies do not specify whether or not the 17-item version or the 23-item version was used, making comparisons across studies difficult.

According to Reynolds & Kobak (1995), the relative lack of standardized administration instructions and scoring criteria for the HRSD represents a weakness of the instrument. These authors point out that there are no standardized probe questions to elicit information from patients, nor are there specific guidelines given for determining each item's rating.

Hamilton (personal communication to E. Beckham, February, 1984) stated that "the rater should not hesitate to record if the symptom is severe, even though he recognizes that other patients are even worse." According to Katz, Shaw, Vallis, and Kaiser (1995), this approach "introduces ambiguity into the ratings, since a given rating may represent different things for two different patients." To increase reliability of the

ratings, Hamilton suggested that each patient should be rated by two different interviewers and their scores be averaged. This solution, however, appears to lack cost effectiveness.

The HRSD was designed for use by experienced clinicians. Interrater reliability is generally considered adequate for clinicians trained at the same facility due to the fact that common guidelines and procedures are generally employed (Kobak & Reynolds, 2000). For example, Hedlund and Vieweg (1979), in their review of nine studies, reported intraclass correlation coefficients of .84 or above, with the exception of one study which found a coefficient of .52. However, as noted by Hooijer et al. (1991), interrater reliability between raters at different sites has been less impressive and harder to achieve because of divergence in the guidelines used and differences in clinician training and experience. Cicchetti and Prusoff (1983), in their study of interrater reliability, found low levels of reliability for individual items, with 14 of 22 items demonstrating intraclass correlation coefficients of less than .40.

Brief Psychiatric Rating Scale (BPRS). The Brief Psychiatric Rating Scale (Overall and Gorman, 1962) is a clinician rated scale for assessing symptom severity in schizophrenia and mood disorders. It was originally developed to fill the need for a tool to measure patient change during a time when new medications were being introduced to treat schizophrenia. The original 16-item scale included the following symptoms: somatic concern, anxiety, emotional withdrawal, conceptual disorganization, guilt feelings, tension, mannerisms and posturing, grandiosity, depressive mood, hostility, suspiciousness, hallucinatory behaviors, motor retardation, uncooperativeness, unusual

thought content, and blunted affect. Two more symptoms, excitement and disorientation were added later, and an expanded 24-item version was introduced in 1995; this version added the symptoms of bizarre behavior, suicidality, self-neglect, motor hyperactivity, distractibility, and elevated mood. Symptoms are rated on a 7-point scale, ranging from "not present" to "extremely severe". On the 18-item version, ratings of 6 items are based on observed behavior during the interview, and the remaining 12 items, including depressed mood, are rated based on the content of the interview. Depressed mood is rated on the basis of expression of discouragement, pessimism, sadness, hopelessness, helplessness, and gloomy theme. Other symptoms commonly associated with depression such as motor retardation, guilt, and somatic complaints are rated as separate items.

The BPRS was designed to be used by trained clinicians in conjunction with an informal clinical interview. There are no mandatory questions for this instrument, but sample questions are provided for each item.

According to Faustman and Overall (1999), a review of available studies on the BPRS showed fairly high interrater reliability, varying around .85. There is no normative data for the instrument because it was developed for the purpose of assessing treatment-related change rather than making clinical diagnoses. The primary constraint in using the BHRS for assessing depression is that fewer than half of the items are relevant to depression; the others are related to thought disorder (Rabkin & Klein, 1987). The HRSD is probably a more reasonable choice for clinicians who are interested solely in assessing depression. However, for those interested in assessing a broader range of psychopathology, or when comorbid conditions are suspected, the BHRS has good clinical utility.

Self-Report Depression Instruments

There are at least 25 self-report measures of depression severity identified in the literature. Of these, 17 were developed specifically to measure depressive symptoms, mood, and/or severity in adult populations. The remaining 8 self-report measures were created for use with special populations (e.g., children, schizophrenia) or in special settings such as primary care. Only a few of these 25 instruments have achieved widespread acceptance. This study will review four of the most commonly cited self-report measures of depression: the Zung Self-Rating Depression Scale, the Carroll Rating Scale for Depression, the Hamilton Depression Inventory, and the Beck Depression Inventory.

Zung Self-Rating Depression Scale (ZSDS). The Zung Self-Rating Depression Scale (Zung, 1965) is a 20-item measure of depressive symptoms which includes various affective (2 items), somatic (8 items), and psychological (10 items) components of depression. Each item is rated on a four point scale and respondents are asked to pick the statement which best describes the amount of time each statement applies to them over the previous several days (i.e., a little, some, a good part, and most of the time). Half of the items are reverse keyed and scored in such a way that a negative answer signifies the presence of a symptom, for example, “morning is when I feel best”. The remaining items are symptomatically positive, for example, “I feel downhearted and blue”. The total possible score is 80. Cutoff scores are: normal = less than 50, mild depression = 50-59, moderate to marked depression = 60-69, and severe depression = 70 or greater. In

addition to the summed score, a depression index may be derived by dividing the summed score by the total possible score of 80. The ZSDS does not include some characteristic symptoms of depression, namely, increased appetite, weight gain, increased sleep, concentration difficulties, and psychomotor retardation. As such, it does not cover the full spectrum of diagnostic criteria for major depressive disorder. The instrument takes about 10-15 minutes to complete.

Few studies have assessed the reliability of the ZSDS, despite its long history of clinical and research use. Gabrys and Peters (1985) reported high internal consistency with a coefficient alpha of .88 for depressed inpatients, but Knight, Waal-Manning, and Spears (1983) reported a coefficient alpha of .79. Some authors (Rablin & Klein, 1987; Schotte, Maes, Cluydts, & Cosyns, 1996) have questioned the factor structure, construct validity, and discriminant validity of the ZSDS. Based on these concerns, and in combination with a relative paucity of psychometric information, the ZSDS may have less merit than other available instruments for assessing depression.

Carroll Rating Scale for Depression (CRSD). The Carroll Rating Scale for Depression (Carroll, Feinberg, Smouse, Rawson, & Greden, 1981) comprises 52 statements that are primarily behavioral and somatic in content. Items are presented in a self-descriptive format, such as “I feel in good spirits”, and are responded to with a “yes” or “no”, based on feelings over the previous few days. Total scores range from 0 to 52, with scores greater than 10 indicating clinically significant depression.

The CRSD was developed with the objective of providing a self-report instrument that corresponds with the clinician-rated HRSD. As such, the 52 statements can be

grouped into the 17 items on the original HRSD: depression, guilt, suicide, initial insomnia, middle insomnia, delayed insomnia, work and interests, retardation, agitation, psychological anxiety, somatic anxiety, gastrointestinal, general somatic, libido, hypochondriasis, loss of insight, and loss of weight (Dozois & Dobson, 2002). In correlational studies, Carroll et al. (1981) found only modest congruency between parallel items on the two instruments (mean r = .60, range = -.06 to .73). Thirteen of the 17 CRSD items correlated most strongly with their HRSD counterpart, but four items assessing agitation, retardation, somatic anxiety, and lost of insight were correlated most strongly with other HRSD items. Total scores for the two instruments correlated at a higher level (.80) than the individual items, providing a reasonable degree of face validity. The CRSD showed positive convergent validity with the Beck Depression Inventory (.86), and positive discriminant validity in differentiating depressed and anxious patients as demonstrated by a correlation of .26 with the State-Trait Anxiety Inventory.

Because it has 52 items, the CRSD is significantly longer and more difficult to score than other self-report measures of depression. Also, the instrument takes about 20 minutes to complete, which is more than other available measures.

Hamilton Depression Inventory (HDI). The Hamilton Depression Inventory (Reynolds & Kobak, 1995) is a self-report version of the HRSD that is available in both a 17-item version and a 23-item version. The 17-item version evaluates the same symptoms as the standard 17-item HRSD, but the 23-item HDI adds items that are consistent with current DSM-IV criteria for depression. The scoring system in both versions is consistent

with the scoring system on the HRSD. Similar to the HRSD, some items are evaluated through the use of multiple questions or probes.

The correlation coefficients between the standard 17-tem HRSD and the full scale (i.e. 23-item) HDI and 17-item HDI were very high at .94 and .95, respectively (Reynolds & Kobak, 1999). The correlation coefficients between the two HDI measures (23-item and 17-item) and the BDI (original version) were strong at .91 and .93, respectively.

The Beck Depression Inventory

History of the Beck Depression Inventory. The original BDI was developed out of Beck's need to have a depression rating instrument for a research study designed to test the psychoanalytic theories of depression. At that time, there were no available instruments that were appropriate for Beck's purposes (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). Therefore, Beck developed a 21-item instrument based on his clinical observations of the characteristic attitudes and symptoms of depressed patients. The items covered: mood, pessimism, sense of failure, lack of satisfaction, guilty feelings, sense of punishment, self-hate, self accusations, self punitive wishes, crying spells, irritability, social withdrawal, indecisiveness, body image, work inhibition, sleep disturbance, fatigability, loss of appetite, weight loss, somatic preoccupation, and loss of libido. Each item consisted of four to five self-evaluative statements in increasing order of severity. As Beck et al. (1961) stated, "the items were chosen on the basis of their relationship to the overt behavioral manifestations of depression and do not reflect any theory regarding the etiology or the underlying psychological processes of depression" (pg. 54). In other

words, the development of the BDI was atheoretical in nature and did not reflect any hypothesized factors of depression. To understand the underlying structure of the BDI, factor analysis has subsequently been widely used to examine item relationships. Construct validity has then been inferred from the identification and categorization of factors that are descriptive of the clinical aspects of depression.

The BDI was found to correlate well with other well-established instruments that rate depression; however, the most significant relationship (.96) was found between clinicians' unstructured clinical interviews and the BDI (Beck, Steer, & Garbin, 1988). Although this is not surprising, given the fact that the BDI was developed on the basis of clinical observation of depressed patients, it does point to an opportunity for a more time and cost efficient means for clinicians to rate their patients.

Ten years after the introduction of the BDI, Beck and his associates developed a revised version, the BDI-IA, based on suggestions made by May, Urquhart, and Tarran (1969). The revised version eliminated alternate wordings of the same responses (e.g., "I don't get satisfaction out of anything anymore" and "I am dissatisfied with everything") and removed double negatives (e.g., "I don't feel I am any worse than anybody else"). Changes of this nature, that is, reducing the number of statements or changing statement wording, were made to 15 of the 21 items. Statements within the remaining 6 items (i.e., irritability, crying, fatigability, appetite, weight loss, and loss of libido) stayed the same. A further revision involved the time frame for the instrument. The original BDI asked respondents to rate themselves "right now", whereas the revised BDI-IA called for respondents to rate themselves in the previous week, including today (Beck, Steer, & Garbin, 1988). Therefore, the time period of focus shifted from current, momentary, or

"state" feelings to more enduring attitudes or "trait" feelings. The BDI-IA was copyrighted in 1978 and published by Beck, Rush, Shaw, and Emery in 1979. The technical manual for the BDI-IA was subsequently published in 1987 (Beck & Steer, 1987).

The BDI underwent further revision in 1996 (Beck, Steer, & Brown, 1996) with the development of the BDI-II, which was a more significant modification than that encompassed in the BDI-IA. The major impetus for the BDI-II was to make the content more consistent with diagnostic criteria for major depressive disorders in the DSM-IV.

Beck Depression Inventory–II. The Beck Depression Inventory–II (Beck, Steer, & Brown, 1996) is a 21-item self-report inventory for assessing the severity of depression in people 13+ years old. Response items now correspond to the criteria listed in the DSM-IV. The BDI-II differs from the earlier versions of the instrument in item content rather than design format. Specifically, items addressing body image change, somatic preoccupation, and work difficulty were replaced with items that explore levels of agitation, worthlessness, concentration difficulty, and loss of energy. In addition, items addressing appetite and sleep patterns were revised to allow for reporting of increases as well as decreases in these symptoms. Also, the time frame for reporting symptoms was increased from the previous week to the previous two weeks. Last, all but three items (suicidal thoughts or wishes, punishment feelings, and loss of interest in sex) were reworded in order to describe patient symptoms better (Beck et al., 1996). Item statements are still rated on a 4-point scale, from 0 to 3, in ascending levels of severity. Interpretation of severity is based on the following cutoff scores: a total score of 0-13

indicates minimal depression and such scores are typical of nonpsychiatric normals; a score of 14-19 suggests mild depression; scores of 20-28 are indicative of moderate depression; and scores from 29 to 63 are indicative of severe depression. High scores are thought to be predictive of suicidal risk. The BDI-II takes about 5-10 minutes to complete, making it less time consuming than most other self-report depression instruments.

The BDI-II has been the subject of extensive psychometric evaluation. Beck, Steer, and Brown (1996) reported internal consistency of .92 for psychiatric outpatients and .93 for college students. Test-retest stability over a one week interval was high at .93 among a sample of 26 outpatients. Construct validity was measured by correlation with the BDI-IA (.93), the Beck Hopelessness Scale (.68), the Beck Scale for Suicide Ideation (.37), and the HRSD (.71). In contrast to the BDI-IA, the BDI-II has improved clinical sensitivity and utility due to its correspondence with the DSM-IV.

Dozois, Dobson, and Ahnberg (1998) investigated the psychometric properties of the BDI-II, compared with the original BDI. These authors concluded that both instruments demonstrated high reliability and validity, but that the BDI-II is a stronger measurement than the BDI in terms of its factor structure.

Although the BDI-II is intended primarily to measure symptom severity in clinically diagnosed patients, it has been used extensively as a screening instrument with medical inpatients and outpatients (Craven, Rodin, & Littlefield, 1988) and with community populations (Whitaker et al., 1990). In addition, Westefeld and Liddell (1994) noted that the BDI-II may be particularly useful in predicting suicide in college students.

Utility of Self-Report and Clinician Rating Scales

The differential merits of self-report versus clinician ratings is a subject of on-going debate; however, a sizeable amount of evidence suggests that the two techniques have strong and weak points of about equal magnitude (Derogatis & Lynn, 1999). Self-report measures tend to be not only brief and inexpensive to administer, but also well tolerated by patients, giving them the advantages of cost-efficiency and high cost-benefit. They may be used in a variety of settings, minimizing professional time and effort. Their administration, scoring, and interpretation require little, and in some cases, no professional input. For example, a number of self-report instruments have been adapted for interactive computer administration in which all aspects of the process are done automatically. A final merit of self-report instruments is that they are completed by the person who is directly experiencing the symptoms. As pointed out by Derogatis & Lynn (1999), clinicians can never know the actual experience of a patient, and must be satisfied with an apparent or deduced representation of the patient's experience. A source of weakness in self-report measures is the potential for patient bias, that is, the potential for the patient, consciously or unconsciously, to distort his or her responses either for personal gain, through acquiescent responding, or by attempts at impression management. However, patient bias is not generally viewed as a major source of error variance in realistic clinical situations. Perhaps a greater area of limitation is that self-report measures capture only information about the specific questions asked. Other aspects of the person such as facial expression, attitudes and postures, and cognitive/emotional status cannot be appreciated.

Clinical rating scales have the opposite strengths and weaknesses of self-report measures. They are more flexible because the clinician has the liberty to pursue lines of questioning in the areas of history, thoughts, and behaviors that might further illuminate the patient's mental health status. However, because clinical rating scales introduce an element of judgement into the rating process, they are subject to powerful clinician bias, as are self-reports subject to patient bias. Despite extensive training in the use of clinical rating scales to minimize clinician bias, it can never be totally eliminated. Another weakness of clinical rating scales is that they are more costly in terms of learning curve, administration time, and scoring.

Composition of the Literature on Self-Report and Observer Correspondence

The majority of studies investigating the correspondence between self-report and observer ratings have focused on personality factor ratings rather than affectivity ratings. A few authors have examined the convergence between affective self-ratings and the ratings made by peer observers, using instruments that measure global affectivity. These will be reviewed later as they pertain to the current study. No authors have compared self-report and observer ratings on specific measures of affect such as depression when the observer was a peer or family member. However, there is a fairly large body of literature regarding the convergence between self-report and clinician ratings of affectivity, and more specifically, depression. It was interesting to note that the subject of correspondence in self-report and observer ratings has not been as active in the literature over the past several years as it had been previously.

Correspondence of Self-Report and Clinician Ratings in Depression

There has been ample discussion in the literature about the degree to which self-report and clinician ratings of depression exhibit convergence. The vast majority of these studies have investigated convergence in ratings between the original versions of the BDI and HRSD. Although most studies have shown that the scores on the original versions of these instruments are positively correlated, the estimate of the size of the correlation has varied considerably. In the mid-1970's, Schnurr, Hoaken, and Jarrett (1976) reported a correlation of .16, but Davies, Burrows, and Poynton (1975) found a correlation of .73. Clark and Watson (1991) found a weighted mean correlation of .72 across 16 studies comparing the BDI and the HRSD. In more recent studies, Schotte, Maes, Cluydts, DeDoncker, and Cosyns (1997) found a correlation between the BDI and the HRSD of .36. Additionally, Enns, Larson, and Cox (2000) found a correlation of .40 between the two original instruments. Results of these investigations show a wide range of concordance, from low to high, between the two measures.

Research has also shown that there can be substantially discordant ratings among a significant number of patients (Domken, Scott, & Kelly, 1994; Sayer et al., 1993). For example, Bailey and Coppen (1976) reported that satisfactory correlations between the two instruments were found in two-thirds of the patients studied; however, very divergent results were characteristic in the remaining third of patients.

Several authors (Brown, Schulberg, & Madonia, 1995; Enns et al., 2000; Moran & Lambert, 1983; Schotte et al., 1997; Steer, Beck, Riskind, & Brown, 1987) have explained these variances as a function of differences in item content between the BDI

and HRSD. In examining this proposition, factor analytic studies have concluded that the BDI and HRSD capture different aspects of the depressive experience; that is, the components of depression assessed by the two scales are very different (Brown et al., 1995; Sayer et al., 1993; Steer et al., 1987). Specifically, the original BDI is posited to emphasize psychological and subjective experiences of depression; the original HRSD focuses more on somatic and vegetative symptoms.

In recent years, the problem of content differences has been addressed through revisions to both instruments; these have resulted in a closer alignment with DSM-IV diagnostic criteria for major depression, and consequently, a closer alignment with each other. However, the only study found (Beck, Steer, Ball, & Ranieri, 1996) which evaluated the equivalence between the newer versions of the two instruments, i.e., the BDI-II and the revised HRSD, reported a correlation of .71 between the two instruments. This suggested that the upgraded instruments have led to a directionally closer correspondence with each other, in contrast to that observed for the older instruments; however, that there are still discrepancies between the two measurements. There is still concern that the obtained differences between the BDI-II and HRSD scores reflect not only differences in item content but also the methodological approach used to obtain the ratings. Steer (1987) suggested that a large part of the discrepancy between the two instruments is due to differences in the method of gathering the data, that is, self-report versus observer ratings.

Opportunities to Address Current Issues

Studies have suggested that the HRSD is a more conservative measure of depressive symptomatology than the BDI for two reasons. First, despite Hamilton's intent that clinicians rate patients independently as described above, clinicians tend to rate patients on a continuum of severity based on the relative symptoms of all patients they have seen over time. In contrast, patients tend to view their symptoms as more extreme than do clinicians because they are comparing their present feeling state with their own baseline experience. As a result of the combined effect of these two factors, there can be wide disparities between the objective ratings of clinicians and the subjective ratings of patients. Domken, Scott, and Kelly (1994) used a matched clinician and self-rating scale, the Inventory for Depressive Symptomatology, to assess concordance rates between self-report and observer ratings of depression. These authors found that patients generally rated their symptoms as being more severe than did the clinicians.

