2010-11-29

Throwing the Baby Out with the Bathwater: When Can We Trust Self Report with the SMI Inpatient Population?

Jeffrey A. Lee
Brigham Young University - Provo

Throwing the Baby Out with the Bathwater: When Can We Trust Self-Report Outcome Assessment from Inpatient SMI?

Jeffrey A. Lee

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Gary M. Burlingame, Committee Chair
Dennis L. Eggett
Sally A. Barlow
Patrick R. Steffen
Scott A. Baldwin
Frank M. Rees

Department of Psychology

Brigham Young University

December 2010

ABSTRACT

Throwing the Baby Out with the Bathwater: When Can We Trust

Self-Report Outcome Assessment from Inpatient SMI?

Jeffrey A. Lee

Department of Psychology

Doctor of Philosophy

Reliability of self-report outcome assessment is often called into question with the severely mentally ill population. In the context of inpatient care, demand characteristics may further complicate self-report measures. Although clinician-completed outcome measures, such as the Brief Psychiatric Rating Scale-Expanded Version (BPRS-E), have become industry standard with this population, self-report assessment may be useful under certain conditions. This study sought to explore the relationship between a clinician-completed, the BPRS-E, and a self-completed measure, the SOQ, within the SMI inpatient population. A total of 357 adult participants with a minimum of three assessment iterations were analyzed. The results of the analysis indicated both measures correlated at all assessment iterations (admission, 90-, 180-, 270-, 360+ days), but when divided into SOQ admission clinical and subclinical groups only the clinical group maintained the correlation at all points. A logistical regression analysis indicated that membership in the subclinical group can be predicted by one subscale (Mood Disturbance) and three items (Hallucinations, Uncooperativeness, and Conceptual Disorganization) from the BPRS-E. The change trajectories of both measures were essentially identical; however, when divided into SOQ admission clinical and subclinical scores the SOQ and BPRS-E change trajectories were significantly different from each other and clinical versus subclinical on the same measure were significantly different. Further examination of the subclinical SOQ group revealed two distinct groups, scores that *eventually* had reliable change and exceeded the cutoff score and those that never did. A logistical regression analyses revealed that membership in these two groups can be reliably predicted by two BPRS-E items (Somatic Concerns and Suspiciousness), in that as each item increases the likelihood of membership in the group that never exceeds the cutoff score also increases. These results suggest that although the SMI inpatient population present with profound limitations, it may be possible to predict those who will eventually provide reliable self-report outcome assessments and those who will not. Although further research is necessary, these results are promising and may provide decision points for clinicians on when and when not to trust self-report outcome assessment with the SMI inpatient population.

Keywords: BPRS-E, inpatient, outcome, SMI, SOQ

ACKNOWLEDGEMENTS

First and foremost, I would like to share my gratitude to my wife, Dina, and my children, Joshua, Lillian, Sophie, and Amy. Your love and support were essential to the completion of my doctoral program. Dina, you have always been a wonderful sounding board a supportive of me. Indeed, it probably felt like you were going through the program as well. Thank you for all the love you have provided. I realize that my education required several significant life adjustments and I wish to express appreciation for your willingness to endure this process with me. I thank each of you for your very important and individual contribution to this graduate program.

I would also like to thank Dr. Gary Burlingame and Dr. Dennis Eggett for their guidance, mentoring, and patience with me. I realize my methods could be described as "gum on a shoe," yet both of you were willing to work with me and provided clear and concise feedback.

I would like to also thank those at Utah State Hospital for their contribution and support, particularly Dr. Frank Rees and Ryan Taylor. In the final hour of this project, you two stood up and made significant sacrifices on my behalf. I appreciate your willingness to work hard and your invaluable contribution to this dissertation project. I thank you for your support.

Lastly, I would like to thank my Heavenly Father for the blessings of these people in my life, for the support provided, and for the challenges that helped me grow.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Throwing the Baby Out with the Bathwater: When Can We Trust Self-Report Outcome Assessment from Inpatient SMI?

In 2008 the United States spent $2.3 trillion on health care services, which was an increase of 4.4 percent from the previous year, and by 2019 health care expenditures is expected to increase to $4.5 trillion (Centers for Medicaid and Medicare Services, 2009). Of those health care expenditures an estimated seven percent are allocated to mental health services (Coffey et al., 2000; Mark, McKusick, King, Harwood, & Genuardi, 1998), of which the cost of inpatient care is disproportionately larger than that of other treatment settings. On a national level in 2004, the cost of inpatient care was five times outpatient care ($113 billion compared to $20 billion), an increase of almost 8% from the previous year (U.S. Census Bureau, 2006). The cost affiliated with the treatment of mental health problems underscores the necessity of assessing the effectiveness of mental health services.

In response to the economic cost, initiatives have emerged to manage the cost of mental health services and hold practitioners, accreditation bodies, public agencies, and consumers accountable for the effectiveness of their services while maintaining equitable care (Lyons, Howard, O'Mahoney, & Lish, 1997; Mirin & Namerow, 1991). In response to such initiatives, outcome management programs were designed to assist in the empirical evaluation of the effectiveness of therapeutic interventions and client change over time (Burlingame et al., 2001; Burlingame, Lambert, Reisinger, Neff, & Mosier, 1995; Lyons et al., 1997). Indeed, Lambert, Bergin, and Garfield (2004) articulated, "…*outcome management makes empiricism a viable part of routine practice rather than a distant abstraction that practitioners find difficult to incorporate in practice*" (p. 9; italics in original). To meet the essential criteria and purpose of

outcome management programs, treatment providers should incorporate both treatment endpoint assessment and continuous patient monitoring. These practices must rely upon outcome instrumentation that is standardized, psychometrically sound, easy to use, and practical (Burlingame et al., 1995; Lambert, 2001; Lambert et al., 2001; Vermillion & Pfeiffer, 1993). Although there are several outcome measures that meet such requirements (e.g., Outcome Questionnaire), fidelity to these recommendations are questionable with some populations, particularly the severely mentally ill (SMI).

By definition, the SMI population consists of individuals whose psychological symptoms represent a lack of or impairment in (1) safety, (2) informal and formal support, (3) diagnosis, (4) disability, and (5) duration ("the SIDDD dimensions;" Slade, Powell, & Strathdee, 1997). Indeed, this population primarily consists of individuals with cognitive and/or reality impairment that prevents them from adequately performing necessary daily functioning skills and primarily consists of people diagnosed with schizophrenia, schizoaffective disorder, bipolar disorder, major depression, and pervasive developmental disorders (Carey & Carey, 1999). Cognitive deficits are considered key symptoms of schizophrenia, which are important determinates of poor social functioning, memory, attention, processing speed, and executive functioning (Aleman, Hijman, de Haan, & Kahn, 1999; Dibben, Rice, Laws, & McKenna, 2009; Galderisi et al., 2009; Gold & Harvey, 1993). Likewise, reality impairment, as understood by distortions in social-, other-, and self-perceptions and self-concepts, are common features of the SMI population (Kim et al., 2007; Nieznanski, 2005). Indeed, many of the inpatient SMI population are court mandated to treatment because they pose a threat to selves or others or are unable to adequately care for themselves. The functional limitations associated with cognitive and reality

impairment of the SMI population poses unique difficulties in reliably assessing change in functioning.

Outcome assessment with the SMI population has changed throughout the years. Lachar et al. (2001, p. 163) noted:

> [Clinician] rating scales such as the BPRS have recently achieved an advantage over self-completed measures in the evaluation of hospital-based treatment because such patients must now exhibit disabling psychopathology to justify their hospitalization (Nelson, Maruish, & Axler, 2000). As a consequence of these contemporary admission standards, newly admitted psychiatric patients are often unable to complete even a brief self-completed diagnostic questionnaire.

Indeed, clinician-completed measures have distinct advantages over self-completed measures with the SMI population because assessors are able to gather reliable data mitigating the serious psychiatric symptomatology or functional impairments by utilizing reliable, trained, professional raters. This outcome assessment process further provides increased reliability, validity, and allows additional exploration into a breadth and depth of symptoms (e.g., the clinician in a semi- or unstructured interview would have allowances to question such symptoms). However, research has indicated that measures completed by different sources (i.e., self, clinician, peer, teacher, and parent) yield different results, particularly when applied to change indices (Hill & Lambert, 2004; Monti, Wallander, Ahern, Abrams, & Monroe, 1983). Furthermore, responses may be influenced by the presence of an assessor (Rosenthal, 1966). Given the context of outcome measures and the presence of the therapist these concerns may be particularly important and influential. Lastly, maintaining inter-rater reliability, minimizing rater drift, training, and the clinician's time associated with making the assessment is often expensive or results in less

frequent or unreliable ratings. The alternative benefits of self-completed measures are that they are quick, cost effective, require minimal training, and can provide immediate feedback, although at the cost of questionable reliability with the SMI population. Hill and Lambert (2004, p. 122) reported:

> Though generally reliable, the accuracy of self-reports when compared with that coming from other assessors seems to depend on the sensitivity of the information sought (e.g., demographics vs. arrest records), specificity of validation criteria (e.g., archival data vs. urine tests), personal characteristics of the informant (e.g., sober vs. intoxicated), reference to time (e.g., immediate past vs. early life), and demand characteristics of the research [or therapeutic] situation (e.g., intake interview vs. program evaluation).

Each of these characteristics are particularly important and potentially problematic given the environmental and psychological state of those deemed SMI when utilizing self-completed outcome measures.

Recognition that outcome assessment is increasingly becoming standard practice and that agencies, therapists, researchers and, perhaps most importantly, clients benefit from such assessment underscore the importance of implementing such practices. However, there remain important limitations to both clinician- and self-completed measures with regard to the SMI population that necessitate further investigation. The purpose of this study is to explore the relationship between two different sources of outcome measures (the Brief Psychiatric Rating Scale [BPRS] and the Severe Outcome Questionnaire [SOQ]) and to identify sub-populations of the SMI that may be able to provide reliable responses on a self-completed measure.

This research utilizes inpatient clinical samples that have been assessed by both a clinician- and self-completed measure over multiple occasions. This population is particularly

relevant because they represent the severest of the SMI population; thus capturing a sample of SMI who have been identified by multiple sources of meeting the functional limitations associated with such a population. Identifying the relationship between self-completed and clinician-completed measures with the SMI population and the potential usefulness of self-completed measures with a believed-to-be unreliable self-reporting population would assist researchers, practitioners, and agencies in determining how to assess the effectiveness of SMI treatment. Such addition of a new level of assessment may not only cut costs but add an introspective level of assessment with the SMI population.

Chapter two contains a review of the literature to elucidate the difficulties of achieving reliable self-completed outcome assessment with the SMI. Specifically, an empirical background will be provided for the two key outcome measures used in this study, the Brief Psychiatric Rating Scale (clinician-completed) and the Severe-Outcome Questionnaire (self-completed), advantages of hierarchal statistical methods and the hypotheses examined herein. In chapter three the specific method used to explore the relationship between the two measures is outlined. In chapter four the results of the analyses performed between the two measures are presented. In chapter five conclusions from the results are discussed and recommendations for future research and uses of self-completed measures with the SMI population are provided.

Literature Review

Historical Context

Psychological testing is rooted in three primary arenas: civil-service exams, school exams, and the study of individual differences (see Du Bois, 1970). Although psychological testing has an extensive history dating approximately 3,000 years ago in China, it has not been until more recently that measurement theory as a discipline, and subsequently assessment

measures, began to blossom. E.L. Thorndike's *An introduction to the Theory of Mental and Social Measurements* marked the first textbook on measurement theory in 1904. Between 1930 and 1950 a number of journals dedicated to measurement theory and psychological assessment began. Indeed, the historical context and development of measurement theory underscores the need to identify, explain, and illuminate the differences between individuals or groups. Presently, this need continues to be a driving force of psychological assessment and the primary purpose of the majority of psychological tests.

Of the thousands of psychological tests available today there are three broad categories: discriminative indexes, predictive indexes, and evaluative indexes (Kirshner & Guyatt, 1985). A discriminative index is a measure which is used to differentiate between two or more groups or individuals on an underlying dimension absent external criterion. Examples of discriminative indexes are the Wechsler Adult Intelligent Scale (WAIS), which its purpose is to discriminate levels of intelligence and the Minnesota Multiphasic Personality Inventory (MMPI-2). A predictive index is a measure used to categorize an individual or group into predefined categories when external criterion is available. An example of this is the Substance Abuse Subtle Screening Inventory (SASSI), which is utilized to identify those that are likely to develop a substance abuse disorder in the future. Finally, an evaluative index is utilized to measure change on a dimension of interest in an individual or group. Outcome measures, such as the two focused on in this study, are an example of these.

Purpose of Outcome Assessment

History and development. Outcome assessment has a relatively short life span relative to other assessment categories. Nevertheless, it grew out of a need to objectively measure the change in psychological phenomenon over time and as a result of intervention. For many years

the field relied on therapist ratings of general improvement that were often based on one's individual theoretical understanding of psychopathology and the change thereof. However, the field has evolved to adopt multiple perspectives and specificity of change to provide clarity, understanding, and equitable communication between practitioners. Hill and Lambert (2004) stated:

> The field has gradually moved from primary reliance on therapist ratings of gross general improvement to the use of outcome indices of specific symptoms that are quantified from a variety of viewpoints, including the client, therapist, trained observers, physiological indices, and environmental data such as employment records. (p. 106)

Furthermore, outcome assessment has developed into a scientific inquiry of itself in an effort to clarify, organize, and set standards that constitutes an acceptable outcome measure. Early devices that relied on inference (e.g., the Rorschach) are currently considered poor measures of outcome due to the prominence of a particular theoretical orientation, the lack of appropriate psychometric properties, and the emphasis of therapist interpretation. Indeed, theoretical specificity has subsided and importance has been placed on symptomatic states as a primary focus of outcome assessment. Hill and Lambert (2004) summarized the essential practices of outcome assessment as: "(1) clearly specifying what is being measured, so that replication is possible; (2) measuring change from multiple perspectives, with several types of rating scales and methods; (3) employing symptom-based, atheoretical measures; and (4) examining, to some extent, patterns of change over time"(p. 107). Certainly, outcome assessment is rapidly developing into an important research inquiry and clinical necessity.

Research. Although there has been significant improvement in standardized outcome assessment and the guidelines thereof, there remain significant problems of appropriate use of

outcome assessment in research. Garfield (1990) and Kiesler (1973) noted the unfortunate common practice of researchers developing an outcome measure for the sole purpose of their specific project and then never use it again. Indeed, 38% to 49% of the measures used in *JCP* and *JCCP* between 1978 and 1992 were new measures developed just for the use of the specific study (Hill, Nutt, & Jackson, 1994). This trend is highly problematic because little is known about the essential psychometric properties and standardization of these measures and it becomes increasingly more difficult to accumulate knowledge across studies. Indeed, many of the studies that utilize such practices may have questionable conclusions when made on such outcome measures. Unfortunately, the use of unstandardized outcome measures has also permeated the clinical field as well.

Clinicians. Research indicates that clinicians' ability to detect change, whether progress or exacerbation, is poor, particularly compared to standardized outcome measures (Dawes, 1996; Hannan et al., 2005; Meehl, 1996). As a result, objectively and empirically assessing change in client symptoms, progress, or exacerbation is critical to ethical and responsible practice. Ideally, clinicians are the primary consumer of outcome research so they can implement it in a way that benefits the therapeutic process and the client. The primary purpose of outcome assessment is to receive objective, empirically-based, immediate feedback on the client's psychological functioning and the progress or exacerbation of symptoms due to therapy (Lambert et al., 2001). Furthermore, researchers (Lambert et al., 2001; Wells, Burlingame, Lambert, Hoah, & Hope, 1996) and clinicians (Hatfield & Ogles, 2004, 2006, 2007) have identified the following as advantages of using outcome assessment:

1. Initial assessment with outcome measures can provide an index of current functioning, initial severity of symptoms, risk factors, and mediating and/or moderating variables that may impact treatment (Lambert et al., 2001; Wells et al., 1996).
2. Outcome measures track client change. Clinicians identified "Tracking client progress" as the most important reason and the most useful information for using outcome measures (Hatfield & Ogles, 2004, 2006).
3. Using standardized measures can provide additional validation of clinical judgment (Hill & Lambert 2004; Lambert et al., 2001).
4. "Determine if there is a need to alter treatment," either to meet the client's immediate psychological needs or adjust for exacerbation effects was the second most common reason provided by clinicians (Hatfield & Ogles, 2004).
5. Outcome assessment provides a therapeutic summary demonstrating the effects of therapy in easily understandable ways (Wells et al., 1996) and can provide critical feedback to affiliated sources (e.g., court, probation, parents, schools, agencies, administration, third-party payers).
6. The use of outcome assessment is considered "ethical practice" by clinicians (Hatfield & Ogles, 2004).

Despite the aforementioned benefits, Hatfield and Ogles (2004) surveyed 874 clinicians of which only 37% use outcome measures in their practice. This is a modest increase from Phelps, Eisman, and Kohout's (1998) survey of 15,918 who found that 29% responded that they relied upon outcome assessment. The low percentage of clinicians who use outcome assessment remains somewhat disconcerting with regard to the previously cited research indicating the difficulty of clinicians to reliably identify client change, especially exacerbation, and the client

benefits outcome assessment can facilitate. Furthermore, Hatfield and Ogles (2004) found that 60% of those who utilized outcome assessment used both standardized and unstandardized measures, 28% used only standardized, and 12% used only unstandardized. These findings further complicate the understanding of outcome assessment because the reliability and validity of those using unstandardized outcome measures may be reporting client benefit when that is not the case.

Several explanations have been provided to better understand the reasons why clinicians opt whether or not to use outcome assessment. In Hatfield and Ogles' (2004, 2006, 2007) research, clinicians reported "Adds too much paper work, Takes too much time, Extra burden on clients, Feel it is not important, and Do not have enough resources" as the top five reasons for not using outcome assessment. These reasons appear to be somewhat practical in nature and may be ameliorated through increased understanding of outcome assessment. For example, in response to "Extra burden on clients" clinicians may find it helpful to know that 57% of clients in the health field "perceived it as valuable" while only 4% disapproved (Nilsson, Wenemark, Bendtsen, & Kristenson, 2007). Furthermore, the context of outcome assessment may influence usage practices. As shown in Table 1, clinicians may be more likely to use outcome measures when there is increased accountability, they're associated with institutional organizations, or when one's theoretical orientation specifically endorses such practices. This summary, however, is somewhat rudimentary and more research is necessary to clarify clinician's motivations and client perceptions of outcome assessment. Regardless of the motivations of clinicians in favor of or opposed to outcome research, understanding the unique psychometric properties of outcome assessment will likely help both clinicians and researchers alike.

Table 1

Percentages of Clinicians that use Outcome Assessment in a Given Context

Work Setting		Source of Income		Theoretical Orientation	
Solo Private Practice	29%	Fee for service	30%	Cognitive-Behavioral	50%
Group private practice	35%	Managed Care / Private insurance	36%	Insight Oriented	24%
Institutional	50%	Institutional sources	48%	Eclectic	36%

Summarized from Hatfield and Ogles (2004, 2007)

Measuring Outcome

Multitrait and monotrait scales. An essential element of outcome measures that effects psychometric properties is whether the measure is designed to rate multiple (i.e., multitrait scales) or single (i.e., monotrait scales) traits. Multitrait scales have an advantage of assessing a breadth of psychological symptoms on a single measure and thus capturing symptoms that may not be readily apparent (Hill & Lambert, 2004). Indeed, clinical experience and research has shown that many clients present with comorbid diagnoses and/or with multiple symptoms. For example, in the event a client has anxiety and depressive symptoms (two pathologies that are often comorbid), a multitrait scale will capture levels of both, whereas a monotrait scale may not. However, multitrait scales may contain items that are irrelevant to some clients and/or have a reduced number of items assessing a particular trait, both of which could affect the overall sensitivity to change of the measure (Vermeersch, 1998). One method of addressing the multitrait problems is to provide sensitivity to change indices at all assessment levels.

Monotrait measures are advantageous when it is clear there is a single trait of interest. These types of measures are brief, only taking minutes to complete, can be repeated multiple times, and measure symptoms that are common across many psychopathologies (e.g., mood,

anxiety, self-esteem, motivation). However, validity is often difficult to establish. Many of these measures have names that "provide an illusion" that they precisely measure the construct of interest, while the constructs themselves (e.g., anxiety, depression) are not as distinct as assumed. Indeed, these measures are often highly correlated with measures presumed to assess a different construct (Hill & Lambert, 2004).

Vermeersch (1998) elucidated the potential affects to sensitivity to change on both multitrait and monotrait scales. He explained that if the monotrait scale does not adequately assess the homogeneous trait of interest or if there is other, related, psychopathological problems that exacerbate the trait of interest then sensitivity to change will be minimal. The multitrait scale, in turn, may have items that are irrelevant to the client which will show little to no change and affect the overall sensitivity to change on the measure. To further illustrate the problem, Froyd, Lambert, and Froyd (1996) identified over 1,430 outcome measures of which the MMPI is in the top 10 most frequently used measures of change, despite its lack of appropriate change indices. Indeed, many of the existing outcome measures are developed according to criteria that are irrelevant to change (Collins & Cliff, 1990). It is clear that measure development depends on the type of measure one is using (e.g., monotrait versus multitrait) and the context in which it is used (e.g., static versus change) when establishing reliability and validity indices.

Reliability. Reliability refers to the degree the test scores are consistent, dependable, or repeatable (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985). In essence, reliability is an estimate of error inherent in a measure around one's true score (Allen & Yen, 1979). If reliability is low then resulting scores represent a large amount of error and a correspondingly poor estimate of true score. Conversely, if reliability is high then the resulting scores represent a small amount of

error and confidence in the estimate of true score is low. Allen and Yen (1979) identified two types of error that affect the results of the scores. *Unsystematic error* is random deviations of the examinee's observed score from a theoretical true score. Minimizing this error is critical in increasing the reliability of a measure.

Although not considered true error in true score theory, *systematic error* affects one's overall confidence of the measure. Reality impairment is an example of systematic error in outcome assessment in the context of the SMI population. Such error is predictable and may be reduced when identified. For example, a 20-point lowering of symptoms due to reality impairment is systematic error. Additionally there may be systematic changes in systematic error with regard to the degree of reality impairment. For example, a low level of reality impairment would result in a low level of systematic error, say 5 points. As reality impairment increases systematic error increases. Thus, a high level of reality impairment may result in a high level of systematic error, say 40 points. Elucidating both types of error are critical in understanding what affects the resulting scores of a measure and the level of confidence one has that the scores represent one's true score.

Allen and Yen (1979) identified the following three common estimates of reliability: test-retest, parallel forms, and internal consistency. As previously mentioned, the selection of the appropriate estimates of reliability depends on the type of measure and the context in which the measure is utilized. In the context of outcome assessment, the reliability of change scores is not equivalent to the reliability of the measure (Allen & Yen, 1979; Cronbach, 1990; Nunally, 1978) because changes in scores are due to the difference of whatever is being measured *and* error (Hill & Lambert, 2004). Thus, relying on internal consistency reliability as a reliability estimate of change is insufficient. In contrast, Lambert and Hill (1994) noted that test-retest reliability was

particularly important for outcome measures because it is critical to capture the reliability of a measure prior to intervention, following intervention, and at some later time. Theoretically, an outcome measure with perfect test-retest reliability will have no change prior to intervention, the expected amount of change following intervention, and the expected amount of maintained gains from an intervention on follow up. However, there are few circumstances in which reliability is perfect; therefore, the estimate of change reflects both the actual change in one's true score and the change in error. Hill and Lambert (2004) explain how critical this concept is in outcome assessment by stating,

> The reliability of a particular measure in outcome assessment is especially critical because low reliability of a measure is compounded with computation of a change score. . . . The use of unstandardized scales makes it difficult, at best, to estimate the amount of change necessary to conclude that the difference between two scores is not due to chance fluctuations in the scores. (p. 116)

Therefore, appropriate estimates of reliability are critical in outcome assessment because the error in the measure will negatively impact the overall reliability and the estimate of change. Researchers recommend internal consistently reliability and test-retest reliability to be at or above 0.80 (Burlingame et al., 1995; Durlak, Wells, Cotton, & Johnson, 1995) and 0.70 (Durlak et al., 1995; Reisinger & Burlingame, 1997), respectively.

The use of judges introduces an additional psychometric consideration in the application of outcome assessment. Utilizing judges (e.g., therapists, independent observers) in psychotherapy outcome assessment is a common practice with several special populations (e.g., children, adolescents, SMI, forensic). Judges theoretically provide an objective perspective; however, Fiske (1977) explained that the use of human judges introduces noise and

undependability in the data. Judges can be considered to be "two-legged meters" because they may interpret the data on the basis of their own reactions. Indeed, when judges are used, they, rather than the measures, are the actual measuring instruments (Mercer & Loesch, 1979). Therefore, it is imperative that the reliability of the judges be assessed and reported.

Reliability among judges ought to be determined through a fixed-effects model form of intraclass correlation when the measure uses an interval scale and the judges are nonrandom (i.e., they are selected because of predetermined criteria; Hill & Lambert, 2004; Shrout & Fleiss, 1979). Similar to other reliability estimates, the higher the reliability coefficient the more confidence one can have that the score contains less variance due to unrelated factors (e.g., aspects of the judges) and is a reflection of interrater agreement and the true score (Finn, 1974). It is recommended that intraclass coefficients be at or above 0.70 to maintain a reasonable level of confidence between interrater agreements (Heppner, Kivlighan, & Wampold, 1999). To achieve such a task, training judges can reduce both contaminating characteristics and the idiosyncratic interpretations of items (Bachrach, Mintz, & Luborsky, 1971; Caracena & Vicory, 1969; Hill & Lambert, 2004; Klein & Cleary, 1967; Mercer & Loesch, 1979). Training recommendations include the following:

1. Sensitize judges to the phenomena and not just to catch phrases (Bordin et al., 1954)
2. Utilize *informal* training. Make independent judgments on a data set other than the one used for the current study and discuss discrepancies until unanimity (Mercer & Loesch, 1979).
3. Use *formal* training. Train judges to a high degree of reliability to a "gold standard" of expert judges (Mercer & Loesch, 1979).
4. Cover a wide range of the severity of the construct of interest (Mercer & Loesch, 1979).

5. Include a plethora of examples representing both likely and unlikely rating circumstances (Mercer & Loesch, 1979).
6. Judges should be aware of social influential processes that may affect the scores (Hill & Lambert, 2004).
7. Protect against "rater drift" by including regular recalibrations of reliability to the expert judges, have regular rating consultations, or have a calibrated independent rater rate the same person of interest consecutively (Ventura, Green, Shaner, & Liberman, 1993).
8. Include procedures that make the judges think they will be checked or monitored (e.g., telling the judges that random informants will be used; Mitchell, 1979; Reid, 1970; Romanczyk, Kent, Diament, & O'Leary, 1973; Taplin & Reid, 1973).

Although judges introduce extra reliability considerations, proper training and instruction will likely increase the confidence in the measure.

Validity. Validity refers to the degree a test measures what it purports to measure. For example, if a test is designed as a monotrait measure of anxiety, such as the Beck Anxiety Inventory (BAI), then the items on the measure should measure the corresponding symptoms. Likewise, if a measure is designed as a multitrait measure, such as the MMPI-2, then the items on the measure should measure each of the constructs that the measure purports to measure. Establishing the validity of measures should be a high priority of psychotherapy research (Lambert & Hill, 1994) and, as mentioned, the established validity should be appropriate for the purported use of the measure. Allen and Yen (1979) identified content, criterion-related, and construct validity as three essential types of validity.

Content validity "is established through a rational analysis of the content of a test, and its determination is based on individual, subjective judgment" (Allen & Yen, 1979). Because it is

determined by subjective evaluation it is more subject to error than other types of validity. Face and logical validity are types of content validity and both require a precise definition of the domain in question. For example, the Beck Depression Inventory has high face validity because the statement, "I am so sad or unhappy that I can't stand it" reflects the specific content of *depression*, as stated in the name of the measure and the definition of depression according to the Diagnostic and Statistical Manual of Mental Disorders-Test Revision, 4th ed. (DSM-IV-TR, APA, 2000). In contrast, the statement "I like mechanics magazines" as a measure of masculinity on the MMPI-2 has low face validity because there is no rational link between the two.

Allen and Yen (1979) explained that criterion-related validity "is used when test scores can be related to a criterion. The criterion is some behavior [or external criteria] that the test scores are used to predict" (p. 97). Predictive and concurrent validity are two types of criterion-related validity. Predictive validity involves using a test score to predict a future behavior. For example, the essential purpose of the SASSI is to predict future alcohol/drug abuse or dependency. Concurrent validity refers to the ability of one instrument to estimate the score on another. For example, the Shipley was designed as a brief measure to estimate the total score of the WAIS which is a more extensive measure of cognitive ability.

A measure's construct validity "is the degree to which it measures the theoretical construct or trait it was designed to measure" (Allen & Yen, 1979, p. 108). Construct validity is particularly important for outcome assessment because there are several facets of change constructs that ought to be established. Kirshner and Guyatt (1985) posited, with regards to outcome assessment the most convincing support for construct validity is the establishment of longitudinal within-subject changes as a result of an intervention. Multitrait-multimethod

validity, factorial validity, and sensitivity to change are all aspects of construct validity and are critical for outcome assessment.