In addition to being more conservative than self-report ratings, clinician ratings derived from an interview process, such as that used by the HRSD, are often thought to be more valid than self-report scales (Katz, Shaw, Vallis, & Kaiser, 1995). This perception is derived from the assumptions that clinicians have more information available to them that they can bring to bear on the ratings, and that they also possess a higher level of assessment accuracy due to their educational training and experience. However, it is also hypothesized that the interrater reliability of interview-based measures may be lower than surmised because of factors such as differences in the level of training received in the use of the instrument, differences in level of exposure to the patient,

differences in the amount of experience using the instruments, and differences in the conceptual framework used to assess depression.

Compounding this effect is the additional impact of using potentially discordant clinician and self-report instruments, such as the HRSD and BDI, to assess severity of depression for the same individual patient. As noted above, the original HRSD and BDI are not purported to measure exactly the same constructs of depression. Hence, clinicians must make subjective assumptions about how the separate scores on the two instruments converge in order to derive a defensible case conceptualization of the patient. Further complicating the picture is the fact that the algorithm or framework used to reconcile the discrepancies between instruments will likely vary between individual clinicians.

The aim of the BDI-II-O was to provide a measure which minimizes both item content dissimilarities and method/form related disparities in scores between the self-report and clinician ratings of depression. The concordance between the BDI-II and the BDI-II-O as compared with that between the BDI-II and the HRSD is of interest, requiring subsequent study using appropriate versions to attain meaningful conclusions.

As pointed out by several authors (Cronbach & Meehl, 1995; Kagan, 1988; and Loevinger, 1957), it is important to consider non-self report data in establishing the validity of psychological constructs. Virtually all of the research to date on the BDI-II involving validation of the depression construct relies on self-report data or the convergence of correlations between the BDI-II and related, but not identical, instruments such as the HRSD. As already established, the latter attempts at validation have produced only modest results. This study extended earlier findings on the construct validity of the BDI-II by examining the convergent validity of depression ratings in relation to

judgements made by observers, specifically clinicians in a clinical setting, through a parallel instrument.

Self- versus Peer-Observer Ratings in Affective Trait Assessment

Returning to the literature on self-report versus observer ratings in affective trait assessment in which the observer is a non-clinician, a number of studies have examined the convergence of affect ratings by using peers or spouses as raters. A few studies have examined the effects of using multiple raters of the same person. The collective results of these studies are now presented as the basis for the selection of demographic variables that were used in the present study, as well as for the determination of the required number of multiple raters to produce robust results. In addition, the conclusions drawn from these studies helped to anchor the current study in previously identified constructs that were employed during statistical analysis or considered during the interpretation of results.

Watson and Clark (1991) conducted a study in which each self-reporter was rated by four peer observers on eight global affect scales such as Sadness, Guilt, Positive Affect, and Shyness. Levels of self-peer agreement ranged from .19 (for Guilt, Fatigue, and Surprise) to .41 for Shyness, with a mean value of .27. The authors noted that this study used groups that were fairly well acquainted and, as such, results may be less robust when less acquainted dyads are used. The study also examined interrater reliability which ranged from .10 (Guilt) to .37 (Positive Affect), with a median value of .27. Interrater agreement tended to parallel the self-peer convergence correlations.

Specifically, scales with the highest self-peer agreement correlations also showed the best interrater agreement. These results mirror those found by Funder and his associates (Funder & Colvin, 1988; Funder & Dobroth, 1987), who also found a strong link between interrater reliability and self-peer agreement.

The literature on self-peer convergence has suggested a number of factors that influence the level of self-observer agreement as detailed below.

Trait visibility effect. A number of authors have demonstrated the fact that some emotional traits are easier to judge than others. Specifically, it has been shown that traits which are more externally visible are easier to judge that those that are more subjective and internal (Albright, Kenny, & Malloy, 1988; Funder & Colvin, 1988; Funder & Dobroth, 1987; Kendrick & Stringfield, 1980; Watson, 1989). For example, happiness is thought to be more generally ratable by others than is guilt. This finding is broadly referred to as the "trait visibility effect" (Watson & Clark, 1991). Research studies have shown that there is greater self-peer agreement on traits that have frequent, observable behavioral manifestations than on those that are less visible and more internal in nature. In addition, these studies have found stronger agreement among multiple peer raters of the same person on traits that are more highly observable. Funder and Colvin (1988) conducted a study in which self-reporters were rated by two friends and two strangers on 100 Q-sort items. They found that convergent validity and interrater reliability were significantly correlated across items, and significantly related to the rated visibility of the trait, with correlations on the latter ranging from .25 to .43 among the five agreement dyads (i.e., self-friend, self-stranger, inter-friend, inter-stranger, friend-stranger). These

findings demonstrate not only that some traits are easier to judge than others but also that agreement is influenced by the degree of trait visibility. The Watson and Clark (1991) study, comparing ratings in which there was one self-reporter and four peer observers, found further evidence of the trait visibility effect. For example, self-peer correlations were higher for more visible traits such as Shyness (.41), Hostility (.32), and Positive Affect (.29), and relatively lower for less visible traits such as Guilt (.19), Fatigue (.19), and Surprise (.19). They also found that rated visibility was significantly correlated with both self-peer agreement (.34) and interrater reliability (.55).

Acquaintance effect. A number of previous studies have found that convergent validity of affective ratings improves with increasing levels of self-peer acquaintance (Funder & Colvin, 1988; Norman & Goldberg, 1966; Watson, 1989; Watson & Clark, 1991). That is, self-peer convergence increases when the dyads are better acquainted. Watson and Clark, in the previously described study, found agreement correlations of .25 for the best-acquainted peer dyads, and .15 for the least acquainted peer dyads. Extending the analyses further to exclude ratings made by the less-acquainted judges, Watson and Clark found a convergence correlation of .40 for seven of the eight scales (the correlation on the eighth scale, Surprise, was insignificant at .14 and was thus excluded from the analysis). The best acquainted peer raters in the Watson and Clark study knew each other "well" or "very well". Because the study was conducted among college dormitory residents, the length of relationship in years, although not measured, was most likely short compared with that of subjects used in two other studies (McCrae and Costa, 1987; Kammann, Smith, Martin, and McQueen, 1984). These studies showed that very long

acquaintance yields even higher convergence between self and observer ratings. Specifically, McCrae and Costa found unusually strong self-peer agreement in a sample of subjects who had known each other for an average of 18.3 years. Likewise, Kammann, Smith, Martin, and McQueen found that self-spouse convergence correlations were somewhat higher than corresponding values obtained with friends or peers. As Watson and Clark noted, the mechanisms underlying the acquaintance effect have not been adequately illuminated. These authors hypothesized that the effect may arise from well-acquainted subjects having more of an opportunity to observe trait relevant behaviors, or alternatively, being more likely to discuss their personalities with each other than are strangers or casual acquaintances.

Aggregation effect. Research has found that self-peer convergence and interrater reliability increase as more peer observers are used (McCrae & Costa, 1987; Watson, 1989; Watson & Clark, 1991). In other words, these measures improve as the responses of multiple observers are averaged. Consistent with previous research in this area (McCrae & Costa, 1987; Norman & Goldberg, 1966; Watson, 1989), Watson & Clark found that self-peer convergence increases systematically as more peer raters are added. Mean convergence between self and peer raters for the eight scales in their study was only .18 for one rater; it rose to .22 for two raters, increased further to .25 for three raters, and reached a final level of .27 for the aggregate of 4 raters. Interestingly, results varied by scales; four scales required only one rater to reach significant correlation and two scales required a second rater for significance. All scales had significant correlations when three raters were aggregated. Watson & Clark concluded that a minimum of three

raters is necessary to achieve a generally significant level of self-peer convergence for affective traits. The authors suggested that the aggregation effect may reflect the fact that different observers view the self-reporter in different settings and under different circumstances. Alternatively, they also suggested that different observers may attend to different aspects of the same situation. Regardless, it was hypothesized that by combining the ratings of multiple observers, the averaged observer rating is based on a more extensive sample of relevant data.

It is interesting to note that self-peer agreement in the above studies was markedly lower than levels found in studies measuring patient and clinician agreement, but was impacted positively by the influences of trait visibility, acquaintanceship, and the use of multiple raters. These findings were woven into the design of the current study to expand the knowledge of these factors as they relate to patient and clinician agreement on depression scales.

The Beck Depression Inventory – II – Observer Version

The BDI-II-O was developed with a number of different goals in mind. The first goal of development was to provide a clinician rated version of the BDI-II that would be consistent with the self-report version in terms of item content.

A second goal was to provide a clinician rated instrument that is more consistent with current definitions of depression as outlined in the DSM-IV. The most widely used clinician rated instrument today, the HRSD, was developed 40 years ago, before the development of modern diagnostic criteria (Kobak & Reynolds, 2000). Although it has

been updated, the newer version does not appear to be widely employed in clinical practice or reflect adequately the prevailing construct of depression. In the past, there has been some discontent regarding the content validity of prior versions of the BDI, but with the introduction of the BDI-II these concerns have been ameliorated.

A third goal was to provide an instrument that can be used by a variety of health professionals without the time and cost of extensive training which is required by other clinician rated instruments.

A fourth goal was to provide a clinician rated instrument with more clinical utility that the HRSD. Katz, Shaw, Vallis, and Kaiser (1995) recommended that the HRSD be administered at intake and then monthly thereafter, and that the BDI be administered an intake and weekly thereafter. It is presumed that the time and cost factor associated with administering the HRSD played a role its being recommended for monthly use only. The more efficient BDI-II-O will overcome this weakness and facilitate more frequent administration of a clinician rated depression measure. This, in turn, will improve the monitoring of treatment progress and outcome.

The BDI-II-O is identical in form, content, and scoring to the BDI-II. Both instruments contain 21 items that are rated on a four-point scale, from 0-3, with higher ratings indicating greater severity of symptoms. The exception to this overall uniformity between the two instruments is that the BDI-II-O uses third person language in each of the item statements to reflect the clinician's point of view, rather than first person language reflecting the patient's point of view, as is the case with the BDI-II. For example, the item "I don't feel particularly guilty" on the BDI-II was reworded to read

"S/he does not feel particularly guilty". Likewise, the item "I am not discouraged about my future" was revised as "S/he is not discouraged about their future".

The language of the BDI-II-O was intentionally designed to be flexible enough for use across multiple types of observers. Specifically, it may be used by clinicians as well as a wide variety of other informants such as spouses, significant others, family members, and friends. In fact, any closely-related individual who has had contact with the self-reporter within the previous two weeks could likely serve as a reliable observer.

The BDI-II-O is consistent with the BDI-II in meeting NIMH guidelines for instrument selection as detailed by Maruish (1999). These guidelines state that instruments used should have the following characteristics: (1) relevance to the target individual/group, (2) simple, teachable methods of administration, (3) objective referents, (4) applicable for use with multiple respondents, (5) usable for assessment of progress and outcome, (6) cost-effectiveness, (7) understandable by nonprofessional audiences such as clients and third party payers, (8) ease of interpretation and conveyance of results to the client, (9) usefulness in clinical services, and (10) compatibility with clinical theories. A final, and obviously critical, characteristic is that the instrument has psychometric strength, which the present study was designed to substantiate. The reliability and validity of the BDI-II-O was compared with the self-report form of the BDI-II in a sample of adult psychiatric patients attending a partial hospitalization/ intensive outpatient clinic.

It is posited that the BDI-II-O, when used in conjunction with the BDI-II for comparing self-report and clinician ratings, will facilitate the collaborative process of therapy because both parties will be examining results across the same points of

reference. Any identified discrepancies will be useful to both the clinician and the patient for setting goals and monitoring progress. The BDI-II-O will assist in the common identification of problems, will offer structured opportunities to establish patient-clinician communication, and will assist in avoiding misplaced therapeutic effort due to disagreement about the goals and tasks of therapy. Information gained from comparison of the BDI-II-O with the BDI-II would be expected to improve not only clinician and patient understanding of the patient's problems, but also their severity, leading to better treatment planning. Agreement on the severity of the problems can aid treatment planning by determining the least restrictive setting for the delivery of therapeutic services, the therapeutic approach to be used, as well as the possible merits of a medication trial to decrease depressive symptoms.

Because the BDI-II-O will be quicker and easier to complete and score in contrast to the currently available clinician rating instruments, it has the advantage of being readily integrated into the clinician's daily workflow, reducing the time it takes to assess depression severity and to monitor progress.

The development of the BDI-II-O is consistent with current trends in the field. In the area of psychological assessment, the use of and reimbursement for psychological assessment, in general, is gradually being curtailed. Moreover, the use of lengthy multidimensional objective and projective instruments is being replaced by the use of brief, inexpensive, albeit well-validated, problem oriented instruments (Maruish, 1999). This change in emphasis reflects the health care industry's move toward time-limited, solution-focused treatment. When patients are being restricted to a limited number of insurance-authorized sessions, the clinician becomes hard pressed to spend a great deal of

time in assessment. The BDI-II-O is expected to address these current industry trends by offering a brief, yet highly valid, instrument for treatment planning, progress monitoring, and outcome assessment.

Purpose of the Study

One purpose of the present study was to initiate the psychometric evaluation of the new parallel form of the Beck Depression Inventory, called the BDI-II-O. In this regard, the study investigated the following psychometric properties of the BDI-II-O: internal consistency, test-retest reliability, and cluster structure. In addition, the study evaluated the correspondence between the BDI-II-O and the BDI-II on the aforementioned psychometric measures in order to establish the construct validity of the BDI-II-O. For purposes of this initial investigation, clinicians served as the observer informants. The use of other observers such as family, significant others, or friends awaits future study.

Another purpose of the study was to investigate the level of agreement between self-reporters and clinicians, and between multiple clinician raters of the same self-reporter in order to assess the prospective clinical utility of the new instrument for treatment planning. It was hypothesized that the BDI-II-O offers an opportunity to improve upon the problematic interrater reliability that is sometimes evidenced by other clinician rating scales, notably the HRSD (Kobak & Reynolds, 1999). Kobak and Reynolds commented, regarding the HRSD, that "even when raters undergo extensive reliability training, adequate interrater reliability is often not obtained" (p. 937). The

authors attributed this problem to differences in clinical training and experience, and to differences in the guidelines used to administer and score the test; this is due to the fact that Hamilton provided only general guidelines for administration and scoring. Because the BDI-II-O is less subject to variability in these areas, it was hypothesized that interrater reliability would be higher than that found with the HRSD. Nevertheless, some variability was expected because of clinician demographic factors and these were examined during the analysis.

A further purpose of the study was to examine the impact of various clinician and patient interaction factors on self-reporter and clinician agreement. Research in these areas, as described previously, has shown that the levels of acquaintanceship and the levels of familiarity between two raters influence the congruency of their ratings. The degree of acquaintanceship and familiarity in the patient-clinician dyads in the present study varied considerably because of patient, clinician, and program characteristics. This variability enabled an evaluation of the acquaintance and familiarity effects further, using multiple observers.

Finally, the effects of trait ratability on self-report and observer ratings were informally analyzed by examining item correlations and drawing appropriate inferences.

Research Hypotheses

1) The factor structure of the BDI-II and the BDI-II-O were expected to be similar, indicating parallel conceptualizations of depression.

2) The factor structure of the BDI-II and the BDI-II-O were expected include two factors: (1) a cognitive factor, and (2) a vegetative/somatic factor.
3) The BDI-II and the BDI-II-O total scores and factor scores were expected to demonstrate high internal consistency (co-efficient alpha) and test-retest reliability of at least .70.
4) Self-reporter and clinician agreement correlations were expected to be significant and positive, averaging in the moderate to high range.
5) The magnitude of self-report and clinician agreement between the BDI-II and the BDI-II-O was expected to vary according to the following patient and clinician interaction factors: length of acquaintanceship, degree of familiarity, hours per week spent together, and determination of whether or not the clinician was the primary therapist for the patient.
6) Self-reporter and clinician agreement correlations were expected to vary across response items, dependent on the trait ratability of the individual item.
7) Items judged more difficult for an informant to rate (i.e., low ratability) were expected to be associated with lower self-reporter and observer agreement.

Chapter 3
Method

Design

The psychometric properties of the BDI-II and the BDI-II-O were compared, based on self-report and clinician ratings of depression in this study, using statistics appropriate for this purpose. The psychometric properties of the BDI-II were also compared against those reported by Beck, Steer, and Brown (1996). In addition, correlations were used to determine the extent of agreement between self-report and clinician ratings, and between the individual ratings of multiple clinicians. This is a descriptive and exploratory study.

Participants

The study sample consisted of 36 self-reporters and four clinician raters. Each clinician rated all of the 36 self-reporters. Therefore the database derived from the sample included 36 self-ratings and 144 clinician ratings. This sample size is corroborated by a previous study on the convergence of self-observer affect ratings by Watson & Clark (1991) which used a sample size of 30 self-reporters and four observer raters. After the initial test administration, half of the self-reporters completed a second administration and were rated by each of the four clinicians a second time to determine test/retest reliability.

Self-reporters were adult psychiatric patients who were attending a partial hospitalization/intensive outpatient clinic in the local Harrisburg area – The Keystone

Center. Adolescents were excluded from the study because they are not represented in The Keystone Center's population. In addition, there was a desire to minimize the potentially confounding effects that might result from the inclusion of adolescents. Patients experiencing active psychotic symptoms or those with an organic brain syndrome were also excluded from the study. Patients whose psychosis was controlled by medication were eligible to participate. The patient's individual therapist made the determination about a patient's qualification for exclusion. In order to be eligible to participate in the study, a patient's duration of treatment at The Keystone Center was specified at a minimum of two weeks.

Patients attending The Keystone Center partial hospitalization program generally have chronic mental illnesses. They typically have received multiple prior mental health treatments, and often have histories of drug and/or alcohol abuse, and legal involvement resulting in incarceration. The usual length of stay in the program is 1-2 years. Patients attending The Keystone Center intensive outpatient program generally have less chronic problems, have a higher level of functioning, and have had less intensive and restrictive prior treatment for mental health or substance abuse reasons. The typical length of stay in the intensive outpatient program is 6-12 months. Depressive symptoms are highly prevalent in both groups of patients.

Clinician participants were mental health professionals who had regular contact with the patient self-reporter at The Keystone Center; this time was defined as at least once per week. Of the four clinicians that participated, two were master's level Licensed Social Workers, and two were bachelor's level psychology graduates. Each clinician served in the dual role of individual therapist for assigned patients and group leader for

assigned therapy groups. Prior to implementing the study, the program director at The Keystone Center confirmed that the majority of patients in the program were capable of completing the BDI-II, and that all program staff were familiar with the administration of the BDI-II. The administrative assistant at The Keystone Center served as witness on the informed consent documents.

Procedures

Recruiting: Clinician raters were recruited first at a small group meeting with the investigator at The Keystone Center. Self-reporters were subsequently recruited over a period of four weeks at The Keystone Center until a large enough sample was obtained.