Multitrait-multimethod validity is used when two or more traits are being measured by two or more methods (Allen & Yen, 1979). The critical elements of this type of validity are to establish convergent and divergent validity. Suppose two traits (e.g., depression and hallucinations) are being measured by different methods (e.g., clinician-completed and self-completed). Ideally, the resulting correlations between the different traits and the different methods of assessment would, in general, match the matrix represented in Table 2. Allen and Yen (1979) explain that the highest correlation should be when the same trait is measured by the same method followed by the same trait measured by different methods. *Convergent validity* in a measure is supported when there are high correlations when the same construct is measured by the same method. Conversely, *divergent validity* is supported when the lowest correlation occurs when different traits are measured by different methods as illustrated in Table 2. This process is particularly important when dealing with multitrait measures because if two or more traits on the

Table 2

Ideal Multitrait-Multimethod Validity Matrix for a 2 x 2 Case

	Trait-1 Method-1	Trait-2 Method-1	Trait-1 Method-2	Trait-2 Method-2
Trait-1 Method-1	highest	low	high	lowest
Trait-2 Method-1	low	highest	lowest	high
Trait-1 Method-2	high	lowest	highest	low
Trait-2 Method-2	lowest	high	low	highest

same measure theoretically do not relate then they should have low correlations. Furthermore, these correlations should also maintain through several iterations of assessment with regard to outcome assessment. If the correlations on the measure do not follow these guidelines then construct validity is in question.

Factorial validity, as explained by Allen and Yen (1997), is established by a factor analysis which analyzes the interrelationships of variables and thereby explaining those relationships with factors. A factor is a hypothetical variable (or construct) that influences the score on one or more questions. For example, Table 3 illustrates how factors one and two influence the hallucination and depression items, respectively, as evidenced by the high loadings. Each factor represents a different hypothetical construct. This example lends evidence of the construct validity of this sample measure because all the hallucination items and depression items loaded highly onto their respective factors and not onto the other factor. In the context of outcome assessment, factorial validity would only be reasonably established if the same factors held up through longitudinal assessments. If through multiple assessment iterations the number

Table 3

Example of Factor Loadings

	Factor	
Test item	1	2
Hallucination 1	.998	.001
Hallucination 2	.812	.111
Hallucination 3	.798	.212
Depression 1	.123	.769
Depression 2	.094	.891
Depression 3	.199	.901

of factors decreased (i.e., given the example, a factor analysis on a subsequent assessment resulted in only one factor) or the loadings became mixed (i.e., in the given example, a factor analysis on a subsequent assessment resulted in some hallucination and some depression items loading high on factor one while other hallucination and depression items loaded onto factor two) then the construct validity of the measure would be in question. Therefore, a one-time factor analysis is insufficient to establish factorial validity of an outcome measure; rather, multiple factor analyses should be conducted through different time intervals to support the factorial validity of an outcome measure.

When the construct theoretically changes as a result of intervention then sensitivity to change is an essential element of construct validity. Lispey (1990, p. 100) explained, "Validity alone is not sufficient to make a measure responsive to treatment effects. What is required is validity for change. A measure can be a valid indicator of a characteristic but still not be a valid indicator of change on that characteristic." The same author continued, "Measurement sensitivity…means that measured values fully reflect any change of interest of the characteristic measured and do not reflect an appreciable amount of variance from any other source" (p.120). Indeed, sensitivity to change is the *most important characteristic* of a treatment outcome instrument (emphasis added; Burlingame et al., 1995). Aiken (1977) explains that the sensitivity of a measure refers to its ability to accurately reflect the change difference on the underlying construct of interest. Recently, several authors have elucidated the concept of sensitivity to change (e.g., Deyo, Diehr, & Patrick, 1991; Guyatt, 1988; Hill & Lambert, 2004; Kazdin, 1992; Lambert & Hill, 1994). In the context of outcome assessment, sensitivity to change can be operationally defined as the degree to which a measure is likely to reflect changes that occur following an intervention.

Vermeersch (1998) elucidated how critical establishing sensitivity to change beyond traditional psychometric indices on an outcome assessment. In his example, he showed how one measure may have excellent intraclass correlations and be very poor at detecting sensitivity to change, while another measure with very poor intraclass correlations may have excellent sensitivity to change. Indeed, in a review of agoraphobia several authors found that different methods of measuring change resulted in drastically different conclusions (Ogles, Lambert, Weight, & Payne, 1990). In their review they found a difference of mean effect sizes of 2.22, based on the type of outcome measure (phobic anxiety and avoidance had an effect size of 2.66 and heart rate had an effect size of .44). The drastic difference of results between the types of outcome measures underscores the importance of establishing the sensitivity to change of the measure. Five primary components of sensitivity to change ought to be evaluated: (1) the degree to how much the construct must theoretically change before the measure detects the change; (2) change on a given item ought to reflect the change in the theoretical direction after an intervention; (3) the change identified by the measure is not attributable to error; (4) the measure rates the total range of change, avoiding floor and ceiling effects; and (5) the measure should reflect significantly more change than that of controls (Burlingame, Nelson, Lee, Thayer, & Lambert, 2008; Tryon, 1991; Vermeersch, Lambert, & Burlingame, 2000).

In summary, clinicians, researchers, health care managers, and program managers should be aware of the essential properties of outcome measures when implementing outcome assessment programs. The following are essential criteria for outcome assessment programs:

1. Outcome measures should have excellent and relevant reliability, particularly test-retest reliability, and interrater reliability when judges are used (Hill & Lambert, 2004; Vermillion & Pfeiffer, 1993).

2. Outcome measures should have excellent and relevant validity, including a neglected but essential characteristic of sensitivity to change (Hill & Lambert, 2004; Vermillion & Pfeiffer, 1993).
3. The sensitivity to change should be established on an item, subscale, and total score basis (Hill & Lambert, 2004; Lipsey, 1990; Vermillion & Pfeiffer, 1993).
4. Outcome measures should be normed so clinicians can make an assessment of the clinical significance of treatment effects in addition to statistical significance (Burlingame et al., 2005; Hill & Lambert, 2004; Vermeersch et al., 2000; Vermillion & Pfeiffer, 1993).
5. Outcome measures should establish cutoff scores and a reliable change index (Hill & Lambert, 2004; Jacobson et al., 1984; Jacobson & Truax, 1991).
6. Outcome measures should be able to be completed by clients in a matter of minutes (Burlingame et al., 1995; Burlingame et al., 2008; Hill & Lambert, 2004; Lipsey, 1990).
7. Outcome measures should not be an extra burden on clients (Hatfield & Ogles, 2004, 2006).
8. Outcome measures should be easily scored and interpretable (Burlingame et al., 1995; Hill & Lambert, 2004).
9. Outcome measures should minimize the paperwork or otherwise taxing available human resources (Hatfield & Ogles, 2004, 2006).
10. Outcome measures should be designed to allow frequent use to track progress or monitor treatment (Burlingame et al., 1995; Hatfield & Ogles, 2004, 2006, 2007) so that treatment can be altered accordingly or problem areas can be identified

dynamically rather than waiting an extended period of time or until the end of treatment before making an assessment.

11. An outcome measure should be cost efficient (i.e., inexpensive) so it does not adversely impact the consumer's financial resources (Burlingame et al., 1995; Hatfield & Ogles, 2004, 2006, 2007).

Although many measures have demonstrated fidelity to these recommendations with the normal or outpatient population, there remains a paucity of research with outcome measures that demonstrate the same fidelity with regards to the SMI population.

The Severely Mentally Ill

Operationalizing severely mentally ill. Although "severely mentally ill" (SMI) connotes particular diagnostic categories (e.g., schizophrenia, bipolar disorder), there remains a lack of consensus of the definition of "severely mentally ill." Perhaps the most widely utilized definition is by the National Institute of Mental Health (NIMH, 1987) which requires the following three criteria: a diagnosis of non-organic psychosis or personality disorder; duration of "prolonged illness and long-term treatment" as operationalized as a two-year or longer history of mental illness or treatment; and disability, which was described as including at least three of the eight specified criteria (as cited in Ruggeri, Leese, Thornicroft, Bisoffi, & Tansella, 2000). Rothbard, Schinnar, and Goldman (1996) reported individuals who are SMI, by definition, have "functional limitations in activities for daily living, social interaction, concentration, and adaptation to change in the environment" for "twelve months or more." After an exhaustive literature review in an effort to operationalization SMI, Slade, Powell, and Strathdee (1997) concluded that definitions of SMI utilized the following five dimensions: (1) safety, (2) informal and formal support, (3) diagnosis, (4) disability, and (5) duration ("the SIDDD dimensions").

SMI at Utah State Hospital. As difficult as SMI is to operationalize, treatment thereof has its own complications. The combination of multiple demands placed on treatment providers with a difficult-to-treat population necessitates exceptional care and programs that meet such obligations. Utah State Hospital (USH) is one such treatment facility that was awarded a three-year grant because of their commitment to research and treatment (Earnshaw et al., 2005). As the only inpatient treatment center in Utah that serves the SMI population exclusively, they are an exemplar treatment facility that incorporates an outcome management program to meet the multitude of demands placed on them. As a result of their valiant efforts they were able to elucidate the challenges and provide recommendations that are inherent in treating the SMI population.

One such challenge was the meaningfulness of self-completed measures for the SMI population. Earnshaw et al. (2005, p. 412) reported, "one-fourth of the patients who were admitted to the facility were either unable (because of the acuity of their illness) or unwilling to complete a self-reported outcome instrument on admission." Questionable reliability and validity were evident because much of the data were either far below normative levels or responses were erratic (items were endorsed at both ends of the range of pathological symptoms severity). Therefore, one of the looming questions was, "…were self-reported outcomes data from our patients meaningful" (p. 412)? Although Earnshaw et al. (2005) found moderate correlations between the BPRS and those that self-completed in the clinical range on the SOQ, there remain questions on how to interpret those who reported in the subclinical range. Indeed, as Figure 1 depicts there are several change trajectories of subclinical scores that have been identified by clinical experience. One change trajectory is that there is no reliable change in report of symptoms from the patient, either they refuse to fill out the questionnaire or their

functional limitations preclude a reliable assessment. A second trajectory has a delayed or shallow rate of change. The last has a reliable change and surpassing the cutoff score ending with a gradual decline in the severity of symptoms.

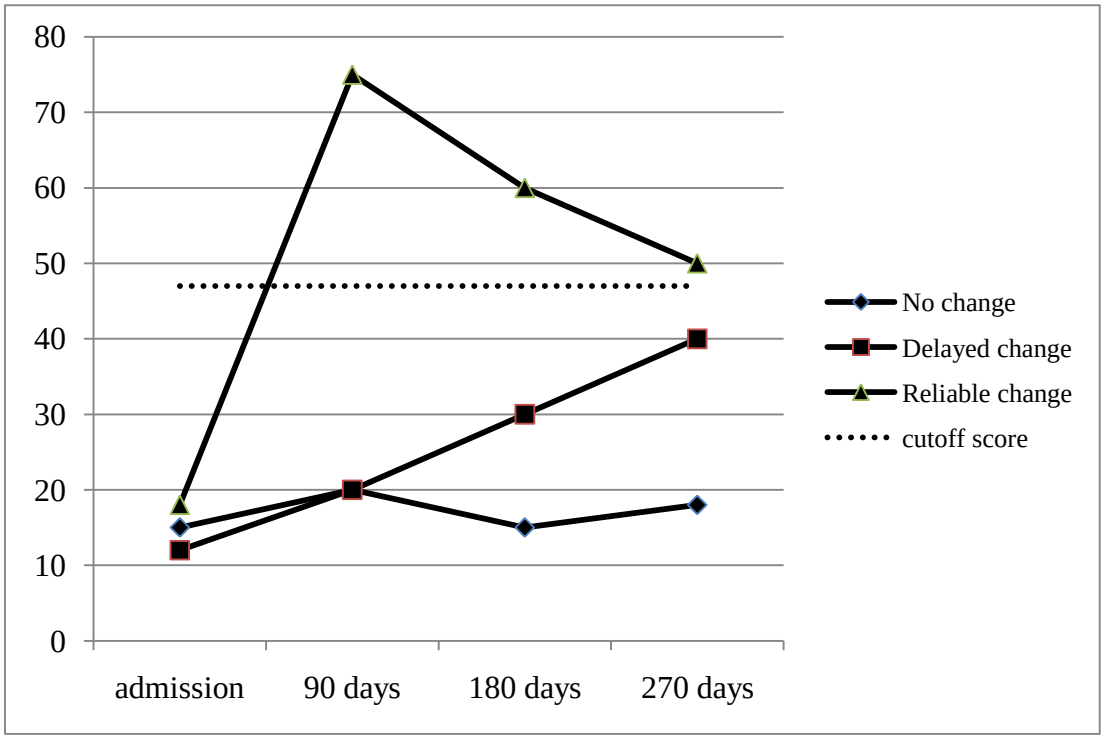


Figure 1. *Hypothesized Subclinical Change Trajectories*

In addition to the question of the application of self-completed measures with the SMI population, Earnshaw et al. (2005) indicated that there was initially significant resistance by the clinical staff to the USH intensive outcome management program. Earnshaw and colleagues concluded, "The greatest factor that contributed to 'buy in' by clinical staff occurred when the data started to empirically demonstrate that treatment was effective and made a difference in the functioning of our patients (p. 412). With regard to outcome assessment, they unexpectedly found that "patients' compliance increased with staff acceptance" (p. 413). With the questionable utility of self-completed measures, it is no surprise that clinicians are skeptical of

utilizing self-completed measures with the SMI. However, as eluded to, if research can show the benefit of using self-completed measures with the SMI, perhaps compliance by the clinical staff, and in turn the patients, will increase.

Outcome with the SMI. The importance of tracking symptom change in the SMI population is underscored by the possible negative pharmacotherapy treatment effects. Indeed, negative drug interactions have been well documented among the SMI population (Gaultieri & Powell, 1978; Kapur & Kambhampati, 1992; Leipzig & Mendelowitz, 1992; ZumBrunnen & Jann, 1998). Kapur and Kambhampati (1992) explain that this problem may result from the increase in multiple medications given and managed by more than one physician, iatrogenic effects, ignorance, or lack of adequate research on potential drug-on-drug interactions. With the increase of pharmacotherapy treatment with the SMI population and the associated rapid response and effectiveness of such treatment, the need for assessment measures that track biopsychosocial change accurately is necessary.