Setting. Most self-reporters completed the BDI-II and a demographic form during one of their daily small group sessions or at their community meeting, in rooms at The Keystone Center where these activities normally take place. Some self-reporters, as the study progressed, were instead assembled in a separate room to minimize disruption of group meetings in which most of the attendees had already participated. Clinicians completed the BDI-II-O, a clinician/patient information form, and a demographic form in the location of their choice. Completed clinician forms were deposited in a locked file cabinet in a locked file room and were picked up by the researcher. Clinicians were responsible for ensuring the confidentiality of the documents at all times.

Administration. The nature and purpose of the study was explained to potential self-report participants by the researcher. The consent form was reviewed and signed by those agreeing to participate. Sufficient time was allowed for all potential participants to read the consent form thoroughly before signing, and to include a question and answer period. Individuals who decided not to participate either remained in the room or rejoined their group session. The procedures for completing the BDI-II and the demographic form were then explained to participants, questions were answered, and the instrument and demographic form were administered. After the administration was completed, participants were debriefed and then resumed their scheduled activities for that day. Self-reporter participants used their first names and last initials in the "name" box on the BDI-II and demographic information forms. The use of first names and last initials was necessary in order to allow for self-reporter and clinician ratings to be matched and correlated.

The researcher met with the clinician participants as a group to explain the study and to obtain their informed consent. Procedures for completing the BDI-II-O were reviewed. Clinicians entered the first name and last initial of the self-reporter in the "name" box on the BDI-II-O and on the clinician/patient information form. The BDI-II-O forms were coded with clinician identification numbers. They used their own first names and last initials on the clinician/patient information form and on the demographic form. The use of first names and last initials or identifying codes on these forms was necessary to allow for self-reporter and clinician ratings to be matched and correlated.

One week after the initial administration of the BDI-II and BDI-II-O, the instruments were completed a second time by one-half of the original self reporters and

each of the clinicians for this subsample of self-reporters; this allowed for the measurement of test-retest stability.

Completed BDI-II and BDI-II-O forms, clinician/patient information forms, and demographic forms were collected by the researcher and kept in a locked file drawer when not being used for analysis. Completed consent forms and release forms were also kept in a locked file drawer. Scoring of the instruments and data entry was completed by the researcher. An independent examiner verified 75% of the data for accuracy.

Materials. Materials consisted of the BDI-II form, the BDI-II-O form, the clinician/patient information form, the self-reporter demographic form, the clinician demographic form, and the consent forms. Permission to adapt the BDI-II for purposes of this study was secured from the publisher before the study was implemented. A copy of the permission letter can be found in Appendix A. Copies of the non-copyrighted materials are included in Appendices B though F.

Measures Completed by Self-Reporters

Demographic form. Self-reporters completed a demographic form specifically designed for this study; this form requested personal information, prescribed medications, and treatment history.

Beck Depression Inventory – II (BDI-II). The BDI-II is a 21-item inventory that assesses the severity of depressive symptoms. Each item is rated on a 0-3 scale with total

scores ranging between 0 and 63. Depression severity is categorized according to the following ranges of scores: Minimally Depressed = 0-13; Mildly Depressed = 14-19; Moderately Depressed = 20-28; and Severely Depressed = 29-63. The instrument has been found to demonstrate high internal consistency. In addition, adequate content validity and factorial structure has been demonstrated, and diagnostic discrimination has been established (Beck, Steer, and Brown, 1996).

Measures Completed by Clinicians

Demographic form. Clinicians completed a specially designed demographic form that requested personal and professional information deemed relevant to the purposes of the present study.

Clinician/patient information. This form, also specially designed for the present study, requested information about the clinician's level of acquaintanceship and level of familiarity with the patient, amount of time spent with the patient on a weekly basis, and whether or not the clinician was the primary therapist for the patient.

Beck Depression Inventory – II – O (BDI-II-O). The BDI-II-O, developed as a new companion instrument to the BDI-II, is the focus of this study. It parallels the content and scoring of the BDI-II and is intended for use by observer raters in reporting on the depressive symptoms of others. For this study, the observers of interest were clinicians. This study represents the first stage of investigation for the BDI-II-O.

Measures Completed by Judges of Trait Ratability

Clinician opinions on the ratability of patient traits. A total of 31 independent judges completed a survey designed specifically for this study which asked respondents to rate the difficulty or ease of rating patients on 21 traits that match those in BDI-II. The scale for trait ratability was: 1 = very difficult to rate, 2 =moderately difficult to rate, 3= a little difficult to rate, 4= a little easy to rate, 5= moderately easy to rate, and 6 = very easy to rate. A copy of the Trait Ratability Survey can be found in Appendix G.

Chapter 4
Results

Descriptive Statistics

For the initial administration, the total number of BDI-II instruments completed by self-reporters was 36, and the total number of BDIO-II-O instruments completed by the four clinicians was 144. For the retest administration, the total number of completed BDI-II instruments was 18, and the total number of completed BDIO-II-O instruments was 72. There were 31 trait ratability surveys completed by independent judges. The database for analysis was created using The Statistical Program for the Social Sciences 10.0 for Windows (SPSS).

Of the 36 self-reporters, twenty-three (63.9%) were female and thirteen (36.1%) were male. Given that the prevalence of depression skews 2:1 on a female to male basis, the study sample was representative of the general population of depressed persons on the dimension of gender. Twenty (55.6%) of the self-reporters were Caucasian, twelve (33.3%) were African American, and four (11.1%) were other races (Hispanic, Native American, Biracial, and African), providing an ethnically diverse sample. Seven (19.4%) self-reporters had less than a high school education, twelve (33.3%) graduated from high school or had a GED, fourteen (38.9%) had some college education, two (5.8%) graduated from college, and one (2.8%) had a Master's degree, providing a sample with a varied level of academic achievement. The mean age of self-reporters was 43.1 years

(SD = 9.65). The youngest and oldest self-reporters were 21 and 59 years old, respectively. A majority of 19 self-reporters were divorced (44.4%) or separated (8.3%), and a third (33.3%) had never been married. Only two self-reporters (5.6%) were married (under-representing the population), and three (8.3%) were widowed.

Self-reporters began receiving mental health services at a mean age of 26.5 years (SD = 13.0). The median was 23 years of age, and the mode was 17 years of age. The age range for commencement of mental health services was broad, extending from 9 to 55 years of age. Self-reporters had received mental health services for a mean of 16.5 years (SD = 12.4), confirming that the sample consisted of patients with chronic mental health problems. The median number of years of service was 19 years. The range was less than one year to 39 years.

All but one self-reporter had previously received some form of mental health treatment. The average number of different treatment modalities received was three. The most predominant forms of prior treatment were inpatient hospitalization (27.9%), partial hospitalization (26.1%), outpatient therapy (19.8%), and intensive outpatient therapy (14.4%). Other forms of treatment included supported living, self-help groups, and residential crisis intervention. Thirty-one (86.1%) self-reporters previously had at least one inpatient hospitalization for mental health reasons. The mean number of hospitalizations for those patients receiving inpatient hospitalization was 13.0; the median was 4.0, and the mode was 1.0. The mean was elevated by four patients who had received between 40 and 120 hospitalizations each. Twenty-nine (80.6%) self-reporters had attended a partial hospitalization program an average of 2.4 times. Eighteen (50.0%) self-reporters had attended an intensive outpatient program an average of 2.2 times.

Twenty-one (58.3%) self-reporters had been treated by an average of four different outpatient therapists.

Slightly over half of the self-reporters (52.8%) had previously received drug and alcohol treatment. The average number of different treatment modalities received was 2.5. The various forms of prior drug and alcohol treatment included inpatient hospitalization (25.6%), partial hospitalization (19.1%), residential treatment (12.8%), halfway house (10.6%), and attendance at Alcoholics Anonymous or Narcotics Anonymous (31.9%). Twelve (33.3%) self-reporters had received inpatient hospitalization for substance abuse problems an average of 2.3 times. Nine (25.0%) self-reporters had attended a drug and alcohol partial hospitalization program an average of 1.6 times. Six (16.7%) self-reporters had attended residential drug and alcohol treatment an average of 1.3 times. Five (13.9%) self-reporters had participated in a halfway house program an average of 1.6 times.

All but one self-reporter was prescribed psychotropic medication at the time of the study. The average number of prescribed medications per patient was three. Of the medications prescribed, 40.1% were antidepressants, 11.8% were anxiolytics, 23.6% were antipsychotics, 12.7% were mood stabilizers, and 11.8% were other medications consisting primarily of hypnotics and anticholinergics. A total of 32 different medications were prescribed. A large majority of self-reporters (88.9%) were taking at least one antidepressant. One-third (33.4%) of self-reporters were taking at least one anxiolytic, and one-third (33.4%) were taking at least one mood stabilizer. Almost two-thirds (63.9%) of self-reporters were taking at least one antipsychotic, and 30.6% were taking an hypnotic and/or an anticholinergic medication.

The four clinician-raters were all female and Caucasian. Their mean age was 33.5 years (SD = 7.55). Two of the clinicians were Licensed Social Workers, and two were Bachelor's level psychology graduates. The clinicians as a group had an average of 5 years, 10 months experience in the mental health field (SD = 2.03). Level of experience ranged from 3 years, 6 months to 7 years, 8 months.

Of the 31 independent judges who responded to the trait ratability survey, seven (22.6%) were faculty members at the Philadelphia College of Osteopathic Medicine, seven (22.6%) were licensed psychologists in clinical practice, six (19.4%) were recent graduates of the Philadelphia College of Osteopathic Medicine Doctorate in Clinical Psychology program, seven (22.6%) were advanced students in the Philadelphia College of Osteopathic Medicine Doctorate in Clinical Psychology program, and four (12.8%) were psychiatrists (three of whom were in clinical practice at Psychological and Behavioral Services of Keystone Children and Family Services; one who was in clinical practice in Philadelphia was also a teaching psychiatrist). Twenty-eight (90.3%) judges were Caucasian, two (6.5%) were African American, and one (3.2%) was Chinese. Thirteen (41.9%) of the judges were male and eighteen (58.1%) of the judges were female.

Test of Normality in the Distribution of Total Scores

The Kolmogorov-Smirnov test of normality was used to determine if the obtained total scores for the BDI-II, the BDI-II-O clinician average, and the BDIO-II-O individual clinician scores approximated a normal distribution. The Komogorov-Smirnov statistic

for each of these total scores indicated that the distribution of the scores resembled a normal distribution.

Analysis of BDI-II and BDI-II-O Total Scores

The mean, mode, and median values for the total scores on the BDI-II were examined. The mean total score on the BDI-II was 33.94 (SD = 13.85), the median was 35.50; there were multiple modes. Total scores ranged from 8 to 61, with a possible range of scores from 0 to 63. The mean value of 33.94 indicates that the study sample was Severely Depressed according to the diagnostic ranges presented by Beck, Steer, and Brown (1996) in the BDI-II manual, confirming a study sample that was describing a high level of self-reported depression.

The mean, mode, and median values for the total scores on the BDI-II-O were examined, using averaged and individual clinician ratings. The mean total score on the BDI-II-O using averaged clinician scores was 23.41 (SD = 9.13), the median was 25.13, and there were multiple modes. Total scores, rounded to the nearest integer, ranged from 5 to 43, with a possible range of scores from 0 to 63. The mean total scores for the individual clinicians were 21.78 (SD = 9.00), 21.86 (SD = 11.35), 23.00 (SD = 8.19), and 27.00 (SD = 12.63) for Clinicians 1 through 4, respectively. The clinician ratings, as a group average and by individual clinician, described a study sample that was Moderately Depressed according to the diagnostic ranges presented by Beck et al. (1996). Table 1 summarizes the means, standard deviations, and mean depression severity ratings for the BDI-II and BDI-II-O.

Table 1

Self-Reporter (BDI-II) and Clinician (BDI-II-O) Total Scores

	Total Scores		
Respondent	Mean	Standard Deviation	Severity of Depression
Self-reporter	33.94	13.85	Severely Depressed
Clinician Average	23.41	9.13	Moderately Depressed
Clinician 1	21.78	8.99	Moderately Depressed
Clinician 2	21.86	11.35	Moderately Depressed
Clinician 3	23.00	8.19	Moderately Depressed
Clinician 4	27.00	12.13	Moderately Depressed

An independent samples t-test was performed to determine if there was a significant difference between the mean total scores on the BDI-II and BDI-II-O, using averaged clinician ratings for the BDI-II-O. This analysis showed that the mean scores of the self-reporters and clinicians were significantly different ($t = 3.81$, $p < .05$).

A one-way ANOVA was performed to determine if a difference in BDI-II-O mean total scores existed among the four clinician raters. This analysis revealed no significant difference in the mean BDIO-II-O scores of the clinician raters, $F (3, 140) = 1.99$, $p > .05$. In other words, the mean BDIO-II-O total scores of the individual clinicians were statistically equivalent.

Distribution of Severity Ratings

Based on the Beck et al. (1996) scoring system, and as presented in Table 2, five self-reporters (13.9%) rated themselves as minimally depressed, seven (19.4%) rated themselves as moderately depressed, and twenty-four (66.7%) rated themselves as severely depressed. No self-reporters rated themselves as mildly depressed.

Clinicians, as a group, rated six (16.7%) patients as minimally depressed, six (16.7%) as mildly depressed, thirteen (36.1%) as moderately depressed, and eleven (30.5%) as severely depressed.

Table 2

Self-Reporter (BDI-II) and Clinician (BDI-II-O) Severity Ratings

	Frequency of Rating			
	Self-reporter (BDI-II)		Clinicians (BDI-II-O)	
Severity Rating	N	%	N	%
Minimal	5	13.9	6	16.7
Mild	0	0.0	6	16.7
Moderate	7	19.4	13	36.1
Severe	24	66.7	11	30.5
Total	36	100.0	36	100.0

Self-Reporter and Clinician Endorsement of Symptoms

As presented in Table 3, patients endorsed depressive symptoms at a higher rate than did clinicians for most items. Endorsement refers to the selection of response categories 1, 2, or 3, with higher numbers denoting greater symptom intensity. A review of the endorsement data shows that the greatest disparity between patient and clinician symptom endorsement occurred on the affective and somatic items shown in italics. The only disparity on the cognitive dimension was for Suicidal Thoughts or Wishes.

Correlations between the BDI-II and BDI-II-O

Pearson product moment correlation coefficients were used to explore the relationship between the total scores of the BDI-II and BDI-II-O, using the averaged clinician ratings for the BDI-II-O. Pearson product moment correlation coefficients were also used to explore the relationship between the total scores of the BDI-II and BDI-II-O using the ratings of individual clinicians.

As shown in Table 4, there was a significant relationship of .59 ($p < .01$) between the total scores of the BDI-II and the BDI-II-O using averaged clinician ratings on the BDI-II-O, suggesting a moderate level of correspondence between self-report and clinician total scores. The correlations between the BDI-II and the BDI-II-O total scores by individual clinician were each at a moderate level of correspondence, and all significant at the $p < .01$ level; however, the correlation for Clinician 3 was at the low moderate level compared with the other clinicians.

Table 3

Endorsement of Depressive Symptoms for Self-Reporters and Clinicians

	Symptom Endorsement (%)	
Symptom (Item)	BDI -II	BDI-II-O
Sadness	.92	.86
Pessimism	.83	.81
Past Failure	.83	.89
Loss of Pleasure	.83	.89
Guilty Feelings	.89	.67
Punishment Feelings	.69	.64
Self-Dislike	.94	.81
Self-Criticalness	.86	.75
Suicidal Thoughts or Wishes	.75	.36
Crying	*.72*	*.25*
Agitation	*.86*	*.28*
Loss of Interest	.83	.81
Indecisiveness	.86	.75
Worthlessness	.78	.83
Loss of Energy	.86	.72
Changes in Sleeping Pattern	*.94*	*.47*
Irritability	*.67*	*.31*
Changes in Appetite	*.86*	*.19*
Concentration Difficulty	.89	.78
Tiredness or Fatigue	.81	.72
Loss of Sexual Interest	*.72*	*.14*

Table 4

Correlations between BDI-II and BDIO-II-O Total Scores

Respondent	Correlation with BDI-II
Clinician Average	.59**
Clinician 1	.62**
Clinician 2	.51**
Clinician 3	.42**
Clinician 4	.54**

** Significant at the $p < .01$ level (one-tailed)

Correlations between Individual Clinicians

Pearson product moment correlation coefficients were used to explore the relationship between the total scores of individual clinicians on the BDI-II-O. These correlations, as detailed in Table 5, show moderate to strong relationships between all six paired combinations of clinician total scores on the BDI-II-O. The strongest relationship was between Clinician 2 and Clinician 4 (.85), both of whom also showed a strong relationship with Clinician 1 at .74 for Clinician 2 and .77 for Clinician 4. Clinician 3 showed a consistently weaker relationship with all of the other clinicians, albeit still at a moderate and significant level.

Table 5

Correlations between Individual Clinician (BDIO-II-O) Total Scores

	Clinician 1	Clinician 2	Clinician 3	Clinician 4
Clinician 1	1.00			
Clinician 2	.74**	1.00		
Clinician 3	.60**	.66**	1.00	
Clinician 4	.77**	.85**	.58**	1.00

** Significant at the $p < .01$ level (one-tailed)

Item Level Correlations

Pearson product moment correlation coefficients were used to explore the relationship between the item scores on the BDI-II and BDI-II-O, using averaged clinician ratings on the BDI-II-O. As shown in Table 6, four items failed to achieve significance, five items were positively correlated at the $p < .05$ level, and 14 items were positively correlated at the $p < .01$ level. This table also identifies the factor dimension of each item, using the factor solution presented by Beck et al. (1996), which identified two factors, a Somatic/Affective factor (Factor 1) and a Cognitive factor (Factor 2).

Table 6

Item Level Correlations between the BDI-II and BDI-II-O

Item	BDI/BDIO Correlation	Beck et al. Factor
Loss of Interest in Sex	.16	Somatic
Changes in Sleeping Pattern	.19	Somatic
Agitation	.20	Affective
Changes in Appetite	.25	Somatic
Indecisiveness	.29*	Affective
Guilty Feelings	.33*	Cognitive
Pessimism	.33*	Cognitive
Loss of Interest	.36*	Affective
Self-Dislike	.39*	Cognitive
Sadness	.40**	Cognitive
Worthlessness	.42**	Cognitive
Suicidal Thoughts or Wishes	.43**	Cognitive
Irritability	.45**	Affective
Crying	.45**	Affective
Punishment Feelings	.45**	Cognitive
Loss of Pleasure	.49**	Affective
Past Failure	.51**	Cognitive
Tiredness or Fatigue	.51**	Somatic
Loss of Energy	.52**	Somatic
Concentration Difficulty	.55**	Somatic
Self-Criticalness	.56**	Cognitive
Total Score	.59**	

** Significant at the $p < .01$ level (one-tailed)
* Significant at the $p < .05$ level (one-tailed)

Internal Consistency of the BDI-II and BDI-II-O

Cronbach's coefficient alpha reliability was used to investigate the internal consistency of the BDI-II and the BDI-II-O, using averaged and individual clinician total scores for the BDI-II-O. In addition, the internal consistency of two item groupings was calculated, according to the factor structure presented by Beck et al. (1996). One grouping consisted of a Somatic/Affective factor (Factor 1) and the second grouping consisted of a Cognitive factor (Factor 2). These item groupings will be later referred to as derived factors. The results of the internal consistency analysis are shown in Table 7.

The coefficient alpha for the total score on the BDI-II was .93. The coefficient alpha for the total score on the BDI-II-O, using averaged clinician ratings, was .96. The coefficient alphas for the total scores of individual clinicians ranged from .85 for Clinician 3 to .94 for Clinician 4. All total score coefficient alphas were in the acceptable range.

The coefficient alphas for the BDI-II and the BDI-II-O averaged and individual clinician factor scores (based on factor groupings according to Beck et al.'s analysis) were also in the acceptable range, with the exception of the alpha for Clinician 3 on Factor 1 (Somatic/Affective) which was .69, falling just below the generally accepted .70 cutoff for adequate internal consistency. Self-reporters showed comparable internal consistency between the two factors; however, clinicians tended to show greater internal consistency on Factor 2 (Cognitive) than Factor 1 (Somatic/Affective).

Table 7

Internal Consistency of BDI-II and BDIO-II-O Total and Factor Scores

	Reliability Coefficients		
Respondent	Total Score	Factor 1 Somatic/Affective	Factor 2 Cognitive
Self-reporter	.93	.87	.89
Clinician Average	.96	.91	.97
Clinician 1	.91	.80	.92
Clinician 2	.96	.92	.93
Clinician 3	.85	.69	.86
Clinician 4	.94	.86	.94

Correlation of the Derived BDI-II Factors

Pearson product moment coefficients of correlation were used to measure the degree of correspondence between derived BDI-II factors; these were obtained by assigning items to two groupings based on the factor structure obtained by Beck et al. (1996). As previously stated, Beck et al. described two factors for the BDI-II, a Somatic/Affective factor and a Cognitive factor. The correlation between the derived factors was strong at .83 ($p < .05$).

Test-Retest Stability

An estimate of the stability of the BDI-II over time was derived from the responses of a subsample of 18 self-reporters who were administered the BDI-II at two different sessions, approximately one week apart. Test-retest reliability was calculated for the BDI-II by correlating the test and retest total scores of self-reporters. The test-retest correlation of .92 was high ($p < .01$), indicating good temporal stability. The first-session mean BDI-II total score of 36.22 (SD = 11.22) and the second-session mean BDI-II total score of 32.89 (SD = 12.44) were comparable [paired $t(17) = 2.87$, not significant].

Test-retest reliability for the BDI-II-O was calculated by correlating the test and retest total scores for each clinician and the clinician average. Each clinician completed the BDI-II-O on the subsample of 18 self-reporters at two different sessions, approximately one week apart. For the averaged clinician total scores, the test-retest correlation of .88 was significant ($p < .01$), and fairly comparable to the test-retest correlation of the self-reporters at .92 ($p < .01$). The first-session mean BDI-II-O total score of 24.88 (SD = 7.47) and the second-session mean BDI-II-O total score of 23.51 (SD = 7.00) were comparable [paired $t(17) = 1.61$, not significant]. The individual clinician test-retest correlations were all significant ($p < .01$) at 80, .61, .88, and .76 for Clinician 1 through Clinician 4, respectively, but showed some variation among the clinicians. Table 8 details the results of the test/retest analysis.

Table 8

Test/Retest Means, Standard Deviations, and Correlations

Respondent	Test		Retest		Test/ Retest Correlation
	Mean	Standard Deviation	Mean	Standard Deviation	
Self-reporter	36.22	11.22	32.89	12.44	.92**
Clinician Average	24.88	7.47	23.51	7.00	.88**
Clinician 1	22.28	8.34	22.94	7.34	.80**
Clinician 2	23.67	8.86	20.33	8.37	.61**
Clinician 3	24.83	8.31	20.94	6.50	.88**
Clinician 4	28.72	9.54	29.83	9.85	.76**

** Significant at the $p < .01$ level of significance (one-tailed)

Cluster Analysis of the BDI-II

Cluster analysis is a multivariate procedure that can be used to place variables into homogeneous groups (i.e., clusters) so that the relationship between the variables is revealed, based on objective interpretation. It detects natural groupings in the variables. An exploratory cluster analysis was performed with item scores from the BDI-II used as the clustering variables. An agglomerative hierarchical clustering procedure was used to produce preliminary cluster solutions. In this procedure, the proximity measure used was squared Euclidean distance, with cluster formation achieved through the average

between-groups method. The average between-groups method uses the average similarity of observations between two groups as the clustering measure between the two groups. Examination of the vertical icicle plot and dendrogram revealed that the most salient cluster solution was composed of one cluster containing 20 items and one single-item cluster. Items in the first cluster included all items except Loss of Interest in Sex, which remained as a single-item second cluster. For subsequent analyses, this second single-item cluster was eliminated due to lack of saliency as a separate cluster, leaving one composite cluster of 20 items, which was labeled Depressive Symptomatology. Cluster solutions containing a greater number of individual clusters, that is, more than two, displayed a higher number of single-item or small clusters that were complex in structure and lacked clarity of interpretation. The one-cluster solution was chosen for the ease and relevance of its interpretation. Cluster membership for the BDI-II items is shown in Table 9, with 1 = Cluster 1 and 2 = Cluster 2.

Cluster Analyses of the BDI-II-O

An exploratory cluster analysis was performed with item scores from the BDI-II-O clinician average used as the clustering variables. An agglomerative hierarchical clustering procedure was used to produce preliminary cluster solutions. In this procedure, the proximity measure used was squared Euclidean distance, with cluster formation achieved through the average between-groups method. For the averaged clinician item scores, examination of the vertical icicle plot and dendrogram revealed that the most salient clustering solution was composed of two distinct clusters. Cluster 1

consisted of 14 items representing a Cognitive/Performance dimension of depression. Cluster 2 consisted of 7 items representing a Somatic/Affective dimension of depression. The items in Cluster 1 included items describing psychological symptoms of a cognitive nature (Sadness, Pessimism, Past Failure, Guilty Feelings, Punishment Feelings, Self-Dislike, Self-Criticalness, and Worthlessness) and psychological symptoms relating to performance impairment (Loss of Pleasure, Loss of Energy, Loss of Interest, Concentration Difficulty, Indecisiveness, and Tiredness or Fatigue). The items included in Cluster 2 included symptoms of affective or somatic disturbance (Crying, Agitation, Changes in Sleeping Pattern, Irritability, Changes in Appetite, and Loss of interest in Sex). Of some interest is the fact that Suicidal Thoughts or Wishes, which is typically thought of as a cognitive symptom, was also included in Cluster 2.

Separate cluster analyses were performed with BDI-II-O item scores from each clinician as the clustering variables. These analyses showed a similar two-cluster solution to be the most salient for the item scores of Clinician 1 and Clinician 4. Cluster membership between these two clinicians varied on only three items (Punishment Feelings, Suicidal Thoughts or Wishes, and Changes in Sleeping Pattern), which did not change the overall interpretation of the clusters. For Clinician 1, Cluster 1 (Cognitive/Performance) contained 15 items and Cluster 2 (Somatic/Affective) contained 6 items. For Clinician 4, Cluster 1 (Cognitive/Performance) contained 14 items and Cluster 2 (Somatic/Affective) contained 7 items.

The cluster analysis for Clinician 3 also produced a salient two-cluster solution; however, the two clusters varied more dramatically in item content compared with those of Clinicians 1 and 4, primarily on the affective symptoms. As such, the cluster analysis

for Clinician 3 produced a Cluster 1 representing a Cognitive/Performance/Affective dimension and a Cluster 2 representing a Somatic dimension. The first cluster contained 18 items and the second cluster contained 3 items, all of which were somatic.

The cluster analysis for Clinician 2 revealed that a three-cluster solution was the most salient. Cluster 1 (Cognitive/Performance) contained Sadness, Pessimism, Self-Dislike, Worthlessness, Past Failure, Guilty Feelings, Self-Criticalness, Tiredness or Fatigue, Loss of Energy, Loss of Interest, and Loss of Pleasure. Cluster 2 (Affective) contained Irritability, Agitation, Crying, Suicidal Thoughts or Wishes, and Punishment. Cluster 3 (Somatic) contained Appetite Changes, Changes in Sleeping Pattern, and Loss of Interest in Sex, Concentration Difficulty, and Indecisiveness.

Cluster membership across clinicians was identical for 71% of the items. Discrepancies in cluster membership at the clinician level did not follow any particular pattern, and included a mix of cognitive, somatic, and affective items. Table 9 details the cluster membership results, with 1 = Cluster 1, 2 = Cluster 2, and 3 = Cluster 3.

Correlation between the Empirical BDI-II Factor and Derived BDI-II Factors

Pearson product moment correlation coefficients were used to measure the degree of correspondence between the empirically derived BDI-II factor (all items less Loss of Interest in Sex) in the cluster analysis, called Depressive Symptomatology, and the derived BDI-II two-factor solution, based on item groupings according to Beck et al.'s factor solution. The correlation between Depressive Symptomatology and the derived Cognitive Factor described by Beck was very strong at .95 ($p < .01$). The correlation

Table 9

Cluster Membership

Symptom (Item)	Clinician				BDI-II-O Average	BDI-II
	1	2	3	4		
Sadness	1	1	1	1	1	1
Pessimism	1	1	1	1	1	1
Past Failure	1	1	1	1	1	1
Loss of Pleasure	1	1	1	1	1	1
Guilty Feelings	1	1	1	1	1	1
Punishment Feelings	2	2	1	1	1	1
Self-Dislike	1	1	1	1	1	1
Self-Criticalness	1	1	1	1	1	1
Suicidal Thoughts or Wishes	1	2	1	2	2	1
Crying	2	2	1	2	2	1
Agitation	2	2	1	2	2	1
Loss of Interest	1	1	1	1	1	1
Indecisiveness	1	3	1	1	1	1
Worthlessness	1	2	1	1	1	1
Loss of Energy	1	1	1	1	1	1
Changes in Sleeping Pattern	1	3	2	2	2	1
Irritability	2	2	1	2	2	1
Appetite Changes	2	3	2	2	2	1
Concentration Difficulty	1	3	1	1	1	1
Tiredness or Fatigue	1	1	1	1	1	1
Loss of Sexual Interest	2	3	2	2	2	2

Note. 1 = Cluster 1, 2 = Cluster 2, 3 = Cluster 3

between Depressive Symptomatology and the derived Somatic/Affective Factor described by Beck was equally strong at .96 ($p < .01$).

Correlation between the Empirical BDI-II Factor and BDI-II-O Clusters

Pearson product moment correlation coefficients were used to explore the degree of correspondence between the empirically derived BDI-II factor in the cluster analysis (all items except Loss of Interest in Sex), called Depressive Symptomatology, and the empirical BDI-II-O factors in the cluster analysis based on averaged clinician scores. The correlation between Depressive Symptomatology and the Cognitive/Performance factor (Cluster 1) was moderately strong at .56 ($p < .01$). The correlation between Depressive Symptomatology and the Somatic/Affective factor (Cluster 2) was also moderately strong at .61 ($p < .01$).

Correlation of the BDI-II-O Clusters

Pearson product moment correlation coefficients were used to measure the degree of correspondence between the two clusters of the BDI-II-O, using the averaged clinician scores. The correlation between the clusters was strong at .79 ($p < .01$).

Internal Consistency of the BDI-II Empirical Factor

Cronbach's coefficient alpha reliability was used to investigate the internal consistency of the empirically derived BDI-II factor in the cluster analysis, Depressive Symptomatology. This analysis yielded a coefficient alpha of .93 for the Depressive Symptomatology scale.

Internal Consistency of the BDI-II-O Empirical Factors

Cronbach's coefficient alpha reliability was used to investigate the internal consistency of the empirically derived BDI-II-O factors from the cluster analysis, based on averaged clinician scores. The Cognitive/Performance factor (Cluster 1) yielded a coefficient alpha of .97 and the Somatic/Affective factor (Cluster 2) yielded a coefficient alpha of .81.

Factor Analysis of the Higher Order Factor

A Higher Order Factor was calculated by summing the two derived BDI-II factors and the two BDI-II-O clusters, based on averaged clinician scores. The Higher Order Factor enables the analysis of four separate scales at the highest level of structure, thus eliminating the influence of version type, that is, self-reporter and observer. A principal components factor analysis of the Higher Order Factor yielded one factor, with an eigenvalue of 2.93, which explained 73.2% of the variance in the combined scales.

Correlation between the Higher Order Factor and Empirical BDI-II Factor

The correlation between the Higher Order Factor and the empirical BDI-II factor, Depressive Symptomatology, was strongly significant at .93 ($p < .01$). This suggests a very high correspondence between self-report and clinician ratings on the overarching, common dimension of depression.

Correlation between the Higher Order Factor and Empirical BDI-II-O Factors

The correlations between the Higher Order Factor and the BDI-II-O empirical factors from the cluster analysis, Cognitive/Performance and Somatic/Affective, were strongly significant at .81 ($p < .01$) and .79 ($p < .01$), respectively. This also suggests a very high correspondence between self-report and observer ratings on the overarching, common dimension of depression.

Internal Consistency of the Higher Order Factor

Cronbach's coefficient alpha reliability was used to investigate the internal consistency of the Higher Order Factor. This analysis yielded a coefficient alpha of .83 for the higher order scale, which combines the derived BDI-II factor scales and the cluster scales for the averaged clinician ratings.

Interrater Reliability

Intraclass correlation coefficients were used to further measure the degree of correspondence between the clinician raters of the BDI-II-O. Because there were a fixed number of raters who were the only raters of interest, a two-way mixed effects model was used to calculate the intraclass correlation coefficients. Because there may have been systematic differences among clinician ratings, the intraclass correlation coefficients were calculated on the basis of absolute agreement rather than consistency, which would be used if the systematic variability due to raters was deemed irrelevant. Two measures of interrater reliability were produced: (1) single measure reliability in which the unit of analysis was the individual rating, and (2) average measure reliability in which the unit of analysis was the mean of all the clinician ratings.

The single measure intraclass correlation coefficient was positive and significant at .65 ($p < .05$). The average measure intraclass correlation coefficient was positive and significant at .88 ($p < .05$). The generally accepted minimum level of intraclass correlation for a single rater is .75 to .80. This suggests that multiple observer raters (e.g., clinician, family member or friend) would be likely to produce better correspondence with self-reporter ratings than would a single observer.

As another measure of clinician agreement, simple agreement among the four clinician-raters was calculated. This measurement consisted of the proportion of cases in which no raters agreed, two raters agreed, three raters agreed, and four raters agreed, expressed as a percentage by item. For example, for the Sadness item, there would be 40% agreement if two raters agreed on 14 of the 36 cases.

Results of the simple agreement analysis are detailed in Table 10. For all items combined, there was no agreement between raters on 2.7% of the cases, two raters agreed on 41.7% of the cases, three raters agreed on 38.9% of the cases, and four raters agreed on 16.7% of the cases. Items with the highest level of four-rater agreement were Sadness, Crying, Irritability, and Loss of Interest in Sex. The first three of these four items were positively correlated with the matching BDI-II item scores, although the fourth item showed no significant correlation. For the total score, there was no agreement on 75% of the cases, but two-raters had agreement on 25% of the cases. There were no instances in which three raters or four raters agreed on the total scores.

Trait Ratability

A group of thirty-one mental health professionals was asked to complete a survey on the trait ratability of each of the 21 BDI-II items on a 6-point scale ranging from1 = Very Difficult to Rate to 6 = Very Easy to Rate. The mean rating for each of the 21 items is shown in Table 11, in ascending order from hardest to easiest to rate.

A scatterplot of the item means for trait ratability and the item correlation coefficients between the total scores of the BDI-II and BDI-II-O is shown in Figure 1. This scatterplot revealed no significant overall pattern of correspondence between these two measures. There were only two instances when trait ratability and item correlation coefficients converged; these were for the Sadness and Crying items.

Table 10

Levels of Agreement between Clinician Raters

Symptom (Item)	No Raters Agreed		Two Raters Agreed		Three Raters Agreed		Four Raters Agreed	
	Cases		Cases		Cases		Cases	
	#	%	#	%	#	%	#	%
Sadness	0	0.0	15	41.6	10	27.8	11	30.6
Pessimism	0	0.0	19	52.8	13	36.1	4	11.1
Past Failure	0	0.0	12	33.3	18	50.0	6	16.7
Loss of Pleasure	1	2.8	21	58.3	10	27.8	4	11.1
Guilty Feelings	0	0.0	15	41.7	17	47.2	4	11.1
Punishment Feelings	2	5.6	21	58.3	11	30.6	2	5.6
Self-Dislike	0	0.0	16	44.4	14	38.9	6	16.7
Self-Criticalness	0	0.0	16	44.4	14	38.9	6	16.7
Suicidal Thoughts or Wishes	0	0.0	7	19.4	20	55.6	9	25.0
Crying	0	0.0	9	25.5	16	44.4	11	30.6
Agitation	0	0.0	11	30.6	19	52.8	6	16.7
Loss of Interest	0	0.0	15	41.7	17	47.2	4	11.1
Indecisiveness	0	0.0	21	58.3	10	27.8	5	13.9
Worthlessness	0	0.0	7	19.4	22	61.2	7	19.4
Loss of Energy	0	0.0	16	44.4	17	47.2	3	8.4
Changes in Sleeping Pattern	1	2.8	21	58.3	12	33.3	2	5.6
Irritability	0	0.0	9	25.5	16	44.4	11	30.6
Changes in Appetite	1	2.8	17	47.2	11	30.6	7	19.4
Concentration Difficulty	0	0.0	22	61.2	12	33.3	2	5.6
Tiredness	0	0.0	18	50.0	14	38.9	4	11.1
Loss of Interest in Sex	0	8.4	11	30.6	11	30.6	11	30.6
Item Average	<1	2.7	15	41.7	14	38.9	6	16.7
Total Score	27	75.0	9	25.0	0	0.0	0	0.0

Table 11

Trait Ratability Item Means

Symptom (Item)	Mean Rating	Standard Deviation
Punishment Feelings	3.42	1.09
Loss of Interest in Sex	3.61	1.50
Concentration Difficulty	3.71	1.35
Worthlessness	4.00	1.29
Indecisiveness	4.10	1.07
Loss of Interest	4.13	1.06
Guilty Feelings	4.16	1.06
Loss of Energy	4.16	1.16
Changes in Appetite	4.16	1.51
Tiredness	4.16	1.19
Loss of Pleasure	4.26	.96
Self-Dislike	4.29	1.07
Past Failure	4.35	1.14
Pessimism	4.42	1.05
Suicidal Thoughts or Wishes	4.42	1.39
Irritability	4.42	1.29
Changes in Sleeping Pattern	4.55	1.34
Self-Criticalness	4.58	1.23
Agitation	4.81	1.22
Sadness	5.03	.98
Crying	5.39	1.15