Although, by definition, significant cognitive distortions and inaccurate reality testing are associated with the SMI, self-completed assessment has shown to be reliable and valid among this population. Several research studies have shown adequate to strong correlations between self-completed and clinician-completed assessment measures when addressing quality of life or patients with schizophrenia (Voruganti, Heslegrave, & Awad, 1998; Wilkinson, Hesdon, & Wild, 2000), drug treatment compliance (Hogan, Awad, & Eastwood, 1983), self-concept (McCay & Seeman, 1998), insight into disorder (Jovanovski, Zakzanis, & Atia, 2007; Van Lieshout & Goldberg, 2007), and severity of symptoms (Katz, Shaw, & Vallis, 1995). Although all the aforementioned self-completed studies reported were comparisons of self-completed to other-completed methods (clinician rated, behavioral observations, video recording), none of the

previously reported comparisons included outcome studies. Outcome assessment presents an additional dilemma because it, theoretically, should reliably assess the patient at all levels of symptom severity. Indeed, some reports suggest significant variability in the reliability of self-completed measures among the cognitively disturbed (i.e., inpatient versus outpatient populations) and that the more impaired the less reliable self-completed measures become (Burlingame et al., 2008; Earnshaw et al., 2005). Although some evidence suggests self-completed assessment with the SMI population is adequately reliable, it should be noted that there is a tendency for negative results to not get published (this is exacerbated in outcome comparisons with the SMI population because of the tendency to include the most severe pathologies). Thus, little is known about non-significant relationships between self-completed measures and other-rated outcome measures with respect to the SMI population. This problem is compounded when SMI patients self-report symptoms at subclinical levels. Therefore, interpretations of self-completed measures with SMI patients should be made with caution until more is known about the reliability, validity, and limitations of self-completed outcome measures with the SMI.

The Brief Psychiatric Rating Scale

Development. Overall and Gorham (1962) developed the original BPRS to answer the call for a reliable and valid assessment tool to monitor SMI patient symptom change. The need of an outcome assessment tool followed from the increase of medication treatments of the 1950s and early 1960s and the subsequent increase in the cost of treatment of the inpatient population. Furthermore, the necessity of showing efficient and effective treatments created the need for a measure that could be employed quickly by the clinician and would track patient change.

Table 4 illustrates the current status of the BPRS-Extended Version (BPRS-E) and is a summary of the development of the BPRS. The original 16-item BPRS was empirically derived by factor analysis from two longer assessment tools, the Multidimensional Scale for Rating Psychiatric Patients (Lorr, Jenkins, & Holsopple, 1953) and the Inpatient Multidimentional Psychiatric Scale (Lorr, McNair, Klett, & Lasky, 1960). The remaining items were then set to a 7-point Likert scale to measure levels of severity within each symptom (Overall & Gorham, 1962). The 7-point Likert scale ranged from *not present* to *extremely severe*, although in some recent versions an additional *N/A* indicator is available for situations in which that item was "not assessed" or information was "not available." Two additional items (excitement and disorientation) were added years later to make the measure an 18-item scale (Overall & Klett 1972). Beller and Overall (1984) reported these items were added to increase the utility among the geropsychiatric patients.

The most significant revision to the BPRS was the addition of six items (suicidality, self-neglect, bizarre behavior, elated mood, motor hyperactivity, distractibility) making the final version a 24-item measure ([BPRS-E] Lukoff, Nuechterlein, & Ventura, 1986). Furthermore, these authors also provided anchor points (behavioral descriptions) to the 7-point Likert scale which were not present in the previous 16-item or 18-item versions. Table 5 is an example of the anchor points used for the hallucination item. As evident, as the scale number increases so does the level of functional impairment or the presence of the particular symptom. In this example, a higher score would be rated if the patient reported an increase of hallucinatory experiences or that the hallucinations had an impact on their functioning. Therefore, one could get a score of seven if they had hallucinations on a daily basis throughout the day with no functional impact, if

Table 4

The Brief Psychiatric Rating Scale

Scale item	Item description
Somatic Concern*	Degree of concern over present bodily health. Rate the degree to which physical health is perceived as a problem by the patient, whether complaints have realistic bases or not. Somatic delusions should be rated in the severe range with or without somatic concern.
Anxiety*	Reported apprehension, tension, fear, panic or worry. Rate only the patient's statements, not observed anxiety which is rated under Tension.
Depression*	Include sadness, unhappiness, anhedonia, and preoccupation with depressing topics, hopelessness and loss of self-esteem. Do not include vegetative symptoms, e.g., motor retardation, early waking, or the amotivation that accompanies the deficit syndrome.
Suicidality	Expressed desire, intent or actions to harm or kill self.
Guilt*	Overconcern or remorse from past behavior. Rate only patient's statements, do not infer guild feelings from depression, anxiety, or neurotic defenses.
Hostility*	Animosity, contempt, belligerence, threats, arguments, tantrums, property destruction, fights and any other expression of hostile attitudes or actions. Do not infer hostility from neurotic defenses, anxiety or somatic complaints. Do not include incidents of appropriate anger of obvious self-defense.
Elevated Mood	A pervasive, sustained and exaggerated feeling of well-being, cheerfulness, euphoria (implying a pathological mood), optimism that is out of proportion to the circumstances. Do not infer elation from increased activity or from grandiose statements alone.
Grandiosity*	Exaggerated self-opinion, self-enhancing conviction of special abilities or powers or identity as someone rich or famous. Rate only patient's statements about himself, not his demeanor.
Suspiciousness*	Expressed or apparent belief that other personas have acted maliciously or with discriminatory intent. Include persecution by supernatural or other nonhuman agencies (e.g., the devil).
Hallucinations*	Reports of perceptual experiences in the absence of relevant eternal stimuli. When rating degree to which functioning is disrupted by hallucinations include preoccupation with the content and experience of the hallucinations, as well as functioning disrupted by acting out on the hallucinatory content. Include thought aloud, or pseudohallucinations if a voice quality is present.
Unusual Thought Content*	Unusual, odd, strange or bizarre thought content. Rate the degree of unusualness, not the degree of disorganization of speech. Delusions that are patently absurd, clearly false or bizarre ideas that are expressed with full conviction. Consider the patient to have full conviction if he/she has acted as though the delusional belief were true. Ideas of reference/persecution can be differentiated from delusions in that ideas are expressed with much doubt and contain more elements of reality. Include thought insertion, withdrawal and broadcast. Include grandiose, somatic and persecutory delusions even if rated elsewhere.
Bizarre Behavior	Reports of behavior which are odd, unusual or psychotically criminal. Not limited to interview period. Include inappropriate sexual behavior and affect.
Self-neglect	Hygiene, appearance, or eating behavior below usual expectations, below socially acceptable standards, or life-threatening.
Disorientation**	Does not comprehend situations or communications, such as questions asked during the entire BPRS interview. Confusion regarding person, place, or time. Do not rate if incorrect responses are due to delusions.

Table 4

The Brief Psychiatric Rating Scale (continued)

Scale item	Item description
Conceptual Disorganization*	Degree to which speech is confused, disconnected, vague or disorganized. Rate tangentiality, circumstantiality, sudden topic shifts, incoherence, derailment, blocking, neologisms, and other speech disorders. Do not rate content of speech.
Blunted Affect*	Restricted range in emotional expressiveness of face, voice, and gestures. Marked indifference or flatness even when discussing distressing topics. In the case of euphoric or dysphoric patients, rate Blunted Affects if a flat quality is also clearly present.
Emotional Withdrawal*	Deficiency in patient's ability to relate emotionally during interview situation. Use your own feeling as to the presence of an "invisible barrier" between patient and interviewer. Include withdrawal apparently due to psychotic processes.
Motor Retardation*	Reduction in energy level evidenced by slowed movements and speech, reduced body tone, decreased number of spontaneous body movements. Rate on the basis of observed behavior of the patient only. Do not rate on the basis of patient's subjective impression of his own energy level. Rate regardless of the medication effects.
Tension*	Observable physical or motor manifestations of tension, "nervousness," and agitation. Self-reported experiences of tension should be rated under the item on anxiety. Do not rate if restlessness is solely akathisia, but do rate if akathisia is exacerbated by tension.
Uncooperativeness*	Resistance and lack of willingness to cooperate with the interview. The uncooperativeness might result from suspiciousness. Rate only uncooperativeness in relation to the interview, not behaviors involving peers and relatives.
Excitement**	Heightened emotional tone, or increased emotional reactivity to interview or topics being discussed, as evidences by increased intensity of facial expressions, voice tone, expressive gestures or increase in speech quality and speed.
Distractibility	Degree to which observed sequences of speech and actions are interrupted by stimuli unrelated to the interview. Distractibility is rated when the patient shows a change in the focus of attention or marked shift in gaze. Patient's attention may be drawn to noise in adjoining room, books on shelf, interviewer's clothing etc. Do not rate circumstantiality, tangentiality or flight of ideas. Also, do not rate rumination with delusional material. Rate even if the distracting stimulus cannot be identified.
Motor Hyperactivity	Increase in energy level evidenced in more frequent movement and/or rapid speech. Do not rate if restlessness is due to akathisia.
Mannerisms and Posturing*	Unusual and bizarre behavior, stylized movements or acts, or any postures which are clearly uncomfortable or inappropriate. Exclude obvious manifestations of medication side-effects. Do not include nervous mannerisms that are odd or unusual.

Note: All items contained herein were taken from the current BPRS-E, (Ventura, Lukoff, Neuchterlein, Liberman, Green, & Shaner, 1993). * = items that comprised the original 16-item BPRS (Overall & Gorham, 1962). ** = items that were added onto the original 16-item BPRS which made up the 18-item BPRS (Overall & Klett, 1972).

they had hallucinations infrequently but it impacted their level of functioning in most areas (e.g., getting a job, maintaining a job, paying bills, basic hygiene), or if both conditions were reported.

Table 5

Example of 7-point Likert scale with Anchor Points

Hallucinations	
Scale	Anchor points
N/A	Not assessed
1-Not present	No symptoms of Hallucinations are present
2-Very Mild	While resting or going to sleep, sees visions, smells odors, or hears voices, sounds or whispers in the absence of external stimulation, but no impairment in functioning.
3-Mild	While in a *clear state of consciousness*, hears a voice calling the subjects name, experiences non-verbal auditory hallucinations, formless visual hallucinations, or has *sensory experiences* in the presence of a modality-relevant stimulus infrequently and with no functional impairment.
4-Moderate	*Occasional* verbal, visual, gustatory, olfactory, or tactile hallucinations with no functional impairment OR non-verbal auditory hallucinations/visual illusions more than infrequently or *with impairment*.
5-Moderately Severe	Experiences *daily* hallucinations OR *some areas of functioning* are disrupted by hallucinations.
6-Severe	Experiences verbal or visual hallucinations *several times a day* OR *many areas of functioning* are disrupted by these hallucinations.
7-Extremely Severe	Persistent verbal or visual hallucinations *throughout the day* OR *most areas of functioning* are disrupted by these hallucinations.

Note: *italicized* parts of the anchor points represent a distinct difference from the previous/lower rating of disturbance.

Research implications. The original authors released the BPRS into the public domain in 1965, which likely increased its use since that time. Indeed, since its inception the BPRS has been recognized as "the most commonly used outcome measure for the [SMI] population" (Burlingame & Lee, 2004). However, several research limitations have arisen since the measure has been employed over the last forty-five years. The first limitation relates to which version of the measure is used in research studies. Lachar, Espadas, and Bailley's (2004) review of a

decade (1990-2001) of literature using the BPRS report that only half of the published articles specified which version was used and of those that specified at least 31 used the original 16-item version. Since a comparison of the different versions has not appeared in any published report conclusions drawn from one version may not apply to other versions. Additionally, due to the lack of reporting which BPRS version was utilized it is impossible to conduct post-hoc comparisons.

A second research limitation with the BPRS is the use of anchor points as opposed to the original version which used a Likert scale with no anchor points. Rhoades and Overall (1988) elucidated potential problems that may arise from the use of anchor points, namely: anchored scales have not been subjected to extensive factor analytic studies; anchored scales may be less sensitive to drug effects; behavioral anchors may reduce the ability to score subtle, but observable, change; and hundreds of studies have demonstrated that the original BPRS, which did not have anchor points, has been used as a sensitive measure of treatment effects. Although research has supported the BPRS's ability to detect "treatment effects," until recently sensitivity to change at the item, subscale, and total score had not been empirically established. In response to Rhoades and Overall's (1988) caution, Burlingame, Seaman, and Johnson (2006) recently reported that indeed the BPRS-E with an anchored, 7-point Likert scale had similar factor analytic results to previous versions and was sensitive to even small changes. Indeed, their research empirically establishes sensitivity to change on the BPRS-E at all levels of assessment. Therefore, in an effort to increase reliability, it is recommended that future research utilize versions of the BPRS with empirically supported anchored scales.

Reliability. Because reliability levels set the upper limits of the validity of a measure, a high reliability level for the BPRS is essential. The BPRS is also unique because it is both an

outcome measure and a judge-rated measure; thus, the variance of the measure does not solely lie within the measure itself; rather, it also lies with the raters using the scale. Therefore, the critical elements of reliability that, at a minimum, must be established are test-retest reliability and interrater reliability.

Test-retest reliability. Despite the noted importance of test-retest reliability of outcome measures, to date there has been no published study that adequately supports this reliability of the BPRS. This is unfortunate, given how extensive this measure is utilized in both research and clinical practice. Indeed, a double-blind, multiple-rater study incorporating two prior-to-treatment and one after-treatment assessment would lend support to the test-retest reliability of the BPRS.

Interrater reliability. Although the test-retest reliability of the BPRS remains elusive, interrater reliability of the BPRS has been addressed at length in the research literature. As Flemenbaum and Zimmerman (1973, p. 784) noted, "the principle factors determining the reproducibility of the ratings are probably associated not with the rating scale but with the rater using the rating scale." They continued with, "These factors include experience in the use of subjective rating scales, rater-patient rapport, the ability to elicit pathology in a structured interview, and general clinical experience." Accordingly, several authors have argued for the necessity of proper training, familiarity of the constructs, experience rating with the BPRS, firm adherence to the definitions outlined for the BPRS items, and rater-drift training (Burlingame et al., 2005; Overall & Gorham, 1962; Overall & Klett, 1972;).

Hedlung and Vieweg (1980) performed an extensive review of the reliability of the BPRS showing typical interrater reliability indices near 0.85. However, Flemenbaum and Zimmerman (1973) identified four sources of error variance when repeatedly rating psychopathology,

namely: day-to-day patient variance and emphasis on specific bits of behavior, systematic differences in the way raters observe and interpret bits of behavior, random errors of observation and rating, and when the rater makes an inference from his/her rating pattern. Thus, Overall and Gorham (1962) insisted that ratings with the BPRS require trained professionals who are knowledgeable of the symptoms and psychopathology as it applies to the SMI population, weekly calibration sessions and, in concordance with the calibration sessions, multiple raters observing a joint interview and independently completing ratings. However, due to practical limitations of the BPRS this method is rarely used.

Faustman (1994) recommended a more ecological-friendly standard of increasing interrater reliably. He suggested the following: 1) adequate training and frequent calibration of ratings; 2) multiple raters observing a joint interview and independently completing ratings that are subsequently averaged; 3) and the development of behavioral anchors for each level of severity. Although the development of behavioral anchors may have helped with the interrater reliability, little control over the anchor descriptions or types of anchors have been employed. Several versions of the anchors, some non-overlapping, have been developed and used in research and practice (Bech, Larsen, & Andersen, 1988; Gabbard et al., 1987; Woerner, Mannuzza, & Kane, 1988). Fortunately, some recent work has been done to standardize the administration, training, and symptom anchors.

Ventura, Green et al. (1993) developed a 4-step, "gold standard" training program (Burlingame & Lee, 2005) using the BPRS-E with symptom anchors. First, raters must become familiar with the BPRS by mastering the definitions of each symptom and associated anchor points. The training ought to be conducted by one who has extensive experience in using the BPRS-E in a clinical setting. Second, raters watch six of eleven training videos produced at

UCLA; these 11 tapes represent one "gold standard" found in the extant literature. The 11 training videos vary in the difficulty of rating patient symptoms. Each video is a recording of one interviewer and three behind-the-scene raters (all whom have extensive experience in using the BPRS-E in a clinical setting). All four rate the patient and come to a consensus on each rating. The trainees are required to demonstrate an interrater reliability of 0.80 at the item and total score level with the "gold standard" raters before continuing to the next step of training. Step three requires raters to evaluate actual patients with an experienced BPRS rater. This step trains skills that include: "(1) establishing and maintaining rapport, (2) follow-up probing, (3) structuring the interview, and (4) eliciting information on symptom frequency and functional impairment." Step four is necessary for optimal reliability. It is recommended that quality assurance checks be conducted at least once a year to prevent rater-drift and to make certain that the trainees' interviewer style remains consistent and unbiased (Burlingame et al., 2005).

Validity. Evidence of the validity for the BPRS has been well established in the literature in all critical areas—content, criterion-related and construct validity.

Content validity. Ventura, Lukoff et al.'s (1993) version of the BPRS-E provides a manual with recommended semi-structured, opening questions. For example, the following are initial assessment questions accompanied with their respective BPRS items: depression—"How has your mood been recently?"; hostility—"Have you been irritable or grumpy lately?"; grandiosity—"Is there anything uniquely special about you?"; suspiciousness—"Are you concerned about anyone's intentions toward you?"; unusual thought content—"Are thoughts put into your head that are not your own?" From these examples, it can be reasonably concluded that face validity is high.

Criterion-related validity. Given that the BPRS is most often utilized with the SMI population, an important predictive validity criterion would be the admittance or readmittance into an inpatient mental health hospital. Lachar et al. (2004) reviewed 2,068 patients hospitalized who were classified into four readmittance groups: within 90 days (n = 279), between 4 to 12 months (n = 295), greater than one year (n = 304) and not readmitted (n = 1,190). The authors found that BPRS total scores, resistance, and positive symptoms predicted readmissions, where the higher the ratings the more likely patients were readmitted. Others authors have found similar predictive results as well (Breier, Schreiber, Dyer, & Pickar, 1991; Hobbs et al., 2000; Hopko, Lachar, Bailley, & Varner, 2001; Nicholson & Feinstein, 1996; Olfson, Mechanic, Hansell, Boyer, & Walkup, 1999; Swett, 1995). Indeed, the predictive validity of the BPRS is well established and ought to be incorporated in the decision process when making discharge and treatment decisions.

Construct validity. Construct validity has been well documented in empirical literature for the BPRS. Specifically, the discriminate and convergent validity, factor analytic validity, and sensitivity to change have been documented in hundreds of journal articles. The aforementioned validities are particularly relevant because the BPRS is a multitrait, outcome assessment measure.

Discriminate validity. Hundreds of articles support the discriminate validity of the BPRS (Faustman, 1994). Because the BPRS is a multitrait measure it is important that ratings on the BPRS that are associated with diagnostically distinct disorders are indeed discriminate from each other. For example, different subscale and total score elevations should be evident when assessing one diagnosis (e.g. schizophrenia) compared to another (e.g. depression). Indeed, patients with the most severe psychopathology (schizophrenia, schizoaffective disorder)

consistently produce higher BPRS total scores, positive symptoms, negative symptoms, resistance, and activation than those experiencing pathologies which are less often associated with such characteristics (e.g., major depression disorder, bipolar disorder, substance abuse; Abel, O'Keane, Muray, & Cleare, 1997; Averill, Hopko, Small, Greenlee, & Varner, 2001; Blanchard, Bellack, & Mueser, 1994; Lachar et al., 2001). Additionally, several studies have documented the discriminate validity on a subscale level comparing schizophrenia and mood disorders (Abel et al., 1997; Averill et al., 2001; Blanchard et al., 1994; Merrin & Floyd, 1997; Lachar et al., 2001; Papassotiropoulos, Hawellek, Frahnert, Rao, & Rao, 1999; Silverstein, Harrow, & Bryson, 1994), schizophrenia and substance abuse (Lachar et al., 2001), and specific types of schizophrenia (e.g., paranoid, undifferentiated, and schizoaffective; Lachar et al., 2004). Lachar et al. (2004) examined a substantial sample of inpatients with schizophrenic-spectrum disorders (n = 728) and nonschizophrenic psychiatric patients (n = 1,410). They found that positive and negative symptoms did not differ among the schizophrenic sample, but were significantly elevated relative to the contrasting sample. As expected, the BPRS distinguishes between symptoms and diagnostic categories which support the discriminate validity of the measure.

Convergent validity. Evidence of the BPRS convergent validity with other measures of corresponding psychopathology is well supported in the literature with studies ranging from depressive symptoms, positive and negative symptoms, brain morphology, neuropsychological functions, and cognitive ability (Faustman, 1994; Lachar et al., 2004). For example, comparisons of the Hamilton Rating Scale for Depression (HAM-D; Hamilton, 1960) and the BPRS have shown high correlations. Three studies comparing the negative symptoms subscale of the BPRS to the HAM-D have produced correlations of 0.79 (Craig, Richardson, Pass, &

Bregmann, 1985), 0.80 (Newcomer, Faustman, Yeh, & Csernansky, 1990), and 0.67 (McAdams, Harris, Bailey, Fell, & Jeste, 1996). Additionally, specific factors of the HAM-D (retardation, loss of insight, decreased work interest) correlated at 0.80 with the BPRS negative symptoms subscale suggesting the parts of the BPRS that theoretically ought to be related to the HAM-D in fact do (Goldman, Tandon, Liberzon, & Greden, 1992). Furthermore, the BPRS psychological discomfort subscale correlated at 0.81 with the Calgary Depression Scale for Schizophrenia, a depression measure specific for schizophrenia (Kontaxakis et al., 2000). Related to depression, self-completed quality of life measures have consistently correlated with the BPRS total score, subscales, and item levels (Begtsson-Tops & Hansson, 1999; Bow-Thomas, Velligan, Miller, & Olsen, 1999; Hansson et al., 1999; Meltzer, Burnett, Bastani, & Ramire, 1990; Packer, Husted, Cohen, & Tomlinson, 1997; Priebe, Huxley, Knight, & Evans, 1999; Rudolf & Priebe, 1999). Clearly, the BPRS exhibits an expected positive relationship to other measures of depressive and positive and negative symptoms.

Given the increase of neurological imaging, testing, neurobiological, and biological epidemiology for psychological disorders and the change thereof, convergent validity with such explanations would add significant support for the validity of the BPRS. Young et al. (1991) reported significant correlations between the BPRS positive symptoms subscale and ventricle volume and the BPRS negative symptoms subscale and the size of the caudate nucleus on the left and right side. These finding agree with earlier research suggesting that ventricle volume of patients with schizophrenia were higher and that they have decreased volume in the caudate nucleus when compared to controls (Pearlson et al., 1989). Furthermore, the ventricle-brain ratio correlate with the BPRS total score (r = 0.46), depressive mood (r = 0.48), blunted affect (r = 0.47), conceptual disorganization (r = 0.41), emotional withdrawal (r = 0.40), and motor

retardation (r = 0.38) of inpatients with major depression (Schlegel, Frommberger, & Buller, 1989). Finally, in a study of major depression, several BPRS items (somatic concern, anxiety, guilt feelings, tension, depressive mood, and motor retardation) correlated negatively with evoked potential amplitudes (Shagass & Roemer, 1992), which was subsequently negatively correlated to severity of depressive symptoms (Mathalon, Ford, & Pfefferbaum, 2000). Clearly, the use of the BPRS is becoming a vital outcome measure of psychological symptoms as they relate to biological and neurobiological changes.

It is widely noted that severe pathology is also related to significant cognitive and neuropsychological functioning deficits. As such, it would be expected that certain patterns of elevation on the BPRS would relate to such deficits as well. Indeed, the BPRS negative symptoms subscale is correlated with poor executive functioning (Poole, Ober, Shenaut, & Vinogradov, 1999), omissions, errors of commission, and poor eye tracking on the Continuous Performance Test (Roitman, Keefe, Harvey, Siever, & Mohs, 1997), as well as neuropsychological deficits in general (Silverstein et al., 1994), poorer neuropsychological competence and adaptive competence (Velligan et al., 1997), temporal sequencing of component actions of social situations (Corrigan & Addis, 1995), social cue recognition (Corrigan, Hirschbeck, & Wolfe, 1995), and WAIS-R digit span (Zakanis, 1998). In contrast, the BPRS positive symptoms subscale is correlated with fewer completed categories and perseverative errors on the Wisconsin Card Sorting Test (Ragland et al., 1996), poorer performance on the Mini Mental Status Examination (Ownby, Koss, Smyth, & Whitehouse, 1994), and longer performance times on the Trails B (Zakanis, 1998). Zakanis (1998) concluded that the BPRS positive symptoms subtest was associated with poorer frontal lobe function while the BPRS negative symptoms subtest was associated with right hemisphere deficits. In general, the BPRS

negative symptoms classification was associated with the broadest and most significant neuropsychological deficits (Mahurin, Vellingan, & Miller, 1998), which corresponds to clinical experience. It is largely substantiated that the BPRS and corresponding subscales correlate with biological and neuropsychological functioning as each are associated with severe psychopathology.

Factor analytic validity. Because the BPRS is a multitrait outcome measure ratings can be tabulated at the global (total score), dimensional (subscales), and item level. Indeed, an inpatient at a psychiatric unit theoretically could have a relatively low total score and still meet admission requirements. However, few studies include multiple levels of reporting. Lachar et al. (2004) examined 813 studies between 1990 and 2001, and found that the total score was reported 83%, subscale 58%, and item 19% of the time. The importance of reports at multiple levels can be found in Dell'Osso et al., (2000) who used a contrasted group design. They report that total score ratings did not differ between groups of patients in the manic, mixed, or depressive episodes of bipolar, but did differ at the subscale level. Furthermore, Burlingame et al., (2006) found no difference in sensitivity to change between subscales but did at the item level. These findings underscore the importance of reporting at each level of analysis.

Nicholson, Chapman, and Neufeld (1995) found that the use of BPRS subscales became standard practice, although different item combinations were used to construct similarly named subscales. For instance, they identified four versions of "Negative Symptoms" and nine versions of "Positive Symptoms" in their review. Four additional complications have been noted in the development of BPRS subscales. The first is the use of BPRS measures with different numbers of items (16, 18, or 24) and the presence of, or lack of, anchor points. Since there are multiple versions of the measure published (Burlingame & Lee, 2004) there is an increased risk of

researchers and practitioners using subscales developed for an alternative version. Another complication is the use of limited sample sizes. Comrey and Lee (1992) recommended a sample size greater than 300 subjects to be an adequate sample size for a factor analysis. However, most of the studies reporting factor analysis on the BPRS fail to meet this requirement (most had between 100 and 200 participants). A third complication relates to the context of outcome assessment and the necessity of performing a factor analysis over time. Indeed, all of the factor analyses reported in the summary by Lacher et al. (2004) fail to perform the analysis over time or fail to confirm the structure of the factor over time. Lachar et al. (2004) provided yet another explanation for item loading differences. They suggested that sample characteristics may contribute to the methodological problems associated with the factor analysis (e.g. characteristics of Alzheimer's disease contributed to patients rating fewer than 90% of the items as reported in Ownby, Koss, Smyth, & Whitehouse, 1994).

Although several methodological hurdles are associated with the development of a consistent factor structure, four or five factors appear to be the most common structure of the BPRS-E (Burger, Calsyn, Morse, Klinkenberg, & Trusty, 1997; Dingemans, Linszen, Lenior, & Smeets, 1995; Long & Brekke, 1999; Morlan & Tan, 1998; Mueser, Curran, & McHugo, 1997; Ownby & Seibel, 1994). For example, Lachar et al. (2004) examined twelve published factor analyses noting that seven report a four-factor solution and four reported a five-factor solution. Ventura, Nuechterlein, Subotnok, Gutkind, and Gilbert (2000) reported a four-factor solution for both the 18- and 24-item versions of the BPRS. Across studies, the most commonly cited factors are Thought Disturbance, Anxiety-Depression, Withdrawal, Hostile-Suspicious, and Activity/Mania as shown in Table 6 (Burlingame et al., 2006). Thus, while discrepancies exist in

particular item loadings and between particular patient populations there is a growing consensus in the general factor structure of the BPRS.

Table 6

Common Factor Dimensions of the BPRS

Factor	BPRS Items
Thought Disturbance/ Positive Symptoms	8 Grandiosity 9 Suspiciousness* 10 Hallucinations 11 Unusual Thought Content 15 Conceptual Disorganization
Anxiety-Depression	1 Somatic Concern 2 Anxiety 3 Depression 5 Guilt
Withdrawal, Negative symptoms	14 Disorientation 16 Blunted Affect 17 Emotional Withdrawal 18 Motor Retardation
Hostile Suspicious/ Paranoid	6 Hostility 9 Suspiciousness* 20 Uncooperativeness*
Activity/ Mania	19 Tension 20 Uncooperativeness* 21 Excitement 24 Mannerisms and Posturing

Note: Adapted from Burlingame et al. (2006). * = item was grouped with different factors on separate studies.

Sensitivity to change. Several studies have shown that individual items and subscales on the BPRS are sensitive to antipsychotic medications (Borison, Sinha, Haverstock, McLarnon, & Diamond, 1989; den Boer et al., 1990; Nair et al., 1986), the exacerbating effects of caffeine with schizophrenic patients (De Freitas & Schwartz, 1979; Lucas et al., 1990), the effects of electroconvulsive therapy (Abraham & Kulhara, 1987), anticholinergic medications (Tandon, Mann, Eisner, & Coppard, 1990), and medicinal treatment for depression (Feigher, Merideth, & Claghorn, 1984). Additionally, Burlingame et al. (2006) demonstrated the sensitivity of the BPRS on both an item and subscale level. They tracked 223 adult psychiatric inpatients from 1999 to 2001 and reported that 22 of the 24 items (excluding elevated mood and mannerisms and posturing items) and all the subscales of the BPRS to be significantly sensitive to change. As evidenced, the BPRS shows exceptional sensitivity to change at the item, subscale, and total

score level and in the context of a multitude of treatment or exacerbating effects. The ability for the BPRS to distinguish diagnostically different patients, severity of different diagnoses, and the sensitivity to change is evidence of robust construct validity.