Note. 1 = Very Difficult to Rate, 6 = Very Easy to Rate

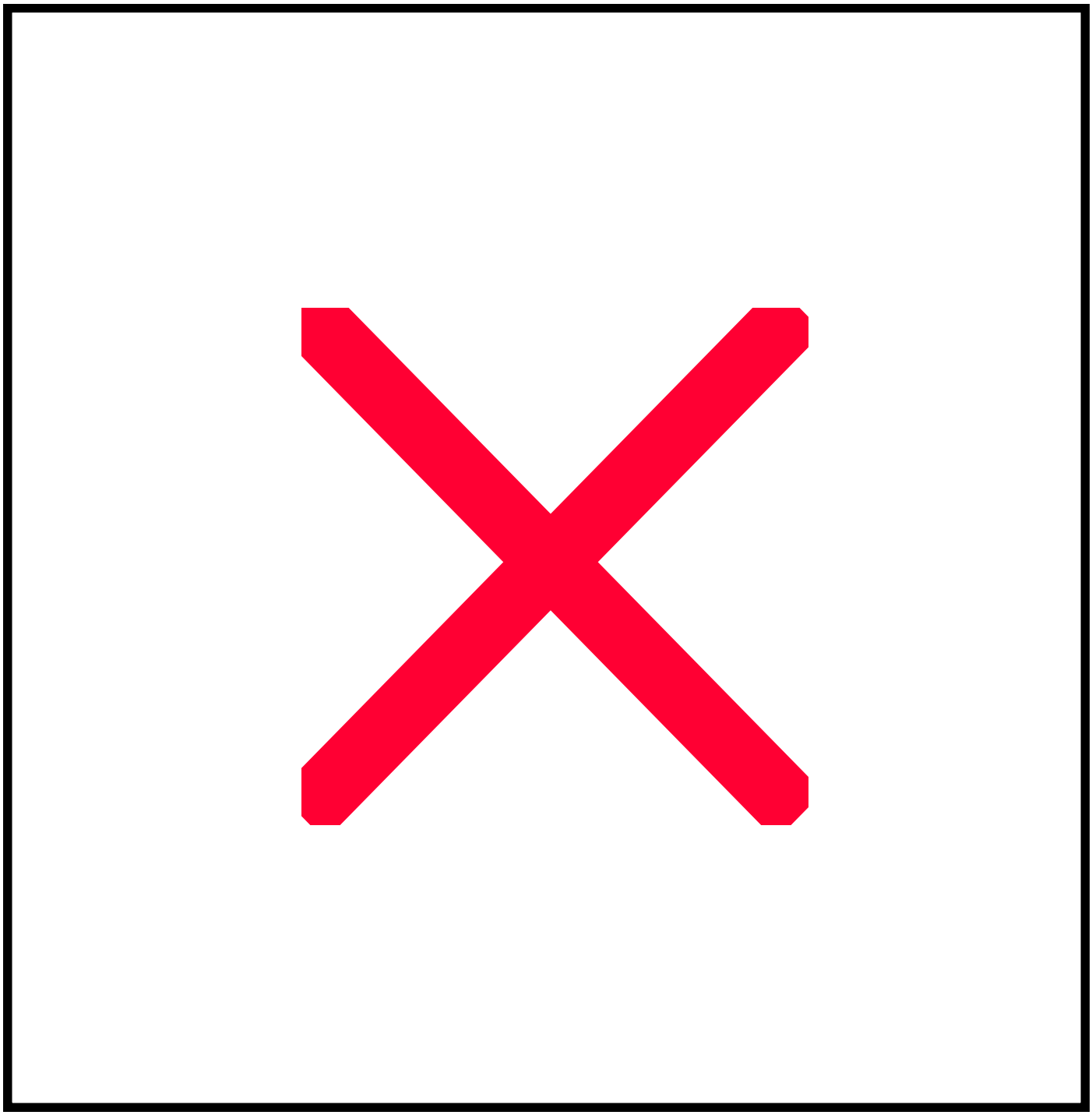

Figure 1. Relationship between trait ratability and the correspondence between total scores on the BDI-II and BDI-II-O.

Acquaintanceship

For purposes of this study, the term, *acquaintanceship*, was used to specify the length of time that a clinician and patient had known each other. Data for this measurement was provided by the clinicians on the patient/client information form.

The average length of acquaintanceship was 7.5 months (SD = 5.07) for the four clinicians combined. Three of the clinicians had been acquainted with the patient group for an average of 8.1 months, with a range of 1 month to 20 months. One clinician (Clinician 2) had been acquainted for an average of 5.6 months, with a range of 1 month to 8 months.

Pearson product moment correlations between the BDI total score and BDIO-II-O averaged and individual clinician total scores were calculated and compared for cases in which acquaintanceship was less than or equal to 6 months versus more than 6 months.

Mean BDI-II total scores and mean BDIO-II-O total scores for the averaged and individual clinician ratings were calculated and compared for cases in which acquaintanceship was less than or equal to 6 months versus more than 6 months. Results of the acquaintanceship analysis are shown in Table 12.

Contact Hours

The term, *contact hours*, was used to describe the amount of time per week that a clinician spent with the patient. Contact hours included time spent with the patient in individual therapy and in group therapy, as well as any contact in the milieu setting or by

Table 12

Acquaintanceship Correlations and Mean Total Scores

	Acquaintanceship			
	Less than or equal to 6 months (n = 18)	More than 6 months (n = 18)	Less than or equal to 6 months (n = 18)	More than 6 months (n = 18)
Respondent	Correlation with BDI-II		Mean Total Score	
Self-reporter			34.06	33.83
Clinician Average	.76**	.46*	23.74	23.08
Clinician 1	.67**	.58**	22.50	21.06
Clinician 2	.79**	.28	22.33	21.39
Clinician 3	.47*	.39	23.28	22.72
Clinician 4	.72**	.40*	26.83	27.17

** Significant at the $p < .01$ level of significance (one-tailed)

* Significant at the $p < .05$ level of significance (one-tailed)

other means such as telephone consultation. The average number of clinician/patient contact hours across the four clinicians was 3.09 (SD = 1.25) per week. The average number of clinician/patient contact hours by individual clinician was: Clinician 1 = 2.86, Clinician 2 = 3.39, Clinician 3 = 2.61, and Clinician 4 = 3.50.

Pearson product moment correlation coefficients between the BDI total score and the BDIO-II-O averaged and individual clinician total scores were calculated and compared for cases in which contact hours were less than or equal to 3 hours per week versus more than 3 hours per week.

Mean BDI-II total scores and mean BDIO-II-O total scores for the averaged and individual clinician total scores were calculated and compared for cases in which contact hours were less than or equal to 3 hours per week versus more than 3 hours per week. Results of the Contact Hours analysis are shown in Table 13.

It should be noted that a separate analysis that used number of contact hours over the length of acquaintanceship, rather than at a specific point in time during the present study, was conducted in an attempt to control for length of acquaintanceship. However, this analysis was not interpretable because the underlying assumption of constant contact hours over time was recognized as invalid due to patient, clinician, and program characteristics which varied over time.

Table 13

Contact Hours Correlations and Mean Total Scores

	Contact Hours			
	Less than or equal to 3 hours (n = 21)	More than 3 hours (n = 15)	Less than or equal to 3 hours (n = 21)	More than 3 hours (n = 15)
Respondent	Correlation with BDI-II		Mean Total Score	
Self-reporter			31.29	37.67
Clinician Average	.47*	.80*	20.27	27.80
Clinician 1	.52*	.81**	20.33	23.80
Clinician 2	.35	.74**	18.38	26.73
Clinician 3	.41	.30	19.76	27.53
Clinician 4	.42	.71**	22.62	33.13

** Significant at the $p < .01$ level of significance (one-tailed)

* Significant at the $p < .05$ level of significance (one-tailed)

Familiarity

The term, *familiarity*, was used to describe how *well* the clinician knew the patient, as distinguished from the acquaintanceship measure, which indicated how *long* the clinician knew the patient. The familiarity measure consisted of a 7-point Likert scale, on a continuum from 1 = Hardly Know at All to 7 = Know Extremely Well. The mean familiarity rating for the four clinicians was 3.65 (SD = .91). The average familiarity rating by individual clinician was: Clinician 1 = 3.92, Clinician 2 = 3.50, Clinician 3 = 3.83, and Clinician 4 = 3.33.

Pearson product moment correlations between the BDI II total score and BDIO-II-O averaged and individual clinician total scores were calculated and compared for cases in which the familiarity rating was less than or equal to 3.5 versus more than 3.5.

Mean BDI-II total scores and mean BDIO-II-O total scores for the averaged and individual clinician total scores were calculated and compared for cases in which familiarity ratings were less than or equal to 3.5 versus more than 3.5. Results of the familiarity analysis are shown in Table 14.

Primary/Not Primary Therapist

All of the 36 self-reporters had a primary therapist who was one of the four clinicians participating in the study. Clinician 1 was the primary therapist for thirteen (36.1%) of the patients, Clinician 2 was the primary therapist for eight (22.2%) of the

Table 14

Familiarity Correlations and Mean Total Scores

	Familiarity			
	Less than or equal to 3.5 rating (n = 18)	Greater than 3.5 rating (n = 18)	Less than or equal to 3.5 rating (n = 18)	Greater than 3.5 rating (n = 18)
Respondent	Correlation with BDI-II		Mean Total Score	
Self-reporter			35.06	32.83
Clinician Average	.46	.73**	23.04	23.78
Clinician 1	.44	.78**	21.22	22.33
Clinician 2	.42	.58*	22.67	21.06
Clinician 3	.44	.45	21.89	24.11
Clinician 4	.39	.69**	26.39	27.61

** Significant at the $p < .01$ level of significance (one-tailed)

* Significant at the $p < .05$ level of significance (one-tailed)

patients, Clinician 3 the primary therapist for five (13.9%) of the patients, and Clinician 4 was the primary therapist for ten (27.8%) of the patients.

Pearson product moment correlation coefficients between the BDI total score and the BDIO-II-O averaged and individual clinician total scores were calculated and compared for cases based on the role of the clinician in the patient's care; that is, whether or not the clinician was the patient's primary therapist or not. For primary therapists, the correlation between the BDI-II and BDI-II-O was significant at .52 ($p < .01$). For non-primary therapists, the correlation between the BDI-II and BDI-II-O was also significant at .59 ($p < .01$).

Mean total scores for the BDI and BDIO-II-O were also calculated based on the role of the therapist in the patient's care; that is, whether or not the clinician was the patient's primary therapist or not. The mean BDI-II-O total score for patients rated by their primary therapist was 24.39 (SD = 9.99). The mean BDI-II-O score for patients rated by all other (i.e., non-primary) therapists was 23.11 (SD = 9.31). There was no significant difference between the BDI-II-O scores of clinicians who were the primary therapist for the patient compared with those who were not [paired $t(35) = 1.35$, ns]. Results of the analysis focusing on the clinician's role in the patient's care are detailed in Table 15.

Table 15

Primary/Not Primary Therapist Correlations and Mean Total Scores

	Role of Clinician			
	Primary therapist	Not primary therapist	Primary therapist	Not primary therapist
Respondent	Correlation with BDI-II (n)		Mean Total Score	
Self-reporter			34.36	33.95
Clinician Average	.52** (36)	.59** (36)	24.39	23.10
Clinician 1	.38 (13)	.74** (23)	22.15	21.57
Clinician 2	.50 (8)	.54** (28)	26.38	20.57
Clinician 3	.37 (5)	.45* (31)	24.20	22.81
Clinician 4	.77** (10)	.47* (26)	25.60	27.54

** Significant at the $p < .01$ level of significance (one-tailed)

* Significant at the $p < .05$ level of significance (one-tailed)

Chapter 5
Discussion

The BDI-II-O was developed not only to provide an observer rated measure of depression that is consistent with current criteria for depressive symptomatology, but also to improve on the clinical utility of other currently available measures.

The present study focused on the initial validation of the BDI-II-O as an observer rated instrument. Our investigation was limited to the examination of clinicians as the observers; however, the instrument is equally well suited for use by other types of observers, including family and peers. The assessment of depression, whether by self-report or clinical observation, is important for purposes of accurate treatment planning and monitoring. It follows that the instruments used to assess depression need to be psychometrically sound.

The BDI-II-O demonstrated high internal consistency and test-retest reliability, a necessary condition for the establishment of validity. The alpha coefficient for internal consistency suggested a high degree of item response homogeneity, indicating that clinicians were responding to scale items in a consistent manner. The results of the test-retest reliability analyses suggested a very high level of temporal stability for the BDI-II-O; there were minimal change in raw scores over a one-week time period.

Evidence for the validity of the BDI-II-O was also examined from the perspective of construct validity. A cluster analysis provided insight into the underlying structure of the instrument, which consisted of two descriptive components of depression. One component, a cognitive/performance dimension, included symptoms related to self-reproach and performance impairment. The other component, a somatic/affective

dimension, included symptoms related to disturbances of a physical or affective nature. Future study, with a larger sample, could help to clarify the construct validity of the BDI-II-O through factor analysis. However, given the fact that the BDI-II has shown a high degree of construct validity, and the BDI-II-O has comparable content, it would be reasonable to assume construct validity for the BDI-II-O based on convergent forms.

A moderate degree of correspondence was found between the total scores on the BDI-II and BDI-II-O. The mean difference between the scores of self-reporters and clinicians of 10.5 points indicated that the two groups were substantially discrepant in their assessments of depression severity. Patients reported a significantly higher level of depressive symptoms than did clinicians for those patients. The discovery of divergence between self-report and clinician ratings is considered to be an important finding because it supports the clinical utility of the BDI-II-O for treatment planning, particularly its use in fostering agreement on problematic symptoms and the goals of therapy. Had the BDI-II-O produced totally equivalent scores with the BDI-II, it could be argued that there would be no need for its further development.

There was a moderate to high degree of correspondence in ratings between individual pairs of clinicians. In addition, interrater reliability for all clinicians combined, as measured by the intraclass correlation coefficient, was found to be in the high range of acceptability. However, it was noted that simple agreement levels were somewhat low on both an individual item and total score basis.

Overall, the results of this study provided strong preliminary support of the reliability and validity of the BDI-II-O as a clinician-rated measure of the severity of depression.

Sample Comparison

The self-reporter sample in the present study was compared with the standardized psychiatric outpatient sample (n=500) used by Beck, Steer, and Brown (1996) to validate the BDI-II. Self-reporters in the present study were 64% female and 36% male, which was almost identical to the psychiatric outpatient sample at 63% female and 37% male. The average age of self-reporters in the present study was 43.1 years (SD = 9.65), which was slightly older than the psychiatric outpatient sample with an average age of 37.2 years (SD = 15.91). The racial/ethnic background of the present sample was fairly diverse with Caucasians accounting for 56% of the sample, African Americans accounting for 33% of the sample, and other races making up the remaining 11% of the sample. In contrast, the psychiatric outpatient sample was heavily skewed toward Caucasians (91%).

The mean BDI-II total score in the present study was 33.94 (SD = 13.85), compared with a mean BDI-II total score in the psychiatric outpatient sample of 20.27 (SD = 10.46). This difference in mean scores reveals that the patients in the present study were reporting a higher level of depressive symptoms than those in the psychiatric outpatient sample. Specifically, the patient group in the present study fell into the Severely Depressed category of scores (range 29-63), but the psychiatric outpatient sample fell into the Moderately Depressed category of scores (range 20 to 28).

There are several possible explanations for the discrepancy in mean scores between the two samples. First, patients in the present study had chronic mental health problems and were attending an intensive outpatient/partial hospitalization program.

They had received multiple forms of prior treatment, including inpatient hospitalization in 86% of the patients. At the time of the study, they had previously received an average of 16½ years of mental health treatment over their lifetimes. Patients in the Beck et al. (1996) study were described as attending one of four different psychiatric outpatient clinics and may have had fewer chronic problems or a higher level of coping skills.

Second, slightly over half of the patients in the present study had a history of comorbid substance abuse problems, which may have contributed to elevated depressive symptoms. However, since substance history was not reported for the psychiatric outpatient sample, a direct comparison on this attribute was not possible.

Third, there may have been diagnostic differences between the two samples. Although patients in the present sample did not report their diagnoses, they did provide information on prescribed medications. Since almost 90% of the patients were prescribed at least one antidepressant medication, and were receiving medication management by a treating psychiatrist at the facility, it can be reasonably deduced that the majority of the patients in the present study had been diagnosed with a mood disorder. In contrast, only 53% of patients in psychiatric outpatient sample had been diagnosed with a mood disorder.

A final alternative explanation for the discrepancy in mean scores could be that the patients in the present study overstated their depressive symptoms compared with the psychiatric outpatient sample. Therefore, it would be useful at this point to consider the clinicians' view of the patient group's depression.

Clinician Assessment of Patient Depression

The clinicians as a group rated the patient sample in the Moderately Depressed category of scores (mean BDI-II-O average total score of 23.41, SD = 9.13). Furthermore, there was a high rate of agreement among the mean scores of the individual clinicians, with each clinician rating the patient group in the same Moderately Depressed category. Therefore, it can be concluded that the clinicians in the present study saw the depression of the patient group as being less severe than did the patients, or similarly, the patients reported a higher level of depressive symptoms than did the clinicians for those patients.

Although the four clinicians agreed on the general categorization of the patient group in terms of depression severity, there was variation in the degree of correspondence between individual clinicians on the BDI-II-O total scores. Clinician 2 and Clinician 4 exhibited the greatest correspondence ($r = .85$). The common and discriminating factor between these two clinicians is an educational degree in psychology. The correspondence level between Clinician 1 and Clinician 2 ($r = .74$) and Clinician 4 ($r = .77$) was comparable, albeit at a slightly lower level than that exhibited between Clinician 2 and Clinician 4. Clinician 1 is an experienced social worker. The correspondence level between Clinician 3 and each of the other clinicians was consistently lower than the levels exhibited among the other clinicians ($r = .60$ with Clinician 1, $r = .66$ with Clinician 2, and $r = .58$ with Clinician 4). Clinician 3 demonstrated correspondence at a moderate level, but the other clinicians exhibited correspondence at a high level. Clinician 3 was a recent (6-month) graduate of a social work program and had the lowest number of contact hours with the patient group.

Interestingly, correspondence among the four clinicians was stronger when the patients rated themselves as mildly depressed. In other words, the relative absence of depression was easier for the clinicians to detect and agree upon than the presence of moderate to severe depression. This finding suggested that the patients in the present study did have an accurate view of their depression and were not overstating. It must be noted, however, that the number of patients who rated themselves as mildly depressed was small, and therefore, this observation may not be replicated in a larger study with a higher prevalence of mildly depressed patients. Nevertheless, and unfortunately, there is no criterion to determine accuracy. One theory suggests that patients are closest to their individual experience and are thus the best judges of their experience. Another theory suggests that clinicians have specialized training and a continuum of experience within which to rate a depressed individual, making them the more accurate raters. There will always be the question of who is right when comparing self-report and clinical ratings.

Accuracy issues aside, and with the recognition that there will always be error in both self-report and clinical ratings, it is now possible to examine the correspondence between the BDI-II and BDI-II-O total scores for this study.

Correspondence between Self-Reporter and Clinician Ratings

The correspondence between self-report and clinician ratings, based on total scores, was moderately high for both the clinician average and the individual clinicians.

The clinician average was r = .59 and the range for the individual clinicians was r = .42 to r = .62. As was the case on most measures throughout this study, Clinician 3 achieved lower levels of correspondence than did the other clinicians, albeit still in the low range of moderately high correspondence. These findings further confirm the fact that there was a difference between the patients' assessments of their depression and the clinicians' assessments; it is also true that the difference was present across multiple raters.

Although the overall degree of correspondence between self-reporter and clinician ratings based on total scores was moderately high, there were varying levels of correspondence on individual items.

Item Level Correlations between the BDI-II and BDI-II-O

There were four items that failed to achieve a significant level of correspondence between self-report and clinician ratings: Loss of Interest in Sex, Changes in Sleeping Pattern, Changes in Appetite, and Agitation. The first three of these items are somatic in nature and tend not to be as readily observable as other scale items, which may account for the low level of agreement on these items. The fourth item, Agitation, may have exhibited low agreement due to differences between patients and clinicians in the way agitation is experienced or interpreted. For example, patients may experience agitation as an internal state that is experienced as physiological hyperarousal; this is similar to a somatic symptom that is not entirely observable by others. In contrast, clinicians may interpret agitation as an affective state with expected and observable behavioral manifestations. Patients endorsed Loss of Interest in Sex at five times the rate of clinician

endorsement, Changes in Sleeping Pattern at almost two times the rate of clinician endorsement, Changes in Appetite at four times, and Agitation at almost three times the rate of clinicians.

Five items achieved a low level of correspondence, that is, a correlation of less than .40. These five items included Indecisiveness, Guilty Feelings, Pessimism, Loss of Interest, and Self-Dislike.

The remaining twelve items showed a moderate level of correspondence, with item correlations between .40 and .56. Three of the top four most highly correlated items were somatic disturbances and included Tiredness or Fatigue, Loss of Energy, and Concentration Difficulty. The somatic items constituting the bottom three of four items that showed weak correlations are assumed to be the result of low observability, unlike the somatic items at the high end of correlation which are more readily visible. The fourth, and most highly correlated item, was Self-Criticalness, a cognitive symptom embodying feelings of negative attitude toward the self. It should be noted that the symptom categories used in the above discussion correspond to the categories used by Beck et al. in their factor analysis. Future discussions will use categorization as deemed appropriate for the particular topic being examined.

Given the fact that there were variations in item level correspondence between the BDI-II and BDI-II-O, it is appropriate to consider the overall level of reliability of the two scales, using measures of internal consistency and test-retest stability.

Reliability of the BDI-II and BDI-II-O

Internal consistency, which is measured using Cronbach's coefficient statistic, is a term that refers to the observed pattern of item responses within a scale. This statistic addresses the question of how well individual items in a scale are correlated with one another, or more simply, the extent to which the items are related to one another. The stronger the items are interrelated, the more likely it is that the scale is consistent. A consistent scale is considered to be reliable. Evidence of a reliable scale suggests that the scale items are measuring the same construct. A high Cronbach's alpha indicates that response patterns are internally consistent and that the scale is reliable. Cronbach's alpha is a measure of the squared correlation between observed scores and true scores, or the ratio of true score variance to observed score variance. The higher the alpha statistic, the more reliable is the scale. An alpha of .70 and above is considered to be acceptable internal consistency, which is one measure of reliability.

Both the BDI-II and BDI-II-O scales exhibited excellent internal consistency in the present study, with alphas of .93 and .96, respectively. Both of these coefficient alphas were slightly higher than the coefficient alpha of .92 achieved by the BDI-II in the psychiatric outpatient sample used by Beck et al. (1996). The coefficient alphas for the individual clinician scores were .91 or higher for three of the four clinicians. The fourth clinician (Clinician 3) showed a lower, albeit still very acceptable, coefficient alpha of .85.

Internal consistency of the BDI-II and BDI-II-O subscales, based on grouping items according to the factor structure presented by Beck et al. for the psychiatric outpatient sample, was also excellent; however, there were some differences between the two instruments. The coefficient alphas for the two BDI-II subscales were comparable at

.87 for Factor 1 (Somatic/Affective) and .89 for Factor II (Cognitive). For the BDI-II-O, the coefficient alpha for Factor 2 (.97) was higher than that for Factor 1 (.91). For the individual clinicians, the coefficient alphas of the subscales were all in the acceptable range, with the exception of the alpha for Clinician 3 on Factor 1 which was .69 and fell just below the generally accepted .70 cutoff. These findings suggested that the patients were responding to a general factor of depression and rated their cognitive and somatic/affection symptoms in the same manner. Clinicians, in contrast, were more consistent on the cognitive symptoms than the somatic/affective symptoms. However, compared with the patients, they exhibited greater consistency on both symptom types.

Test-retest reliability is a measure of the stability of a scale over time, from one administration of a test to the next on the same group of people, generally within a short period of time. For the present study, the interval between the test and retest administrations was one week. Half of the total sample, or 18 patients, were retested a week after they completed the initial BDI-II. The four clinicians completed a second BDI-II-O on the same 18 patients one week after the completion of their first BDI-II-O.

The test-retest reliability coefficient for the BDI-II total score was .92, indicating very good temporal stability after one week. This reliability coefficient is consistent with that reported by Beck et al. (1996) on the BDI-II (.93) for a subsample of 26 psychiatric outpatients. Both of the BDI-II subscales, derived from Beck et al.'s factor analysis, also showed very good temporal stability. Factor 1 (Somatic/Affective) had a test-retest reliability coefficient of .87, while Factor 2 (Cognitive) had a test-retest reliability coefficient of .92.

The test-retest reliability coefficient for the BDI-II-O averaged total score was .88, indicating very good temporal stability for this instrument after one week. The test-retest coefficients for the individual clinicians were as follows: Clinician 1 ($r = .80$, $p < .05$), Clinician 2 ($r = .61$, $p < .01$), Clinician 3 ($r = .88$, $p < .01$), and Clinician 4 ($r = .76$, $p < .01$). Both of the BDI-II-O subscales, derived from Beck et al.'s factor analysis, also showed very good temporal stability. Factor 1 (Somatic/Affective) had a test-retest reliability coefficient of .77, and Factor 2 (Cognitive) had a test-retest reliability coefficient of .91. Comparing these results with those achieved on the BDI-II suggested that both clinicians and patients were less consistent over time in their ratings of somatic/affective items than they were in their ratings of cognitive items. This effect was more pronounced among the clinicians who appeared to have a comparatively difficult time rating some of the somatic/affective items. However, despite this finding, the BDI-II-O still demonstrated very good stability over time.

A separate analysis of test-retest stability compared the mean total scores between the two administrations of the BDI-II and BDI-II-O. For the BDI-II, there was no significant difference in the mean total scores of patients between administrations, as was the case in the Beck et al. (1996) study. For the BDI-II-O, there was also no significant difference in the mean BDI-II-O total scores of clinicians between administrations. This suggested that, on the basis of total scores, self-reporters and clinicians tended to rate depression severity within the same range from one week to the next, providing further evidence of the stability of both instruments.

The results of the reliability analyses described above indicated that both the BDI-II and BDIO-II-O provided highly consistent measures of the construct of depression, which were stable over time.

Up until this point, we have been comparing the BDI-II and BDI-II-O subscale results based on the factor groupings derived from Beck, Steer, and Brown's (1996) factor analysis. We will now examine the empirically derived factors based on the cluster analyses of the two instruments.

Cluster Analysis of the BDI-II and BDI-II-O

The cluster analysis of the BDI-II produced one cluster, Depressive Symptomatology, which included 20 of the 21 BDI-II items. The remaining item, Loss of Interest in Sex, did not adequately define a separate second cluster and was, therefore, excluded from further analysis. The fact that Loss of Interest in Sex was an isolated item suggested that self-reporters rated their level of sexual interest in a manner different from that with which they rated their other symptoms. An examination of this item revealed that, for about one-third of the patients in this sample, sexual interest did not appear to wane even when other depressive symptoms were seriously present. Results of this cluster analysis varied significantly from the factor analysis on the BDI-II by Beck et al. (1996), which produced two factors, a Somatic/Affective factor and a Cognitive factor as previously described. The cluster analysis of the BDI-II in the present study suggested that the sample did not differentiate between symptoms, experiencing their depression as one consistent set of associated symptoms.

The cluster analysis of the BDI-II-O, using the combined ratings of the four clinicians, produced two clusters. Although this is the same number of groupings produced by Beck et al. (1996) in their factor analysis of the BDI-II, the items included in each of the two clusters differed from the items included in the two factors.

Cluster 1, called the Cognitive/Performance cluster, consisted of fourteen items. These items described psychological symptoms related to self-reproach and performance impairment. High ratings on this dimension of depression would suggest that the clinician viewed the patient as having: (1) a negative assessment of their personal worth in multiple areas, and (2) performance difficulties related to cognitive abilities (concentration and decisiveness), physical resiliency (energy and alertness), and enjoyment of pleasurable activities (interest in individual pursuits and the company of others).

Cluster 2, called the Somatic/Affective cluster, consisted of seven items. These items included various symptoms related to somatic or affective disturbance, plus suicidal ideation. High ratings on this dimension of depression would suggest that the clinician viewed the patient as experiencing difficulties with appetite and sleep, feelings of agitation and irritation, loss of sexual interest, and episodes of crying and suicidality.

As stated above, both the cluster analysis of the BDI-II-O in the present study and the factor analysis of the BDI-II in Beck et al.'s study produced two separate and distinct item groupings, which varied from one another. A comparison of Cluster 1 (Cognitive/Performance) in the present study with Beck et al.'s Factor 2 (Cognitive) revealed that Cluster 1 shared all of the items in Factor 2 except suicidal ideation, adding, however, all items related to performance impairment. Cluster 2 (Somatic/Affective) in

the present study was strictly composed of somatic and affective items, unlike Factor 1 which contained items related to performance difficulty. It was interesting to note that the clinicians were evenly split about how they viewed the suicide item. The two clinicians with degrees in psychology rated this item with the somatic/affective items, but the two social workers rated this item with the cognitive/performance items.

Cluster membership for the individual clinicians differed, in some cases significantly (i.e., the cluster solutions for Clinician 2 and Clinician 3 varied from each other and from Clinician 1 and Clinician 4, who had similar cluster membership). The cluster analyses for both Clinicians 1 and Clinician 4 produced two clusters, called Cognitive/Performance (Cluster 1) and Somatic/Affective (Cluster 2), identical in name to the clusters for the average of the clinicians, but differing slightly in content between the two clinicians. The cluster solutions for these two clinicians shared all but three items in common (i.e., Punishment Feelings, Suicidal Thoughts or Wishes, and Changes in Sleeping Pattern), which did not change the overall interpretation of the cluster analysis results.

The cluster analysis for Clinician 3 also produced a two cluster solution but the affective items (i.e., Crying, Agitation, and Irritability) shifted to Cluster 1, called Cognitive/Affective/Performance, leaving only three items in the second cluster, called Somatic. The three somatic items were the hardest to observe, and included appetite and sleep changes as well as loss of sexual interest. It was interesting to note that the cluster solution for Clinician 3 was the most similar to that produced by the self-reporters. However, this fact did not seem to influence positively the correspondence between the total scores of Clinician 3 and the self-reporters.

The cluster analysis for Clinician 2 produced a three-cluster solution consisting of a Cluster 1, called Cognitive/Performance, a Cluster 2, called Affective, and a Cluster 3, called Somatic. This clinician appeared to view depressive symptoms in categories which were the most discrete of all the clinicians.

An overall examination of the separate clusters produced by the ratings of the individual clinicians showed that cluster membership was identical for 71% of the items. The greatest divergence in cluster membership between clinicians was on the following items: Punishment Feelings, Suicidal Thoughts or Wishes, and Changes in Sleeping Pattern. No particular pattern of divergence emerged among the items relative to cluster categories. It can be noted, however, that Punishment Feelings was judged as the most difficult item to rate by respondents completing the trait ratability survey.

It is now feasible to examine the empirically derived BDI-II and BDI-II-O clusters in comparison with various other measures and with each other.

Relationships between Empirical Results and Derived Results

The BDI-II cluster, Depressive Symptomatology, was strongly correlated with both the Cognitive factor (.95) and Somatic/Affective factor (.96) derived from grouping items according to Beck et al.'s (1996) factor solution for the BDI-II. This suggests that, for the sample in the present study, there is an overarching second order factor, Depressive Symptomatology, which is composed of two first order factors, a Cognitive factor, and a Somatic/Affective factor.

The BDI-II cluster, Depressive Symptomatology, was moderately correlated with the BDI-II-O Cognitive/Performance Cluster 1 (.56) and the BDIO-II-O Somatic/Affective Cluster 2 (.61), based on averaged clinician ratings. This result further substantiated the moderate level of correspondence found between self-report and clinician ratings based on the item scores for each instrument. It also suggested that the elimination of the Loss of Interest in Sex item did not materially affect correspondence levels.

The two BDIO-II-O clusters were strongly correlated with each other at .79, suggesting that they share a relatively high level of common variance. This correlation coefficient was higher than the .66 correlation coefficient found between the factors in Beck et al.'s study sample. This indicated that the factors found by Beck et al. were somewhat less interrelated in that sample versus the present sample, making the factors slightly more distinct from one another. Said another way, items in the clusters were viewed as being more similar, and less discrete. Given the fact that the clusters of the BDI-II-O were expected to mirror the different dimensions of depression found by Beck et al., it was predicted that the correlation between the factors would be in the moderate range. Instead, they were found to be in the high range.

Given the fact that we have identified empirical clusters for the BDI-II and BDI-II-O, it is appropriate to consider the internal consistency of the clusters.

Internal Consistency of the BDI-II and BDI-II-O Clusters

The internal consistency of the Depressive Symptomatology cluster for the BDI-II was strong, having a coefficient alpha of .93, which is identical to that of the total scale. This again suggested that the removal of Loss of Interest in Sex did not materially affect the internal consistency of the scale in the present study.

For the BDI-II-O, the internal consistency of Cluster 1 (Cognitive/Performance) was .97, and equal to the internal consistency reported previously for the derived cognitive factor. The internal consistency of Cluster 2 (Somatic/Affective) was .81, and lower than the .91 for the derived somatic/affective factor. This variation was due to the different composition of the clusters versus the derived factors and the fact that clinicians had relatively more difficulty rating the somatic/affective components of depression, as opposed those of a cognitive and performance nature.

Results of the Higher Order Factor Analysis

In order to assess the construct validity of the BDI-II-O, a higher order factor analysis was performed on the combined subscales of the BDI-II and BDI-II-O, using the derived BDI-II subscales and the empirical clusters. This analysis produced one factor which accounted for 73.2% of the variance in the combined scales. The correlation between the empirical BDI-II factor, Depressive Symptomatology, and the higher order factor was strong (.93), as were the correlations between the empirical clusters and the higher order factor (Cluster 1 = .81, Cluster 2 = .79). The internal consistency of the higher order factor was high at .83. These results suggested that there was a high correspondence level between self-reporter and clinician ratings on the overarching,

common dimension of depression, which helps to substantiate the construct validity of the BDI-II-O.

Interrater Agreement

Clinicians in this study showed moderate to high levels of agreement between themselves based on the paired correlations of individual clinician ratings on the BDI-II-O. This result was substantiated by a high intraclass correlation of .88, which measured the average correlation for all pairs of clinicians. It is important to note that the average single measure correlation was only .65, and below the acceptable range for use of a single rater. This finding suggested that there is benefit to having a number of different raters providing information on a patient; it also speaks to the clinical utility of the BDI-II-O in this regard, because it can be used by a wide variety of informants.

As a further measure of clinician agreement, simple agreement levels were examined. This measure evaluated the proportion of cases in which various combinations of raters agreed, expressed as a percentage by item. Results of this analysis showed the following simple agreement rates: no raters agreed on 2.7% of the cases, two raters agreed on 41.7% of the cases, three raters agreed on 38.9% of the cases, and four raters agreed on 16.7% of the cases. Items with the highest level of four-rater agreement were Sadness, Crying, Irritability, and Loss of Interest in Sex. The first three of these four items were moderately correlated with the matching BDI-II score (and speculated to be fairly easy to rate); however, the fourth item showed no significant correlation. On the basis of clinician total scores, there was no agreement on 75% of the cases, and two-rater

agreement on 25% of the cases. There were no cases in which three or four raters agreed on the total scores.

The simple agreement rates suggested that there is a fair amount of inconsistency in individual item ratings, and more so in the total scores. This finding also supports the notion that it is preferable to have more raters, and that an observer rated instrument to complement the BDI-II would probably have particular clinical utility if a combination of raters, such as a therapist, family member, and friend simultaneously rated the patient.

Trait Ratability

It was hypothesized that items with low trait ratability would show the least amount of self-reporter and clinician agreement. However, in the present study, there was practically no relationship between self-reporter and clinician agreement based on the results of the trait ratability survey. The only two exceptions were for the Sadness and Crying items, which were judged as the two easiest traits to rate. These items showed a moderate level of self-reporter and clinician agreement, albeit at a lower rate of correspondence than some other items judged more difficult to rate. Despite the fact that the survey failed to detect differences materially in item correspondence levels attributable to trait ratability, it was apparent in other parts of the study analysis that clinicians did have more difficulty rating some items which could legitimately be assessed as low versus high visibility traits (e.g., loss of interest in sex and changes in appetite or sleep).

The Influence of Patient/Clinician Interaction on Levels of Agreement

Four dimensions of patient/clinician interaction were studied to assess their impact on agreement levels between self-reporter and clinician total scores: Acquaintanceship, Contact Hours, Familiarity, and Primary/Not Primary Therapist.

Acquaintanceship was defined as the length of time a patient and clinician had known each other. Results from the Acquaintanceship analysis suggested that the degree of correspondence between self-reporter and clinician ratings did not increase with greater acquaintanceship. In fact, correspondence levels were higher for patients who had been in the program less than six months versus those who had been in the program more than six months. A possible explanation for this finding is that patients with a shorter duration of stay may have exhibited more obvious symptomatology, which increased the clinician's ability to assess the patient's depression better. The significance of this finding was less so for Clinician 3 compared with the other clinicians.

Mean total scores for the BDI and BDIO-II-O were calculated based on length of acquaintanceship and suggested that self-report and clinician depression severity ratings were not influenced by acquaintanceship. Specifically, self-reporters who had been in the program longer tended to rate themselves with the same mean level of depression severity as those who had been in the program a shorter amount of time. In addition, clinicians tended to rate patients with the same mean level of depression severity regardless of length of acquaintanceship.

Contact Hours was defined as the amount of time per week that a clinician spent with a patient, including time spent in individual and group therapy, milieu therapy, and

telephone consultation. Results of the Contact Hours analysis suggested that the degree of correspondence between self-report and clinician ratings increased with contact hours; however, this finding was not significant for Clinician 3.

Mean total scores for the BDI-II and BDIO-II-O were calculated based on number of contact hours, and suggested that self-report and clinician depression severity ratings were related to contact hours. Depression severity ratings for both self-reporters and clinicians were higher when contact hours were greater. However, this finding may simply be a reflection of differences in attendance rates among patients, because it is likely that the more severely depressed patients attend the program for a greater number of hours per week.

Familiarity was defined as a measure of how well rather than how long the clinician knew the patient. It was interesting to note that familiarity ratings were lower than expected, with no clinicians reporting that they knew the patients extremely well. Results of the Familiarity analysis showed that the degree of correspondence between self-reporter and clinician ratings increased with familiarity, although this finding was not significant for Clinician 3.

Mean total scores for the BDI and BDIO-II-O were calculated based on familiarity ratings, and suggested that self-reporter and clinician depression severity ratings were not substantially influenced by the degree of familiarity between the self-reporter and clinician. This suggested that clinicians were rating the patients' depression based on criteria other than familiarity; this would be appropriate.

Primary/Not Primary Therapist referred to the role of the clinician in relation to the patient, and is self-explanatory. Results of this analysis showed that there was no

difference in degree of correspondence between self-report and clinician ratings based on the role of the therapist. However, it must be noted that the correlation coefficient produced from the averaged clinician ratings, upon which this conclusion is based, was highly influenced by the performance of one clinician and can be somewhat misleading. Specifically, Clinician 4 demonstrated significant correspondence with the patient as both a primary and non-primary therapist, although the other clinicians demonstrated significant correspondence only in the role of non-primary therapist. Clinician 3 demonstrated the lowest level of correspondence in both roles.