Summary. One way to summarize the BPRS empirical literature is to use an evaluative framework for examining the strengths and weaknesses of outcome measures. The following is a summary of the BPRS using Faustman's (1994) and Ciarlo, Brown, Edwards, Kiresak, and Newman's (1986) analysis:

1) An ideal outcome measure is useful in a wide range of settings and client samples (Ciarlo et al., 1986). Although the BPRS was originally developed for the use of inpatient populations, the scale can be used in outpatient settings (Pull & Overall, 1977) and is becoming a favored measure in research settings treating those as identified as SMI (Burlingame et al., 2005). Furthermore, the measure can be applied to several different diagnostic categories.
2) With the advent of Ventura, Lukoff et al.'s (1993) BPRS-E, which provides behavioral anchors, the scale has "clear and objective referents (meanings) that are consistent across clients, to insure interpretability of the individual and group scores and score changes" (Ciarlo et al., 1986, p. 52). Although more work on anchored versions is needed, Burlingame et al., (2006) found that the BPRS-E factor structure, reliability, and sensitivity to change apply to an anchored, 7-point Likert scale.
3) Measures should reflect "the perceptive of all relevant participants in the treatment process" (Ciarlo et al., 1986, p. 52). In this respect the BPRS is somewhat unique. Although the BPRS is primarily clinician-based, Ventura, Lukoff et al., (1993) emphasized that clinicians should take into consideration behavioral and weekly notes

of those who have more frequent contact with the patient (e.g., psych tech, psychiatric nurse). Thus, the BPRS is specifically designed and training encourages the use of multiple sources from the treatment team.

4) Outcome measures should have demonstrated reliability, validity, sensitivity to change, and freedom from bias (Ciarlo et al., 1986; Hill & Lambert, 2004; Lipsey, 1990; Vermillion & Pfeiffer, 1993). The BPRS demonstrates exceptional reliability and validity. Ongoing monitoring of interrater reliability is encouraged and built into the extant BPRS training modules, although test-retest deserves empirical support. Hundreds of studies have demonstrated the validity of the BPRS in several different settings and with a broad base of psychiatric inpatient, outpatient, and community participants. The validity has been tested with a breadth of methods and populations. More recently, the BPRS has demonstrated excellent sensitivity to change, an often overlooked validity aspect, on the item, subscale, and total scale levels. However, the BPRS lacks in its ability to be free from bias because "clinician-based instruments may be influenced by the expectations and hopes of the rater" (Faustman, 1994, p. 391). It is recommended that raters be blind to the treatment condition, severity of the pathology, and diagnosis of the patient (Faustman, 1994; Ventura, Lukoff et al., 1993) whenever possible to minimize potential biases.
5) The scale can be utilized in a variety of ways and situations (Ciarlo et al., 1994). The symptom constructs are familiar to those with a wide range of experience, it is easy to score and interpret, does not require sophisticated statistical analyses for scoring, can be graphed easily, and is easily integrated into a standard interview (Faustman, 1994).

6) An outcome measure should be compatible with a wide range of theoretical orientation and theories of psychopathology (Ciarlo et al., 1994). In its original development and when properly used the BPRS can be easily modified to fit one's particular theory of pathology and practice.
7) Outcome measures should be normed so clinicians can make an assessment of the clinical significance of treatment effects in addition to statistical significance by establishing cutoff scores and a reliable change index (Burlingame et al., 2005; Hill & Lambert, 2004; Jacobson et al., 1984; Jacobson & Truax, 1991; Vermeersch et al., 2000; Vermillion & Pfeiffer, 1993). Although some research has identified a reliable change index and cut scores (Lee, Rees, Burlingame, Hwang, & O'Neal, 2005), more research is necessary to empirically establish these indices for the BPRS.
8) Outcome measures should be able to be completed in a matter of minutes and should not be an extra burden on clients (Burlingame et al., 1995; Burlingame et al., 2008; Hatfield & Ogles, 2004, 2006; Hill & Lambert, 2004; Lipsey, 1990). This is one of the shortcomings with the BPRS. It takes, with an experienced clinician, approximately 30 to 40 minutes to conduct a thorough BPRS assessment. Although some of the information may be gleaned from a clinical interview, much of the specific information, follow-up questioning, and review of collaborative sources adds a burden on the clinical staff.
9) Outcome measures should be easily scored and interpretable (Burlingame et al., 1995; Hill & Lambert, 2004). The BPRS requires minimal mathematical procedures. Indeed, the measure can be scored while assessment takes place and immediate feedback is available on a total, subscale, and item level if necessary.

10) Outcome measures should minimize the paperwork or otherwise taxing the available human resources (Hatfield & Ogles, 2004, 2006). Once the BPRS manual is mastered by an experienced rater, paper work is quite minimal.
11) Outcome measures should allow frequent use to track progress, or monitor treatment (Burlingame et al., 1995; Hatfield & Ogles, 2004, 2006, 2007) so that treatment can be altered accordingly or problem areas can be identified dynamically rather than waiting an extended period of time or until the end of treatment before making an assessment. Although the BPRS was developed with the expectation of frequent use, the time and resources it takes to complete often necessitate that assessment periods be spaced out. This is problematic because the further spaced out the assessment time the less likely it will detect acute exacerbation of pathological symptoms.
12) An outcome measure should be cost efficient (i.e., inexpensive) so it does not adversely impact the consumer's financial resources (Burlingame et al., 1995; Hatfield & Ogles, 2004, 2006, 2007). Largely due to the resources required to administer the BPRS, it remains somewhat cost inefficient.

The Severe Outcome Questionnaire

Development. As part of the OQ family, the SOQ was conceptualized and developed to address the specific limitations the OQ had with the SMI population in a variety of settings (e.g., inpatient, outpatient). Specifically, the SOQ is a 45-item, self-completed measure composed of 30 items from the OQ-30 (Lambert, Finch, Okiishi, & Burlingame, 2005) and an additional 15 items based on literature guidelines for SMI patients, as shown in Table 7 (Burlingame et al., 2008; Carey, 2000). The items adopted from the OQ-30 represent the most sensitive items on the original OQ-45 (Lambert et al., 2005; Vermeersch et al., 2000). Based on Lambert's (1983)

Table 7

Severe Outcome Questionnaire Items

1. I have trouble falling asleep or staying asleep.	24. I have trouble at work/school or other daily activities because of drinking or drug use.
2. I feel no interest in things.	25. I feel that something bad is going to happen.
3. I feel stressed at work, school or other daily activities.	26. I feel nervous.
4. I blame myself for things.	27. I feel that I am not doing well at work/school or in other daily activities.
5. I am satisfied with my life.	28. I feel something is wrong with my mind.
6. I feel irritated.	29. I feel blue.
7. I have thoughts of ending my life.	30. I am satisfied with my relationships with others.
8. I feel weak.	31. I see or hear things that other people don't.*
9. I find my work/school or other daily activities satisfying.	32. I can't stop talking, moving or doing things.*
10. I feel fearful.	33. I have been told by others that my behavior is out of control.*
11. I use alcohol or a drug to get going in the morning.	34. I must do things like was my hands or hurt myself to feel better.*
12. I feel worthless.	35. I have difficulty with my unstable moods.*
13. I am concerned about family troubles.	36. My temper leads me to act without thinking, or say things that I don't mean.*
14. I feel lonely.	37. I am not in control of my life.*
15. I have frequent arguments.	38. I am forgetful.*
16. I have difficulty concentrating.	39. I have been told that I have difficulty keeping myself neat and clean.*
17. I feel hopeless about the future.	40. I think people are trying to make it difficult for me to succeed.*
18. I am a happy person.	41. I feel confused.*
19. Disturbing thoughts come into my mind that I cannot get rid of.	42. I have difficulty completing my household chores like shopping, cooking and cleaning.*
20. People criticize my drinking (or drug) use.	43. I think I am really ill.*
21. I have an upset stomach.	44. I have problems making daily decisions.*
22. I am not working/studying as well as I used to.	45. I have difficulty keeping jobs or managing money.*
23. I have trouble getting along with friends and close acquaintances.	

* indicates SMI specific items.

conceptualization of psychopathology the items derived from the OQ-30 measure patient change along four dimensions: 1) subjective discomfort (intrapsychic functioning), 2) interpersonal relationships, 3) social role performance, and 4) severe functional impairment. Burlingame et al.

(2008) indicated, "The S-OQ is conceptualized as having three levels of usage for a severely mentally ill (SMI) population: 1) To measure current level of distress; 2) as an outcome measure to be administered prior to and following treatment interventions or to monitor ongoing treatment response; and 3) to accompany computerized decision support tools to improve the quality of patient care." Overall, the SOQ was developed to address ceiling effects associated with the OQ-30 when used with the SMI population and to provide an efficient, reliable, and valid method of monitoring change within the SMI population.

Normative data. Normative data for the SOQ have been drawn from a limited number of studies compared to its predecessor (the OQ-30). Table 8 depicts the means and standard deviations for different patient settings. As indicated, SOQ mean scores for the non-patient

Table 8

Admission Scores by Setting

	Non-patient (n = 228)		Inpatient (n = 312)		Outpatient (n = 1110)	
Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*
SOQ	36.00	17.26	75.27	27.07	93.06	26.34
OQ-30	27.61	12.73	54.05	19.34	65.93	17.27
SMI-15	8.39	5.60	23.70	11.59	27.13	10.78

Note: Extracted from Burlingame et al. (2008)

population are significantly lower than the inpatient and outpatient samples, as expected. Of note, outpatient means are higher than inpatient means. This difference may be explained by the demand characteristics of the inpatient population (e.g., reality impairment, resistance), which could lower the overall mean when patients deny pathological symptoms; however, more research is needed to support this explanation. Table 9 shows the comparative change from

Table 9

Change Scores by Setting

		SOQ					
		Admission		Discharge		Change	
Setting	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Inpatient	255	76.52	28.12	45.23	24.67	32.13	30.36
Outpatient	287	91.12	25.70	86.80	30.87	4.31	38.25

		OQ-30					
		Admission		Discharge		Change	
Setting	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Inpatient	255	53.55	20.81	32.74	16.58	20.81	20.70
Outpatient	287	64.90	16.51	61.98	20.16	2.92	25.01

Note: Adapted from Burlingame et al. (2008)

admission to discharge between the SOQ and the OQ-30. As expected the SOQ has greater admission and discharge means and greater change means compared to the OQ-30, which suggests that the goal of the SOQ to extend the ceiling effects may have been accomplished.

Reliability. Although test-retest reliability of outcome measures is recommended as an essential psychometric property (Hill & Lambert, 2004), this has yet to be empirically established for the SOQ. Lambert, Hatfield et al. (2001) reported the OQ-30 (a subset of the SOQ) has a test-retest reliability of 0.80 to 0.84 with a non-patient sample. Because the first 30 items of the SOQ are composed of the same items as the OQ-30, this can be used as an estimate of the test-retest reliability. However, it is stressed that empirical support on the test-retest reliability on the SOQ should be completed in the context of the SMI population. Table 10 shows the internal consistency reliabilities (ICR; Chronbach's Alpha) of the SOQ, OQ-30, and the 15

SMI items. As shown the ICR of the SOQ is high, which suggests all the items measure the same construct. Although these results suggest high reliability, more work on test-retest reliability is needed.

Table 10

Internal Consistency Reliabilities of the SOQ

	Non-Patient (n = 144)	Inpatient (n = 312)	Outpatient (n = 1110)	Total (n = 1586)
SOQ	0.93	0.93	0.94	0.95
OQ-30	0.91	0.91	0.91	0.94
SMI-15	0.82	0.87	0.86	0.89

Note: extracted from Burlingame et al., (2008)

Validity. Validity studies continue to be a major research focus on the SOQ. In the most thorough study of the SOQ to date, Carey (2000) reported the convergent validity of the SOQ with the Behavior and Symptom Identification Scale (BASIS-32) and the Nurses' Observation Scale for Inpatient Evaluation (NOSIE-30). As shown in Table 11, the SOQ as a complete measure, the 30 items from the OQ-30 and the 15 SMI items correlated significantly to each criterion measure. Although some research has shown the validity of the SOQ, more is necessary to fulfill outcome assessment recommendations. Indeed, the purpose of this project is to explore the relationship between the SOQ and the gold standard outcome measure for the SMI population, the BPRS.

Table 11

SOQ Correlations with Criterion Measures

	BASIS-32 (*n* = 147)	BPRS-E (*n* = 83)					NOSIE-30 (*n* = 79)
	Total	Thought	Anergia	Total	Disorg.	Total	Total
SOQ	.90*	0.04	0.07	-.44*	0.00	.43**	-.44*
OQ-30	.86*	-0.03	0.03	-.44*	-0.04	.37**	-.44*
SMI-15	.87*	0.15	0.15	-.38*	0.08	.47**	-.38*

Note: Extracted from Burlingame et al. (2008).
* Significant at $p < .01$

Statistical Analyses

Accurate statistical methodology is of utmost importance in outcome research since different statistical models can result in different conclusions with the same set of data (Speer & Greenbaum, 1995). Thus, inaccurate statistical methods may lead to inaccurate results and conclusions. For instance, two commonly used analyses, univariate or multivariate analysis of variance (ANOVA or MANOVA; Raudenbush & Chan, 1993), have been inappropriately used for the evaluation of change where unbalanced designs, missing data, time-varying covariates, or continuous predictors were employed (Bryk & Raudenbush, 1987; Huttenlocher, Haight, Bryk, & Selter, 1991; Mason, Wong, & Entwistle 1983; Ware, 1985). Bryk and Raudenbush (1987) summarized the inadequacies of measuring individual change as the following:

1. *Conceptualization*. In any research context, a model of the phenomena under study is an important heuristic for guiding inquiry. Yet in most previous research on individual change, the model of individual growth is rarely addressed explicitly.
2. *Measurement*. Studies of change typically use tests that are developed to discriminate among individuals at a fixed point in time. Their adequacy for distinguishing the *rate*

of change among individuals is rarely considered during the instrument design process. Further, statistical procedures routinely applied to these instruments, such as standardizing the scores to a common mean and variance over time effectively eliminate the essence of individual growth (Rogosa, Brandt, & Zimowski, 1982). Psychometric procedures are needed that enable assessment of the adequacy of instruments for measuring both status and change.

3. *Design*. Much of the research on change has been based on data on individual status at two points, for example, scores on a pretest and posttest. In general, two time points provide an inadequate basis for studying change (Bryk & Weisberg, 1977; Rogosa et al., 1982). Further even in instances in which data have been collected on multiple occasions, researchers have typically analyzed the data as a series of separate designs with two time points.