Treatment Implications

Accurate assessment of patient problems constitutes critical input to the treatment process. The existence of discordant views among patient and staff, or among staff members, relative to the patient's severity of depressive symptoms, and/or the specific categories of symptoms pertaining to the patient, has negative implication for developing effective treatment planning and producing a successful outcome. Illuminating discordant views through the use of a common measurement system, such as provided by the BDI-II and BDI-II-O, offers the following opportunities and benefits: (a) consensual validation of the patient's depressive symptomatology among staff members, (b) development of a common case conceptualization of the patient, (c) identification of appropriate treatment planning goals, (d) a more focused treatment plan, (e) identification of discrepant perspectives relative to specific symptoms measured by the common instruments, (f) facilitation of communication between staff and patient relative to these differing

perspectives, (g) enhanced agenda setting for therapy sessions, and (h) a closer therapeutic alliance. With respect to this last point, the merits of a strong therapeutic alliance have been abundantly described in the literature and will not be further elaborated here, other than to suggest the proposition that comparing and sharing the results of the patient's BDI-II and BDI-II-O with the patient may provide a unique tool for building rapport, as well as providing a springboard for further discussion of depressive symptoms from two perspectives. In turn, a common vision of the patient's problems could be developed. This, combined with all of the other factors above, supports the likelihood of better treatment and outcome.

Nezu, Ronan, Meadows, and McClure (2002), cited additional benefits of a brief, standardized depression instrument, such as the BDI-II-O for clinicians. First, by using a standardized format, the reliability of serial assessments over time is enhanced. In addition, use of a standardized clinician rated instrument ensures a complete assessment that includes all of the various symptoms of depression so that none are inadvertently overlooked. Lastly, a standardized clinician rated instrument may be preferred when comordid disorders may reduce the reliability of, or may interfere with, the completion of self-reports. For example, Addington, Addington, and Matick-Tyndale (1993) found convergent results in a study of psychiatric inpatients with schizophrenia using the BDI-II and a clinician rated scale, the Calgary Depression Scale; however, one third of the patients had difficulty completing the BDI-II and 6% were unable to complete it at all. This finding suggests that the BDI-II-O would have the further utility of stand-alone capability in circumstances in which patients can not provide self-report evaluations.

Limitations of the Study

The scope of this study was admittedly limited in a number of important areas. The study was conducted among a sample of patients attending an intensive outpatient/partial hospitalization program in only one facility and, therefore the sample may not be representative of the entire population of patients in this type of program. The study was conducted in one type of treatment setting and may not be generalizable to other treatment settings such as inpatient programs and outpatient clinics. The study was conducted in one regional location and may not be representative of results that might be achieved in other geographic areas. The study was conducted among adults and may not be generalizable to adolescent populations. And lastly, the study was conducted using one group of clinicians who did not represent the total spectrum of clinicians relative to professional discipline, years of experience, or gender.

The study did not attempt to assess whether clinicians or patients were the more accurate raters, given the fact that there is no existing criterion for accuracy. Despite this limitation, it is believed that there is still merit in using parallel versions of the Beck Depression Inventory due to the inherent value in identifying disparate ratings for purposes of more consensual treatment planning. Determining who is "right" is perhaps not as critical as formulating a common perspective on the goals and tasks of therapy.

The study assessed correspondence levels between patients and clinicians from a cognitive-behavioral theoretical orientation, given the fact that this is most widely researched and accepted approach to treatment at the present time. The study did not

examine correspondence levels that might be achieved between other existing, or yet to be developed, instruments for depression based on alternative theoretical orientations.

Recommendations for Future Research

The present study examined the reliability and validity of the BDI-II-O in comparison with the standard self-reporter version of the Beck Depression Inventory in a sample of adult psychiatric patients. The focus of the study was to begin establishing reliability and validity evidence for this new instrument. Given that initial psychometric results were positive, there are a variety of opportunities to extend this research into other areas. For example, additional psychometric measures could be examined, such as convergent validity with the updated version of the clinician rated HDRS. In addition, contrasted groups validity, that is, examining correspondence rates using a sample of patients with varying degrees of depression, non-depressed patients, or patients with other psychiatric disorders, could help to determine the ability of a clinician rated version to be specific to depression as opposed to generalized distress or negative affectivity.

There is a broad opportunity to extend the study using other groups of clinicians or mental health professionals (e.g., licensed psychologists, psychiatrists, nurses), other observers (e.g., spouses, friends, relatives, significant others), and other settings (e.g., hospitals, outpatient offices). Of particular interest, to assess agreement levels in a greater degree between different clinicians, one might conduct a study in which the setting was an inpatient unit, the clinicians were nurses, and patients were rated during each shift

over a 24 hour period. This study would allow for an assessment of whether or not clinician ratings vary depending on the patient's presentation over a short period of time.

Given the fact that the clinician group in this study had an average experience level on the low end of the experience spectrum, it might be interesting to include clinicians with more widely varying levels of experience in the next study to see if differences in agreement levels are detected.

The relatively small sample size in this study prevented a replication of Beck's principal axis factor analysis with orthogonal rotation. Therefore, a goal of future studies might be to obtain a large enough sample to conduct a comparative factor analysis, similar in composition to that described in this study, using clusters as the basis of comparison.

Future studies may want to examine in more depth the influence of demographic and/or sample characteristics on correspondence levels (e.g., using fewer chronic patients, using a patient group profile that is more consistent with general population statistics, using clinicians of both genders). The influence of other characteristics of both the patient and clinician groups may be relevant to understanding how correspondence levels vary.

Given the fact that depression is a large risk factor for suicide, and the fact that clinicians did not endorse the Suicidal Thoughts and Wishes item as frequently nor with the same intensity as patients did, future studies may want to examine the reasons for the disparity in ratings between the two groups. The finding of low clinician ratings in this study was unusual, given that clinicians generally err on the conservative side relative to the assessment of suicidality; in addition, patients in this study were known to have

previously received a high level of repeated inpatient hospitalizations, presumably associated with suicidal ideation. The finding of very high patient ratings regarding suicidality could also be considered curious, given that severely depressed patients, such as those found in this study, typically have neither the awareness of their suicidal risk, nor the energy to plan and carry out a suicidal act.

A more extensive analysis of the convergence of ratings among multiple observers would help to clarify the optimal number of raters to use in assessing a patient. This analysis would examine correspondence as a function of number of raters, and would demonstrate whether or not agreement levels increase as ratings from additional observers are included.

Regarding the trait ratability survey, which failed to detect differences materially in patient/clinician agreement levels, it may be worthwhile in future studies to replicate this investigation using a group of independent judges who have a consistently high level of expertise and training in the field of psychology. The group of independent judges used in this study consisted of a mix of highly qualified professionals from psychology and medicine, as well as advanced psychology students who were still in training; the combination of these may have confounded the overall results. In addition, future trait ratability surveys should add language in the instructions which specifies a time period for ratings, that is, “ratings should be based on how hard or easy it would be to rate the patient’s symptoms during a two week period”. In this way, the specified timeframe would be consistent with the BDI-II, and would obviate any error in ratings due to a possible assumption among judges that ratings should be based on a specific moment in time.

Given that clinicians in this study had a relatively more difficult time rating the somatic/ affective items than the cognitive/performance items as detected by lower correspondence levels on the former, there may be an opportunity to use a shorter version of the BDI-II-O containing fewer items. This approach has already been implemented in the medical setting with the Beck Depression Inventory – Primary Care (BDI – PC), which contains just seven items.

As a final opportunity for future study, it would no doubt be interesting and relevant to conduct follow-up research with clinicians after they have discussed rating disparities with the patient, in order to determine the usefulness of the discussions for treatment planning purposes in a real world setting. This investigation would assist in further clarifying the clinical utility of the BDI-O-II.

References

Addington, D., Addington, J., & Matika-Tyndale, E. (1993). Rating depression in schizophrenia. A comparison of a self-report and an observer report scale. *Journal of Nervous and Mental Disorders, 181(9),* 561-565.

Albright, L., Kenny, D.A., & Malloy, T. E. (1988). Consensus in personality judgements at zero acquaintance. *Journal of Personality and Social Psychology, 55,* 387-395.

American Psychiatric Association. (1980). *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed.). Washington, DC: Author

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.). Washington, DC: Author

American Psychological Association. (1992). *Ethical Principles.* Washington, DC: Author.

Bailey, J. & Coppen, A. (1976). A comparison between the Hamilton rating scale and the Beck inventory in the measurement of depression. *British Journal of Psychiatry, 128,* 486-489.

Beck, A. T. (1987). Cognitive models of depression. *Journal of Cognitive Psychotherapy, 1,* 2-27.

Beck, A. T., & Steer, R. A. (1987). *Beck depression inventory manual.* San Antonio, TX: The Psychological Corporation.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression.* New York: The Guilford Press.

Beck, A. T., Steer, R. A., Garbin, M. G., (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8,* 77-100.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory* (2nd ed). San Antonio, TX: The Psychological Corporation.

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of Beck Depression Inventories –IA and –II in psychiatric outpatients. *Journal of Personality Assessment, 67(3),* 588-597.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 53-63.

Beckham, E. E., & Leber, W. R. (1995). *Handbook of depression.* New York: The Guilford Press.

Brown, C., Schulberg, H. C., & Madonia, M. J. (1995). Assessing depression in primary care practice with the Beck Depression Inventory and the Hamilton Rating Scale for Depression. *Psychological Assessment, 7(1),* 59-65.

Carroll, B. J., Feinberg, M., Smouse, P. E., Rawson, S. G., & Greden, J. F. (1981). The Carroll Rating Scale for Depression: I. Development, reliability, and validation. *British Journal of Psychiatry, 138,* 194-200.

Chiles, J. A., & Strosahl, K. D. (1995). *The suicidal patient: principles of assessment, treatment, and case management* (pp. 50-105). Washington, DC: American Psychiatric Press.

Cicchetti, D. V., & Prushoff, B. A. (1983). Reliability of depression and associated clinical symptoms. *Archives of General Psychiatry, 40,* 987-990.

Clark, L. A. & Watson, D. (1991). Theoretical and empirical issues in differentiating anxiety from depression. In J. Becker, & A. J. Kleinman (Eds.), *Psychosocial aspects of mood disorder* (pp. 39-65*)*. Hillsdale, NJ: Erlbaum.

Craven, J. L., Rodin, G. M., & Littlefield, C. (1988). The Beck Depression Inventory as a screening device for major depression in renal dialysis patients. *International Journal of Psychiatry in Medicine, 18,* 365-374.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Davies, B., Burrows, & G. Poynton, N. (1975) A comparison study of four depression rating scales. *Australia and New Zealand Journal of Psychiatry, 9,* 21-24.

Derogatis, L. R. & Lynn, L. L. (1999) Psychological tests in screening for psychiatric disorder. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed., pp. 41-79). Mahwah, NJ: Lawrence Erlbaum Associates.

Dozois, D. J., Dobson, K. S., & Ahnberg, J. L. (1998). A psychometric evaluation of the Beck Depression Inventory-II. *Psychological Assessment, 10(2),* 83-89.

Dozois, D. J. & Dobson, K. S. (2002). Depression. In M. M. Antony, & D. H. Barlow (Eds.), *Handbook of assessment and treatment planning for psychological disorders* (pp. 259-291). New York, NY. The Guilford Press.

Domken, M., Scott, J., & Kelly, P. (1994). What factors predict discrepancies between self and observer ratings of depression? *Journal of Affective Disorders, 31,* 253-259.

Eddicott, J., & Spitzer, R. L. (1978). A diagnostic interview: The Schedule for Affective Disorders and Schizophrenia. *Archives of General Psychiatry, 35,* 837-844.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., et al. (1989). National Institute of Mental Health treatment of depression collaborative research program. *Archives of General Psychiatry, 46,* 971-983.

Enns, M. W., Larsen, M. A., & Cox, B. J. (2000). Discrepancies between self and observer ratings of depression: The relationship to demographic, clinical, and personality variables. *Journal of Affective Disorders, 60,* 33-41.

Faustman, W. O., & Overall, J. E. (1999). Brief Psychiatric Rating Scale. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed., pp. 791-830). Mahwah, NJ: Erlbaum.

Finn, S. E., & Tonsager, M. E. (1992). Therapeutic effects of providing MMPI-2 test feedback to college students awaiting therapy. *Psychological Assessment, 4,* 278-287.

First, M. B., Gibbon, M., Spitzer, R. L., & Williams. J. B. W. (1996). *User's Guide for the Structured Clinical Interview for DSM-IV Axis I Disorders – Research Version (SCID-I, Version 2.0, February, 1996, Final Version).* New York: Biometrics Research.

First, M. B., Spitzer, R. L.,Gibbon, M., & Williams. J. B. W. (1997). *Structured Clinical Interview for DSM-IV Axis I Disorders – Clinician Version (SCID-CV).* Washington, DC: American Psychiatric Press.

Funder, D. C. & Colvin, C. R. (1988). Friends and strangers: acquaintanceship, agreement, and the accuracy of personality judgement. *Journal of Personality and Social Psychology, 55,* 149-158.

Funder, D. C. & Dobroth, K. M. (1987). Differences between traits: properties associated with interjudge agreement. *Journal of Personality and Social Psychology, 52,* 409-418.

Gabrys, J. B, & Peters, K. (1985). Reliability, discriminant and predictive validity of the Zung Self-Rating Depression Scale. *Psychological Reports, 57,* 1091-1096.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology and Neurosurgical Psychiatry, 23,* 56-62.

Hamilton, M. (1967). Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology, 6,* 278-296.

Hays, R. D., Wells, K. B., Sherbourne, C. D., Rogers, W., & Spritzer, K. (1995). Functioning and well-being outcomes of patients with depression compared to chronic general medical illnesses. *Archives of General Psychiatry, 52,* 11-19.

Hedlund, J., & Vieweg, B. (1979). The Hamilton Rating Scale for Depression: A comprehensive review. *Journal of Operational Psychiatry, 10,* 149-162.

Hesselbrock, M., Easton, C., Buchotz, K. K., Schuckit, M., & Hesselbrock, V. (1999). A validity study of the SSAGA – a comparison with the SCAN. *Addiction, 94(9),* 1361-1370.

Hollon, S. D., DeRubeis, R. J., & Seligman, M. E. (1992). Cognitive therapy and the prevention of depression. *Applied and Preventive Psychology, 1,* 89-95.

Hooijer, C., Zitman, F. G., Griez, E., van Tilburg, W., Willemse, A., & Dinkgreve, M. A. (1991). The Hamilton Depression Rating Scale (HDRS): Changes in scores as a function of training and version used. *Journal of Affective Disorders, 22,* 21-29.

Kaelber, C. T., Moul, D. E., & Farmer, M. (1995). Epidemiology of depression. In E. E. Beckam, & W. R. Leber (Eds.), *Handbook of depression* (pp. 3-35). New York: The Guilford Press.

Kagan, J. (1988). The meanings of personality predicates. *American Psychologist, 43,* 616-620.

Kammann, R., Smith, R. Martin, C. & McQueen, M. (1984). Low accuracy in judgements of others' psychological well-being as seen from a phenomenological perspective. *Journal of Personality, 52,* 107-123.

Katz, R., Shaw, B. F., Vallis, T. M., & Kaiser, A. S. (1995). The assessment of severity and symptom patterns in depression. In E. E. Beckam, & W. R. Leber (Eds.), *Handbook of depression* (pp. 61-85). New York, NY: The Guilford Press.

Kenrick, D. T. & Stringfield, D. O. (1980). Personality traits and the eye of the beholder: crossing some traditional philosophical boundaries in the search for consistency in all of the people. *Psychological Review, 87,* 88-104.

Kessler, R. C., McGonagle, K. A., Swartz, M., Blazer, D. G., & Nelson, C. B. (1993). Sex and depression in the National Comorbidity Survey: I. Lifetime prevalence, chronicity, and recurrence. *Journal of Affective Disorders, 29,* 85-96.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M, Eshleman, S., et al. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. *Archives of General Psychiatry, 51,* 8-19.

Knight, R. G., Waal-Manning, H. J., & Spears, G. F. (1983). Some norms and reliability data for the State-Trait Anxiety Inventory and the Zung Self-Rating Depression Scale. *British Journal of Psychiatry, 131,* 49-52.

Kobak, K. A., & Reynolds, W. M. (1999). The Hamilton Depression Inventory. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment.* (935-969) Mahwah, NJ: Lawrence Erlbaum Associates.

Kobak, K. A., & Reynolds, W. M. (2000). The Hamilton Depression Inventory. In M. E. Maruish (Ed.), *Handbook of psychological assessment in primary care settings* (pp. 423-461). Mahwah, NJ: Lawrence Erlbaum Associates.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* (monograph no. 9), 635-694.

Maruish, M. (1999). Introduction. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment, 2nd Edition*, (pp. 1-39). Mahwah, NJ: Lawrence Erlbaum Associates.

May, A. E., Urguhart, A., & Tarran, J. (1969). Self-evaluation of depression in various diagnostic and therapeutic groups. *Archives of General Psychiatry, 21,* 191-194.

McCrae, R. R. & Costa, P. T. (1987). Validation of a five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 57,* 691-706.

Moran, P. W. & Lambert, M. J. (1983). A review of current assessment tools for monitoring changes in depression. In: M. S. Lambert, E. R. Christensen, S. S. Dejulio (Eds.), *The assessment of psychotherapy outcome.* New York: Wiley

Nezu, A. M., Ronan, G. F., Meadows, E. A., & McClure, K. S. (2000). *Clinical assessment series: Vol. I. Practitioner's guide to empirically-based measures of depression.* New York: Kluwer/Plenum.

Norman, W. T. & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology, 4,* 681-691.

Overall, J. E., & Gorman, D. R. (1962). The Brief Psychiatric Rating Scale. *Psychological Reports, 10,* 799-812.

Rabkin, J. G., & Klein, D. F. (1987). The clinical measurement of depressive disorders. In A. J. Marsella & R. M. A. Hirschfeld (Eds.), *The measurement of depression* (pp. 30-83). New York: The Guilford Press.

Reynolds, W. M., & Kobak, K. A. (1995). Development and validation of the Hamilton Depression Inventory: a self-report version of the Hamilton Depression Rating Scale. *Psychological Assessment, 7,* 472-483.

Riskind, J. H., Beck, A. T., Berchick, R. J., Brown, G., & Steer, R. A. (1987). Reliability of DSM-III diagnoses for major depression and generalized anxiety disorder using the Structured Clinical Interview for DSM-III. *Archives of General Psychiatry, 44,* 817-820.

Sayer, N. A., Sackeim, H. A., Moeller, J. R., Prudic, J., Devanand, D. P., Coleman, E. A., et al. (1993). The relations between observer rating and self-report of depressive symptomatology. *Psychological Assessment, 5(3),* 350-360.

Segal, D. L., Hersen, M., Van Hasselt, V. B. (1994). Reliability of the Structured Clinical Interview of DSM-III-R: An evaluative review. *Comprehensive Psychiatry, 35(4),* 316-327.