Developments in statistical theory of hierarchal linear modeling (HLM) enable simultaneous analysis of individual growth, reliability of outcome measures, correlates of change status, and testing hypotheses about individual growth (Bryk & Raudenbush, 1987, 1992; Raudenbush & Bryk, 2002).

HLM operates on a two-level hierarchal model where level one is *nested* in level two. In studying individual change the level one data is represented by individual growth trajectories which are dependent on specific parameters. The individual change parameters then become level two (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002). For example, level one analysis may be the downward trend of distress of individuals in the study. Those downward trends become the dependent variables and are contained in parameters such as age, sex, or diagnosis, which is represented by the level two analyses. Another way to conceptualize the

hierarchal function of HLM is that the regressions used to represent individual can be used as the dependent variables in higher level regressions (i.e., level two; Arnold, 1992).

A number of advantages of using HLM analysis have been noted in the literature. One type of advantage is the flexibility of HLM. The flexibility of the model allows for easy handling of missing data; specifically, no data needs to be discarded or imputed to make the analysis work (Raudenbush & Chan, 1993). This flexibility allows comparison of change data with time-interval differences (Bryk & Raudenbush, 1987, 1992; Raudenbush & Chan, 1993; Speer & Greenbaum, 1995). Additionally, repeated measures increases the accuracy of regression equations by decreasing the standard errors while providing reliable estimates of parameter correlations (Speer & Greenbaum, 1995). HLM capitalizes on the use of hierarchal, unit nested within unit, method to better estimate regressions in repeated measures (Bryk & Raudenbush, 1992, Raudenbush & Bryk, 2002; Speer & Greenbaum, 1995). Speer and Greenbaum (1995) summarized the advantages of HLM by stating,

> Its greater precision in estimating individual change parameters as a result of using empirical Bayes estimation, the EM [expectation-maximization] algorithm for missing data, and parameter estimates based on mulitwave data. . . using all of the available information to provide better estimates of significant change data.

Hypotheses

Based on the literature review herein presented, the following are the problems this study will attempt to address:

1) The BPRS is the gold standard tool for assessing outcome in SMI patients. However, several drawbacks are noted; namely, it is time consuming to train and maintain reliable raters, it is more costly to collect outcome data than self-completed measures,

and some settings may not have the resources to use a clinician-completed tool. Thus, practical limitations may necessitate less frequent BPRS assessments.

2) Self-completed measures address many of the drawbacks from the BPRS (they are inexpensive, easily implemented, and require minimal time), but a self-completed measure has its own drawbacks when assessing the SMI population. Notably, some of the patients may be too impaired to provide accurate information and malingering and feigning will lower the reliability of the scores. Indeed, experience at the USH with the SOQ suggests that approximately 25% of SMI inpatients produce scores at or below the community normal range. This clearly calls into question the information drawn from some of the patients and calls for an understanding of when to trust and distrust self-reported data from the SMI.
3) Previous outcome research with the SMI population has been methodologically inadequate because they have not explicitly measured individual change, distinguishing on a basis of rate of change, and only compared single-point or pre-post comparisons rather than use HLM.

The purpose of this study is to use a gold standard, clinician-completed outcome measure, the BPRS, to explore the relationship between self- and clinician-completed outcome measures with an SMI population and to further understand those that score at the subclinical range in the SOQ. Thus the following hypotheses are proposed:

1) There will be a positive correlation between the total score of the BPRS-E and SOQ at all measurement points (intake, 90-, 180-, 270 and 360+ days; see Figure 2).

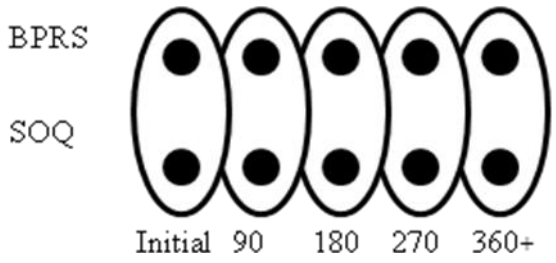


Figure 2

2) Patients who score at the subclinical level on the SOQ at admission will have elevated BPRS-E subscales (e.g. Positive Symptoms) or items associated with cognitive and reality impairment when compared to those in the clinical range at admissions (see Figure 3).

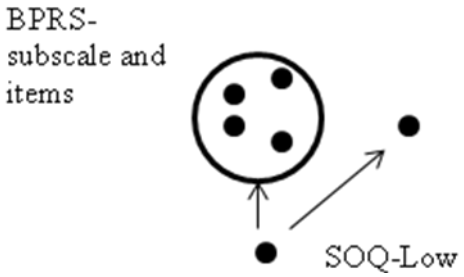


Figure 3

3) The change trajectory for the SOQ will not be significantly different than the change trajectory for the BPRS (see Figure 4).

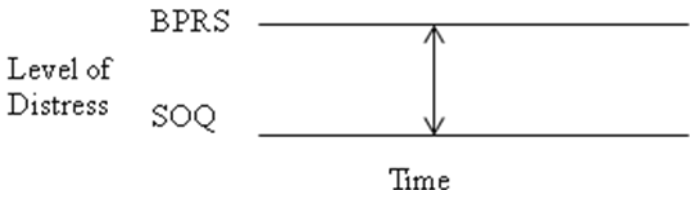


Figure 4

4) Patients who score at the subclinical level on the SOQ at admission will have a less reliable and shallower SOQ change trajectory when compared to those who score in

the clinical level. The corresponding BPRS-E change trajectories will be unaffected (see Figure 5).

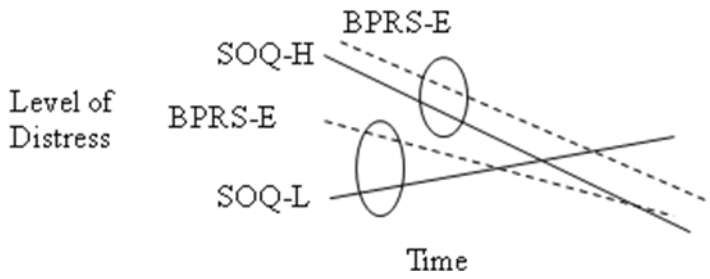


Figure 5

5) There will be distinct subgroups of patients who initially score subclinical at admission. One group will show a reliable increase in scores that meet or exceeds the cutoff score at 90 days. A second group will show reliable increase and meet or exceed the cutoff score after 90 days. A third group will never score in the clinical range and show no reliable change, essentially staying in the subclinical range (see Figure 6).

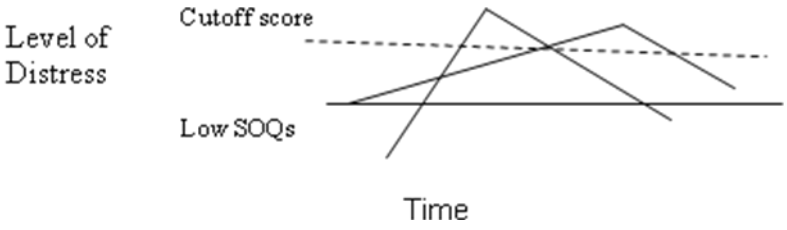


Figure 6

6) There will be individual items on the BPRS that predict membership in one of the groups identified in hypothesis five (see Figure 7).

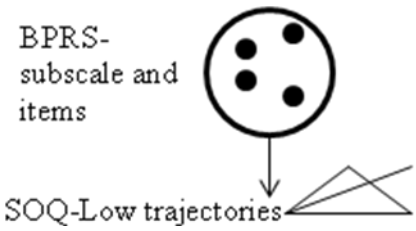


Figure 7

Method

Setting

Utah State Hospital (USH) is Utah's largest government-operated, inpatient mental health hospital with 354 beds. Referrals come from 11 community based facilities across the state. USH provides services for approximately 285 adults, 22 children (ages 6 through 13 years old), and 50 youth (ages 14 through 18 years old). These services include general inpatient care, intensive care, acute rehabilitation, transition living, and forensic services. The hospital provides individual, family, and group therapy; medication evaluation and management; work opportunities; occupational therapy; physical therapy; and crisis interventions. These services are provided by psychologists, psychiatrists, clinical social workers, nurses, physical therapists, psychological technicians, and supporting staff.

USH assessment procedures

The Brief Psychiatric Rating Scale-Expanded Version (BPRS-E; Ventura, Lukoff et al., 1993) and the Severe Outcome Questionnaire (SOQ; Burlingame et al., 2008) are given to patients of USH as part of the hospital's standard outcome assessment procedures. USH's outcome assessment procedures necessitate that an "admission" assessment be completed within 72 hours of admittance. Due to the intensive treatment regimen at USH the time frame of 72 hours is important because patient's symptoms may change as a result of hospital admittance and

treatment. Following an admission assessment, subsequent BPRS-E and SOQ ratings are given approximately on 90-day intervals.

Interviewers consist of licensed psychologists and clinical psychology interns who have completed reliability training for the BPRS-E at USH with an interrater reliability of equal to or greater than 0.80. Training on the BPRS-E follows the rubric provided by Ventura, Green, et al. (1993), which has been recognized as the "gold standard" training protocol for the BPRS-E (Burlingame & Lee, 2005; Burlingame et al., 2005). In addition to the initial training, each interviewer receives "rater drift" training every six months.

Participants

The archival data set for this study was provided by USH as part of their ongoing mission to contribute to research. Adult (18 years old or older) BPRS-Es and SOQs gathered from January 1, 2000 through December 31, 2008 were included in this study. As a result, this data set had a total of 2,180 participants with a total number of 13,139 BPRS-E ratings and 10,219 SOQ ratings.

Measures

The BPRS-E is a 24-item, semi-structured, multivariate, anchored outcome measure (see Appendix A). The semi-structured interview provides brief example questions to begin the interview, while allowing the flexibility of adjusting the interview appropriately. The multivariate characteristic of the measure allows for a quick assessment on a breadth of pathological symptoms. The anchored, 7-point Likert scale ranges from one (*not present*) to seven (*extremely severe*) with an optional N/A (*not assessed*) to measure the severity of each symptom. The patient is asked to consider only the previous 14 days when answering questions.

The SOQ is a 45-item, self-completed, multivariate, anchored outcome measure. The self-completed format allows for quick assessment (approximately five to ten minutes) on a breadth of symptoms common among those who experience psychological distress. Specifically, the SOQ is a multitrait measure that covers a variety of symptoms associated with general psychological functioning with specific items that assess common symptoms among the SMI population. The 4-point Likert scale from zero (*never*) to four (*almost always*) measures the frequency of each symptom. The patient is asked to consider the previous 14 days when answering the questions, although the time frame can be modified to be a weekly assessment measure.

Procedures

Inclusion criteria. Prior to inclusion into the statistical analysis, participants' data were evaluated on several inclusion criteria rules. First, it was necessary that each BPRS-E entry be accompanied by an SOQ entry within a 14-day period. The guideline was chosen because it was necessary for there to be overlap in assessment time periods of each outcome assessment. Second, each participant must have a minimum of three data points. This minimum number was selected because researchers have concluded that when change is non-linear two-time point estimates are inadequate (cf. Bryk & Weisberg, 1977; Rogosa et al., 1982) and the benefits of hierarchical analyses require a minimum of three within-subject data points (Arnold, 1992). Third, there could be no ratings of zero (a numerical indication of N/A) on any of the BPRS-E items. This is because the BPRS-E instructions allow for the use of collaborative information and encourages it when possible. Missing values for the BPRS-E represent, according to the researcher, an inadequate assessment and questionable reliability and fidelity to the outcome

assessment protocol. If all previously outlined criteria were met then that participant's data were included.

There were several occasions where participants were discharged and then readmitted. If two BPRS-E assessments were missing (representing more than 180 days) and there were no SOQ assessments during that time period, then it was assumed that the participant had been discharged on the last BPRS-E date. The following BPRS-E was considered to be a new admission assessment. Each admission was considered an independent rating period; thus, if the participant did not meet the inclusion criteria on one admission but did on a subsequent, then the data on the subsequent data set was included.

Working with the SMI population presents unique problems in research design and data accumulation. One challenge in this archival study was that many of the patients exhibited a biased response style. For example, some patients drew a line down the "never" column (a clear indication that they did not read the items). Anticipating these response sets, the creators of the SOQ worded and reverse scored four items.

An additional problem was data input. When utilizing archival data the researcher often has little control on data input. The data input problem in this study was that some of the data were inputted incorrectly. Where "never" was marked on the SOQ zeros where inputted instead of reverse scoring the proper items. Thus, the client with a response style of marking "never" for each item should have a score of 16, yet few did. Prior to addressing the following hypotheses it was necessary to rationally eliminate the data with biased response styles or that were likely coded incorrectly. Two primary decision points were considered, either eliminate all patients that had SOQs with a total score of zero at any time in the archival data set or only eliminate the SOQs and corresponding BPRS-Es at the point when the SOQ total score was zero (e.g., if

participant X has an SOQ total score of zero on their third iteration but had good data on all other iterations, then only their third iteration would be eliminated and the rest would be retained; thus, retaining the maximum number of participants in the study.). A total of 106 participants had, at some point, an SOQ total score of zero. Figure 8 depicts the percentage of remaining data from those participants after eliminating the SOQs with a total score of zero and the corresponding BPRS-E. At face value there may be some concerns because many of the participants had a low percentage of data points included in the data analysis. However, these patients' data were used in point-in-time comparisons and were not included in the change analyses; thus, maximizing the utility of each participant's data. Therefore, it was decided when the total score of the SOQ equals zero, only that episode and the corresponding BPRS-E were eliminated, preserving the rest of the patient's data for analysis.