Schnurr, R., Hoaken, R. C. S., & Jarrett, F. J. (1976). Comparison of depression inventories in a clinical population. *Canadian Psychiatric Association Journal, 21,* 473-476.

Schotte, C. K. W., Maes, M., Cluydts, R., & Cosyns, P. (1996). Effects of affective-semantic mode of item presentation in balanced self-report scales: Biased construct validity of the Zung Self-Rating Depression Scale. *Psychological Medicine, 26,* 1161-1168.

Schotte, C. K. W., Maes, M., Cluydts, R., DeDoncker, D. & Cosyns, P. (1997). Construct validity of the Beck Depression Inventory in a depressive population. *Journal of Affective Disorders, 46,* 115-125.

Simon, R., Endicott, J., & Nee, J. (1987). Intake diagnoses: How representative? *Comprehensive Psychiatry, 28,* 389-396.

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

Steer, R. A., Beck, A. T., Riskind, J. H., & Brown, G. (1987). Relationship between the Beck Depression Inventory and the Hamilton Rating Scale for Depression in depressed outpatients. *Journal of Psychopathology and Behavioral Assessment, 9(3),* 327-339.

Steiner, J. S., Tebes, J. K., Sledge, W. H., & Walker, M. L. (1995). A comparison of the Structured Clinical Interview for DSM-III-R and clinical diagnoses. *The Journal of Nervous and Mental Disease, 183(6),* 365-369.

Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria. *Archives of General Psychiatry, 35,* 773-782.

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

Watson, D. (1989). Strangers' ratings of the five robust personality factors: evidence of a surprising convergence with self-report. *Journal of Personality and Social Psychology, 57,* 120-128.

Watson, D., & Clark, L. A. (1991). Self- versus peer ratings of specific emotional traits evidence of convergent and discriminant validity. *Journal of Personality and Social Psychology, 60,* 927-940).

Westfield, J. S., & Liddell, D. L. (1994). The Beck Depression Inventory and its relationship to college student suicide. *Journal of College Student Development, 35,* 145-146.

Whitaker, A., Johnson, J., Shaffer, D., Rapoport, J., Kalikow, K., Walsh, B., et al. (1990). Uncommon trouble in young people: prevalence estimates of selected psychiatric disorders in a non-referred adolescent population. *Archives of General Psychiatry, 47,* 487-496.

Williams, J. B. W., Gibbon, M., First, M. B., Spitzer, R. L., Davies, M., Borus, J. Howes, M. J., Kane, J., Pope, H. G., Rounsaville, B., Wittchen, H. (1992). The Structured

Clinical Interview for DSM-III-R (SCID), II. Multisite test-retest reliability. *Archives of General Psychiatry, 49,* 630-636.

World Health Organization, World Bank, & Harvard University (1996). In C. L. Murray, & A. D. Lopez (Eds.), *The global burden of disease: A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020.* Cambridge, MA: Harvard University Press.

Zimmerman, M., Coryell, W., Corenthal, C., & Wilson, S. (1986). A self-report scale to diagnose major depressive disorder. *Archives of General Psychiatry, 43,* 1076-1081.

Zung, W. W. K. (1965). A self-rating depression scale. *Archives of General Psychology, 12,*63-70.

Appendix A: Permission Letter for the BDI-II-O

Appendix B: Patient Demographic Form

DEMOGRAPHICS – PATIENT _____

Please answer the questions below by checking the appropriate box or filling in the blank line. Please print legibly so we can read your answers. These questions are being asked so that we will be able to describe the characteristics of the study group as a whole.

Gender

K Male
K Female

Race

K Caucasian
K African American
K Hispanic
K Asian
K Native American
K Other (specify) _____

Education

K Less than High School
K High School or GED
K Associates degree
K Some college
K College graduate (BS/BA)
K Master's/Professional degree
K M.D./D.O./Other Doctorate

Age

Enter your age _____

Marital Status

K Married
K Never married
K Separated
K Divorced
K Widowed

Psychotropic Medications (please list)

_____
_____
_____
_____
_____
_____
_____
_____
_____

Mental Health Treatment History

□ Inpatient hospital # of times ____
□ Partial program # of times ____
□ Outpatient therapy # of times ____
□ Supported living # of times ____
□ Self-help groups # of times ____
□ Other (specify) _____

Drug and Alcohol Treatment History

□ Inpatient hospital # of times ____
□ Residential program # of times ____
□ Halfway house # of times ____
□ Partial program # of times ____
□ AA/NA # of times ____

Age when you received mental health services for the first time

_____

Appendix C: Clinician Demographic Form

DEMOGRAPHICS – CLINICIAN _____

Please answer the questions below by checking the appropriate box or filling in the blank line. Please print legibly so we can read your answers. These questions are being asked so that we will be able to describe the study group as a whole on demographics variables.

Gender

K Male
K Female

Race

K Caucasian
K African American
K Hispanic
K Asian
K Native American
K Other (specify) _____

Education

K Less than High School
K High School or GED
K Associates degree
K Some College
K College graduate (BA/BS)
K Master's/Professional degree
K M.D./D.O./Other Doctorate

Professional Discipline

K Social Work (MSW/LSW)
K Psychology (BA/BS)
K Psychology (MS)
K Psychiatrist (MD/DO)

Experience Post-Degree

Enter number of years _____ and months _____

Age

Enter your age _____

Appendix D: Clinician/Patient Information Form

CLINICIAN/PATIENT INFORMATION FORM

Clinician's First Name: _____

Patient's First Name/Last Initial _____

Patient's Admission Date _____

Please answer the questions below by checking the appropriate box or filling in the blank line.

How long have you known this patient?

Number of months _____

How many hours per week do you spend with this patient? (If you only see the patient once per month for an hour or less, please check "less than 1")

K Less than 1
K 1 hour
K 2 hours
K 3 hours
K 4 hours
K 5 hours
K 6 hours
K 7 hours
K 8 hours
K 9 hours
K 10 hours
K 11 hours
K 12 hours
K 13 hours
K 14 hours
K 15 hours
K 16 hours
K 17 hours
K 18 hours
K Greater than 18 hours (specify) _____

How well do you know this patient? (Circle your response with 1 being hardly at all and 7 being extremely well)

Hardly at all						Extremely Well
1	2	3	4	5	6	7

Are you this patient's individual therapist? (i.e., you provide individual psychotherapy)

K Yes
K No

Appendix E: Patient Informed Consent Form

INFORMED CONSENT FORM
PATIENTS

TITLE OF STUDY

Correspondence of Self and Observer Ratings of Depression using the BDI-II and BDI-II-O

PURPOSE

The purpose of this research is to find out how you view your feelings of depression compared to how others might see your depression.

You are being asked to be in this research study because you are depressed or you have other mental health problems that result in feelings of depression. If your therapist feels that it would be better for you not to be in this study, then you can not be in the study.

INVESTIGATOR(S)

Principal Investigator

Name: Dr. Arthur Freeman, Ed.D.

Department: PCOM, Department of Psychology

Address: 4190 City Avenue
Philadelphia, Pa. 19131

Phone: 215 871-6456

Responsible Investigator

Name: Leah Longan, M.S., Psy.D. Candidate

Department: PCOM, Department of Psychology
Address: 4190 City Avenue
Philadelphia, Pa. 19131

Phone: 717 871-6442

The doctors and scientists at Philadelphia College of Osteopathic Medicine (PCOM) do research on diseases, new treatments, and psychological issues. The study you are being asked to volunteer for is part of a research project on depression.

Even though this research project is to study depression, no one can say whether the results will result in better treatment.

If you have any questions about this research, you can call Dr. Arthur Freeman at (215) 871-6456.

If you have any questions or problems during the study, you can ask Dr. Freeman, who will be available during the entire study. If you want to know more about Dr. Freeman's background, or the rights of research subjects, you can call Dr. John Simelaro, Chairperson, PCOM Institutional Review Board at (215) 871-6337.

DESCRIPTION OF THE PROCEDURES

If you agree to be a part of this study, you will be asked to fill out the Beck Depression Inventory – II (BDI-II). This is a questionnaire that asks about your feelings, thoughts, and behaviors that have to do with depression. You will also be asked to fill out a form that asks for your age, education level, race, gender, and marital status. You may also be asked to fill out the BDI-II a second time, one week later.

You will fill out the questionnaire and information form during the course of your normal day at The Keystone Center. It will take about 10 minutes to complete. There is nothing you need to do to prepare for participation in this research project.

You will be asked to put your first name and last initial on the questionnaire, as well as on the form that asks for personal information as described above. The reason we need you to use your name is so that we can match your answers with those from similar questionnaires that will be filled out by the professional staff at The Keystone Center. Program staff will not have access to your questionnaires unless you decide to discuss your answers with your individual therapist. If you decide to share your answers with your therapist, copies of your forms will be given to the therapist, who will keep them in a locked drawer. You will be asked to sign a form giving your permission for your individual therapist to receive copies.

POTENTIAL BENEFITS

You may not benefit from being in this study. Other people in the future may benefit from what the researchers learn from the study.

RISKS AND DISCOMFORTS

There are no known risks from being in this study, however, you may feel upset or uncomfortable during or after answering the questions because you will be thinking about your feelings, thoughts, and behaviors. In the unlikely event that this happens, one of the group leaders or the researcher will be available to talk to you. If you become seriously

upset while not at The Keystone Center, please go to your local hospital emergency room.

ALTERNATIVES

The other choice is to not be in this study.

PAYMENT

You will not receive any payment for being in this study.

CONFIDENTIALITY

All information and medical records relating to your participation will be kept in a locked file. Only the research team, members of the Institutional Review Board and the U.S. Food and Drug Administration will be able to look at these records. If the results of this study are published, no names or other identifying information will be used.

REASONS YOU MAY BE TAKEN OUT OF THE STUDY WITHOUT YOUR CONSENT

If health conditions occur that would make staying in the study possibly dangerous to you, or if other conditions occur that would damage you or your health, Dr. Freeman or his/her associates may take you out of this study. In addition, the entire study may be stopped if dangerous risks or side effects occur in other people. You will also be taken out of the study if you develop mental health symptoms that would make it impossible for you to complete the questionnaire.

NEW FINDINGS

If any new information develops that may affect your willingness to stay in this study, you will be told about it. If you are interested in the results of this study, you may ask the investigators for a copy of the results for the group as a whole.

INJURY

If you are injured as a result of this research study, you will be provided with immediate necessary medical care.

However, you will not be reimbursed for medical care or receive other payment. PCOM will not be responsible for any of your bills, including any routine medical care under this program or reimbursement for any side effects that may occur as a result of this program.

If you believe that you have suffered injury or illness in the course of this research, you should notify John Simelaro, D.O., Chairperson, PCOM Institutional Review Board at

(215) 871-6337. A review by a committee will be arranged to determine if your injury or illness is a result of your being in this research. You should also contact Dr. Simelaro if you think that you have not been told enough about the risks, benefits, or other options, or that you are being pressured to stay in this study against your wishes.

VOLUNTARY PARTICIPATION

You may refuse to be in this study. You voluntarily consent to be in this study with the understanding of the known possible effects or hazards that might occur while you are in this study. Not all the possible effects of the study are known.

You may leave this study at any time by notifying Dr. Freeman, Leah Longan, or one of The Keystone Center staff that you no longer wish to participate.

You also understand that if you drop out of this study, there will be no penalty or loss of benefits to which you are entitled.

I have had adequate time to read this form and I understand its contents. I have been given a copy for my personal records.

I agree to be in this research study.

Signature of Subject: _____

Date: _____/_____/_____ Time:_____AM/PM

Signature of Witness:_____

Date: ____/_____/_____ Time:_____AM/PM

Signature of Investigator:_____

Date: ____/____/_____ Time:_____AM/PM

Appendix F: Clinician Informed Consent Form

INFORMED CONSENT FORM
CLINICIANS

TITLE OF STUDY

Correspondence of Self and Observer Ratings of Depression using the BDI-II and BDI-II-O

PURPOSE

The purpose of this research is to find out how self-report and clinician ratings of depression might be related, how the ratings of multiple clinicians might be related, and what factors influence these relationships.

You are being asked to be in this research study because you are a clinician who has regular contact (i.e., at least once per week for therapists and once every two months for psychiatrists) with patients who have a diagnosis of depression or who have mental health problems that are usually accompanied by depression. If you do not meet this requirement for regular contact, you can not be in the study.

INVESTIGATOR(S)

Principal Investigator

Name: Dr. Arthur Freeman, Ed.D.

Department: PCOM, Department of Psychology

Address: 4190 City Avenue
Philadelphia, Pa. 19131

Phone: 215 871-6456

Responsible Investigator

Name: Leah Longan, M.S., Psy.D. Candidate

Department: PCOM, Department of Psychology

Address: 4190 City Avenue
Philadelphia, Pa. 19131

Phone: 717 871-6442

The doctors and scientists at Philadelphia College of Osteopathic Medicine (PCOM) do research on diseases, new treatments, and psychological issues. The study you are being asked to volunteer for is part of a research project on depression.

Even though this research project is to study depression, no one can say whether the results will result in better treatment.

If you have any questions about this research, you can call Dr. Arthur Freeman at (215) 871-6456.

If you have any questions or problems during the study, you can ask Dr. Freeman, who will be available during the entire study. If you want to know more about Dr. Freeman's background, or the rights of research subjects, you can call Dr. John Simelaro, Chairperson, PCOM Institutional Review Board at (215) 871-6337.

DESCRIPTION OF THE PROCEDURES

If you agree to be a part of this study, you will be asked to fill out, at two separate times, the Beck Depression Inventory-II-O (BDI-II-O). This is a clinician rated questionnaire that asks about patient feelings, thoughts, and behaviors that are related to depression. You will fill out a BDI-II-O for each participating patient during the first week of the study, and you will fill out a BDI-II-O for a smaller group of patients during the second week of the study. You will also be asked to fill out a demographic form that asks for your age, education level, race, gender, marital status, professional discipline, and years of experience. In addition, you will be ask to fill out a clinician/patient information form that asks about your level of acquaintanceship and familiarity with the patient, in addition to information about the number of hours you spend with the patient per week, and whether you are the patient's individual therapist. The demographic form and the clinician/patient information form will be filled out only once, which will be during the first week of the study.

The BDI-II-O will take about 5 minutes to complete per patient. There is nothing you need to do to prepare for participation in this research project.

You will be asked to put your first name and last initial on the BDI-II-O forms, as well as on the forms that ask for personal information as described above. The reason we need you to use your name is so that we can match your answers to those of the patients, and distinguish your answers from those of the other clinicians participating in the study. You will not have access to the patients' questionnaires or your questionnaires unless a patient chooses to discuss their answers with you during an individual therapy session. Should this be the case, the patient will be required to sign a release form before you are given copies of the questionnaires, and you will be required to keep the questionnaires and release forms in a locked drawer.

POTENTIAL BENEFITS

You may not benefit from being in this study. You could benefit from comparing ratings with a patient if they agree to do so. Other people in the future may benefit from what the researchers learn from the study.

RISKS AND DISCOMFORTS

There are no known risks from being in this study, however, you may feel upset or uncomfortable during or after completing the BDI-II-O because you will be thinking about your patient's depression. In the unlikely event that this happens, other group leaders and the researcher will be available to talk to you. If you become seriously upset while not at The Keystone Center, please go to your local hospital emergency room.

ALTERNATIVES

The other choice is to not be in this study.

PAYMENT

You will not receive any payment for being in this study.

CONFIDENTIALITY

All information and medical records relating to your participation will be kept in a locked file. Only the research team, members of the Institutional Review Board and the U.S. Food and Drug Administration will be able to look at these records. If the results of this study are published, no names or other identifying information will be used.

REASONS YOU MAY BE TAKEN OUT OF THE STUDY WITHOUT YOUR CONSENT

If health conditions occur that would make staying in the study possibly dangerous to you, or if other conditions occur that would damage you or your health, Dr. Freeman or his/her associates may take you out of this study. In addition, the entire study may be stopped if dangerous risks or side effects occur in other people. You will also be taken out of the study if you develop mental health symptoms that would make it impossible for you to complete the BDI-II-O.

NEW FINDINGS

If any new information develops that may affect your willingness to stay in this study, you will be told about it. If you are interested in the results of this study, you may ask the investigators for a copy of the results for the group as a whole.

INJURY

If you are injured as a result of this research study, you will be provided with immediate necessary medical care.

However, you will not be reimbursed for medical care or receive other payment. PCOM will not be responsible for any of your bills, including any routine medical care under this program or reimbursement for any side effects that may occur as a result of this program.

If you believe that you have suffered injury or illness in the course of this research, you should notify John Simelaro, D.O., Chairperson, PCOM Institutional Review Board at (215) 871-6337. A review by a committee will be arranged to determine if your injury or illness is a result of your being in this research. You should also contact Dr. Simelaro if you think that you have not been told enough about the risks, benefits, or other options, or that you are being pressured to stay in this study against your wishes.

VOLUNTARY PARTICIPATION

You may refuse to be in this study. You voluntarily consent to be in this study with the understanding of the known possible effects or hazards that might occur while you are in this study. Not all the possible effects of the study are known.

You may leave this study at any time by notifying Dr. Freeman or Leah Longan that you no longer wish to participate.

You also understand that if you drop out of this study, there will be no penalty or loss of benefits to which you are entitled.

I have had adequate time to read this form and I understand its contents. I have been given a copy for my personal records.

I agree to be in this research study.

Signature of Subject: _____

Date: _____/_____/_____ Time:_____AM/PM

Signature of Witness:_____

Date: ____/_____/_____ Time:_____AM/PM

Signature of Investigator:_____

Date: ____/____/_____ Time:_____AM/PM

Appendix G: Trait Ratability Survey
Clinician Opinions on the Ratability of Patient Traits

As a part of my dissertation study, I am gathering information about the ratability of patient traits from the clinician's perspective. It is hypothesized that some traits are easier to rate than others.

Please choose the number which best describes how difficult or easy you think it would be to rate a patient on the following characteristics. See completion options on the second page.

	Very Difficult	Moderately Difficult	A Little Difficult	A Little Easy	Moderately Easy	Very Easy
Sadness	1	2	3	4	5	6
Pessimism	1	2	3	4	5	6
Feelings of Past Failure	1	2	3	4	5	6
Loss of Pleasure	1	2	3	4	5	6
Guilty Feelings	1	2	3	4	5	6
Feeling of Punishment	1	2	3	4	5	6
Self-Dislike	1	2	3	4	5	6
Self-Criticalness	1	2	3	4	5	6
Suicidal Thoughts or Wishes	1	2	3	4	5	6
Crying	1	2	3	4	5	6
Agitation	1	2	3	4	5	6
Loss of Interest	1	2	3	4	5	6
	Very	Moderately	A Little	A Little	Moderately	Very

	Difficult	Difficult	Difficult	Easy	Easy	Easy
Indecisiveness	1	2	3	4	5	6
Feelings of Worthlessness	1	2	3	4	5	6
Loss of Energy	1	2	3	4	5	6
Sleep Changes	1	2	3	4	5	6
Irritability	1	2	3	4	5	6
Appetite Changes	1	2	3	4	5	6
Concentration Difficulty	1	2	3	4	5	6
Tiredness or Fatigue	1	2	3	4	5	6
Loss of Interest in Sex	1	2	3	4	5	6

Thank you for your time. Please return this survey to Leah Longan's mailbox at Keystone Psychological and Behavioral Services at 3700 Vartan Way, or fax to 1 419 831-2688.