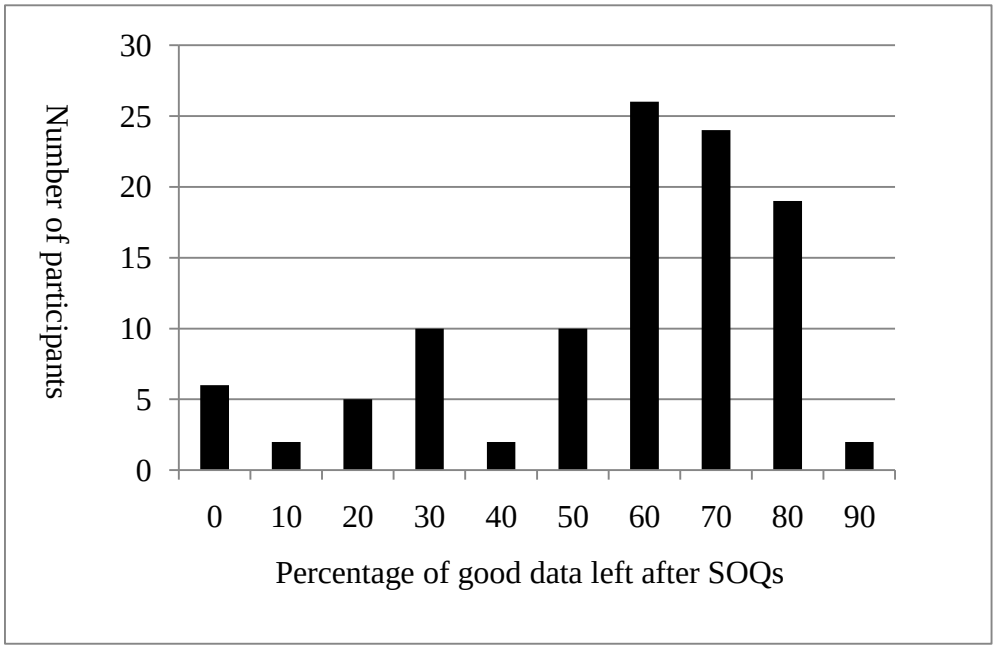


Figure 8. *Histogram of Remaining SOQ Data*

Establishing the cutoff score. Jacobsen and Truax (1991) provided a formula to calculate an empirical cutoff score that best separates two distinct populations. This cutoff score is used to identify whether a score is more likely to fall into a clinical or nonclinical population. The statistical formula is as follows:

$$C = \frac{\left(M_{clinical} \times SD_{non-patient}\right) + \left(M_{non-patient} \times SD_{clinical}\right)}{SD_{clinical} + SD_{non-patient}}$$

where M_{clinical} and SD_{clincial} is the mean and standard deviation of the patient population, respectively, and $M_{\text{non-patient}}$ and $SD_{\text{non-patient}}$ is the mean and standard deviation of the non-patient population, respectively. The non-patient *M* of 36.00 and *SD* of 17.26 were adopted from the SOQ manual. The *M* of 64.73 and *SD* of 29.89 from this data set were used for the clinical variables (The outpatient *M* of 93.06 and *SD* of 26.34 from the SOQ manual were not included as part of the clinical variables after a two-tailed *t*-test indicated they were statistically different from one another.). The final calculation of the empirical cutoff score was as follows:

$$C = \frac{\left(64.73 \times 17.26\right) + \left(36.00 \times 29.89\right)}{29.89 + 17.26} = 46.52$$

resulting in the empirical cutoff score of 47 (i.e., anything below 47 was considered subclinical and anything equal to or above was considered clinical).

Data Analysis

Hypothesis 1. The primary purpose of the first hypothesis was to understand the relationship between the two outcome measures. Thus, the Pearson Product Correlation Coefficient (PPCC) was calculated to test the correlation between the two measures. Because both are outcome measures that were administered across time it was important to explore the correlation of the two measures in multiple ways. The first was to calculate the overall correlation of the BPRS-E and the SOQ. This initial analysis essentially put all the data points,

from admission to discharge, in the correlation pool and allows for a baseline comparison for further evaluations. Next, the correlation was calculated between the two measures for each time period (admission, 90-, 180-, 270-, and 360+ days). Finally, the correlation was calculated for the clinical and subclinical SOQs to their corresponding BPRS-Es at each time period.

Essential to the interpretation of the correlation between the BPRS-E and SOQ is to recognize the assumptions of the PPCC calculations. The first assumption is that of bi-variate normality. This assumption necessitates that each distribution of the variables is approximately normal and is not skewed (Howell, 2002, p. 267). The second assumption is that of linearity. This assumption requires a linear relationship between the two variables. If a nonlinear relationship exists then the PPCC will not capture the true relationship between the two distributions of variables. Finally, the third assumption is that of homogeneity. This assumption requires that there is the same amount of variance throughout the distribution by each distribution of variables (Howell, 2002, p. 267-268).

Hypothesis 2. The primary objective of hypothesis two was to identify the relative strength of predictors from the BPRS-E on clinical and subclinical self-completed measure in the SMI population. Specifically, in this study the low and high SOQ scores were the predicted variables while the four subscales and individual items of the BPRS-E were the predictor variables. When dichotomous predicted variables are used logistical regression is preferred to discriminate analyses and linear regression. It is preferred to the discriminate analysis because discriminate analyses can produce probabilities outside of the zero-to-one range, which is theoretically impossible. Second, discriminate analyses depend on restrictive normality assumptions, which, in the case of special populations such as the SMI, are often not realistic. Although linear regression may provide a good estimate, logistical regression is preferred

because, like discriminate analyses, it is possible to get probability estimates less than zero and greater than one, especially at the extreme ends of estimation. Second, the distribution of variances is small for the extreme ends of the dependent variables and large in the middle. This is a violation of the regression assumptions based on a lack of homogeneity of variance.

The logistical regression can be represented by the following notation:

$$P = (e^{a + bX}) / (1 + e^{a + bX}),$$

where P is the probability of scoring in a particular category, *e* is the base of the natural logarithm (i.e., exponential variable), X is the predictor variable (the predictor value), *a* is the y-intercept and *b* is the amount of increase in log odds. Evident in Figure 9, the binary variables are not normally distributed and a sigmoidal curve best represents the predictability of dichotomous, dependent variables where there is little change at the extremes and a drastic change in between the binary variables.

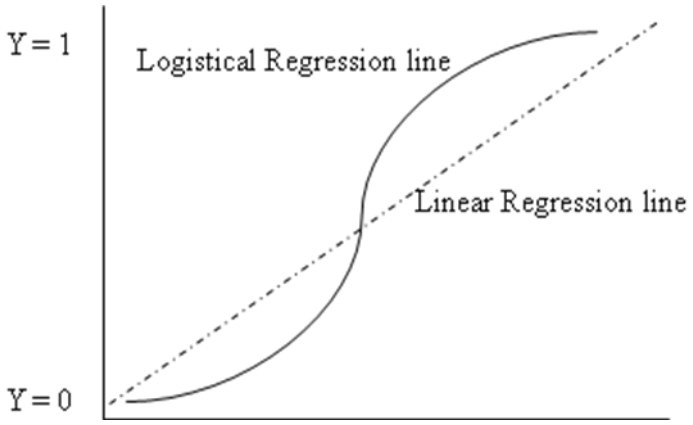


Figure 9. *Example Comparison of Logistical and Linear Regression Lines.*

Hypothesis 3. Hypothesis three was a comparison of the change trajectories of the BPRS-E and the SOQ. The HLM analysis for this hypothesis was computed using the SAS statistical program. This method offers several advantages over traditional end-point

comparisons, because it accounts for intraindividual change when change is nonlinear (refer to pages 50 through 52 for a more detailed explanation of HLM advantages).

Although there are several advantages of hierarchical analyses, there are important considerations to remember. First, modeling intraindividual change, particularly when the change is nonlinear, requires a minimum of three within-subject data points. Second, because HLM analyses are essentially a regression of regressions, the same assumptions of regression analyses also apply (Arnold, 1992). Third, because the there are hierarchical levels in the analysis, it is necessary that the units are nested in groups. Fourth, although it is unclear how much is sufficient within and between subject data (Willett, 1989), there needs to be sufficient data in both conditions to conduct the analyses. Researchers suggest that the reliability of the analysis increases as the number of within-subject and between-subject observations increase (Willett, 1989).

Hypothesis 4. Similar to hypothesis three, HLM analysis was used to model the change patterns of the SOQ scores in the clinical and subclinical range (cutoff score of 47) with their corresponding BPRS-E scores.

Hypothesis 5. This hypothesis was conceptualized to gain further understanding of subclinical SOQ scores. As discussed above (see pages 24 through 26), this is based on the assumption that there are three different trajectories. Burlingame et al., (2008) reported the reliable change index for the SOQ at a 90% confidence interval to be at 14.9; thus, an RCI of > 14 was used to assume reliable change. The empirical cutoff score of 47 on the SOQ was used to separate the clinical and subclinical population. Following the identification of the different groups, each group was modeled using HLM analysis and compared to their corresponding BPRS-E change trajectories.

Hypothesis 6. As a continuation from the previous hypothesis, the goal of hypothesis six is to identify predictors from the BPRS-E (subscales or individual items) to the groups of subclinical SOQ scores indentified as a result of the previous analyses. As with hypothesis two and for the same reasons, a logistical regression was used estimate the probability of membership in one of the groups.

Results

Sample demographics

Of the total 2,180 participants, 357 (16.38%) met all inclusion criteria and were included in the data analysis, producing 1,650 BPRS-E (12.56%) and SOQ (16.15%) ratings. The primary reasons for exclusion were exclusively or a combination of missing data on the BPRS-E (i.e., scores of zero; 54.31%), the BPRS-E did not have a corresponding SOQ (48.49%), and the participant had less than three data points (31.15%). Of the 357 participants, 152 (42.57%) were female and 205 (57.42%) were male. The age ranged from 20 years old to 90 years old, with a *M* of 43.44 and a mode of 30. The most common primary diagnoses were schizoaffective disorder (27.65%), schizophrenia, paranoid type (23.46%), bipolar with psychotic features (6.42%). As shown in Table 12, the number of assessments ranged from three to 16, with a mode of three.

As part of archival data research, it is necessary to work within the parameters of the available data. In this study, the dates of the data ranged from January 1, 2000 through December 31, 2008. Because each measure was administered on 90-day intervals and it is unclear whether the scores within 90 days of each endpoint were indeed admission or discharge scores. As shown in Figure 10, of the 357 participants, 10 (3.57%) had first scores within 90 days of the beginning archival data set (March 30, 2000 and earlier) and 81 (22.67%) had last

Table 12

Distribution of the Number of Assessments per Patient

Number of Assessments	Number of days at USH	N	Percent
3	270	149	41.74
4	360	89	24.93
5	450	39	10.92
6	540	29	8.12
7	630	13	3.64
8	720	8	2.24
9	810	12	3.36
10	900	6	1.68
11	990	4	1.12
12	1080	3	0.84
13	1170	1	0.28
14	1260	1	0.28
15	1350	2	0.56
16	1440	1	0.28

scores within 90 days of the end (October 3, 2008 and later). A two-tailed t-test analysis revealed there was not a significant difference between admission scores within 90 days of January 1, 2000 and the rest of the data (the BPRS-E total score [$F(2, 357) = 0.39$, $p = 53$], Thought Disturbance [$F(2, 357) = 0.03$, $p = .87$], Animation [$F(2, 357) = 0.13$, $p = .98$], Mood disturbance [$F(2, 357) = 2.96$, $p = .09$], Apathy, [$F(2, 357) = 0.00$, $p = .98$], and the SOQ total [$F(2, 357) = 0.12$, $p = .73$]), although there was a significant difference of some scores at discharge (BPRS-E total score [$F(2, 357) = 26.80$, $p < .01$], Mood Disturbance [$F(2, 357) = 21.93$, $p < .01$], and Apathy [$F(2, 357) = 29.03$, $p < .01$]).

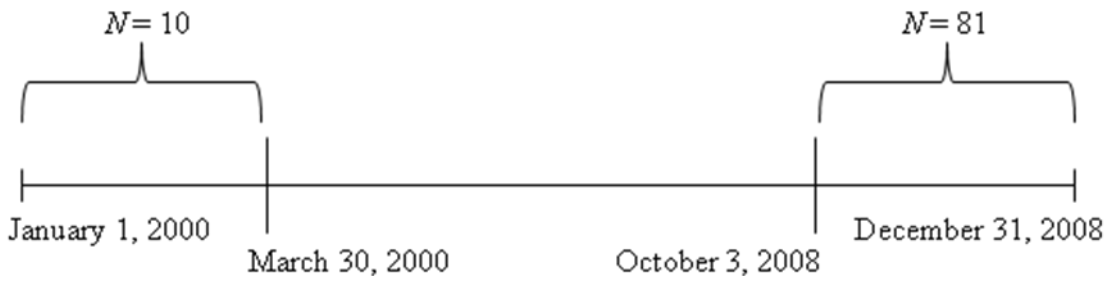


Figure 10. *Number of Participants within 90 days of Data Set End Points*

Descriptive data

Of those included in the analysis, 299 (83.75%) were first admissions, 43 (12.04%) were second admissions, and 15 (4.20%) were third admissions to USH. A two-tailed, t-test analysis with a Tukey-Kramer adjustment for multiple analyses was used to test for significant differences between these three groups. The means of each admission were not significantly different from one another, thus the three admissions were combined for a total of 357 participants [$F(2, 351)$ $p > .9$].

Table 13 displays the range, *M*, and *SD* of each item, subscale, and total score of admission BPRS-E. Nearly every item achieved full range with ample distribution properties, suggesting adequate variability for subsequent analyses. As can be noted, at the item level responses on some items do not have a normal distribution. Furthermore, as shown in Figure 11, the distribution of the admission scores was slightly positively skewed and leptokurtic. However, these values are within reasonable means to meet the assumptions of the statistical procedures.

Table 13

BPRS-E Descriptive Statistics

Variable	Min	Max	*M*	*SD*	Skew	Kurtosis
Total Score	26	95	54.19	12.31	0.86	0.42
Subscales						
Thought Disturbance	4	26	11.9	5.74	1.2	.92
Animation	4	19	6.86	3.29	1.89	3.33
Mood Disturbance	4	26	10.13	5.22	1.52	3.21
Apathy	4	22	8.59	3.65	1.75	3.45
Items						
Somatic Concern	1	7	2.63	1.62	1.08	.37
Anxiety	1	7	3.07	1.67	0.71	-0.03
Depression	1	7	2.97	1.89	1.20	.83
Suicidality	1	7	2.00	1.66	3.63	14.10
Guilt	1	7	2.09	1.49	1.97	3.17
Hostility	1	7	2.49	1.61	1.34	1.09
Elevated Mood	1	7	1.72	1.28	2.82	8.95
Grandiosity	1	7	2.55	2.05	2.14	3.46
Suspiciousness	1	7	3.19	1.90	1.20	0.22
Hallucinations	1	7	2.72	2.08	1.46	0.76
Unusual Though Content	1	7	3.44	2.05	1.20	0.16
Bizarre Behavior	1	7	2.50	1.59	2.18	4.04
Self-neglect	1	7	2.49	1.40	1.60	2.08
Disorientation	1	7	2.10	1.43	2.21	4.61
Conceptual Disorganization	1	6	2.41	1.44	2.06	4.22
Blunted Affect	1	7	2.33	1.31	1.28	1.27
Emotional Withdrawal	1	7	2.11	1.25	2.18	6.76
Motor Retardation	1	6	1.66	1.00	2.45	6.61
Tension	1	6	1.82	1.01	3.06	13.33
Uncooperativeness	1	7	1.61	1.19	4.51	21.14
Excitement	1	7	1.72	1.20	3.71	14.77
Distractibility	1	6	1.73	1.09	6.34	49.09
Motor Hyperactivity	1	6	1.61	1.02	4.95	26.03
Mannerisms and Posturing	1	5	1.25	0.72	11.69	141.90

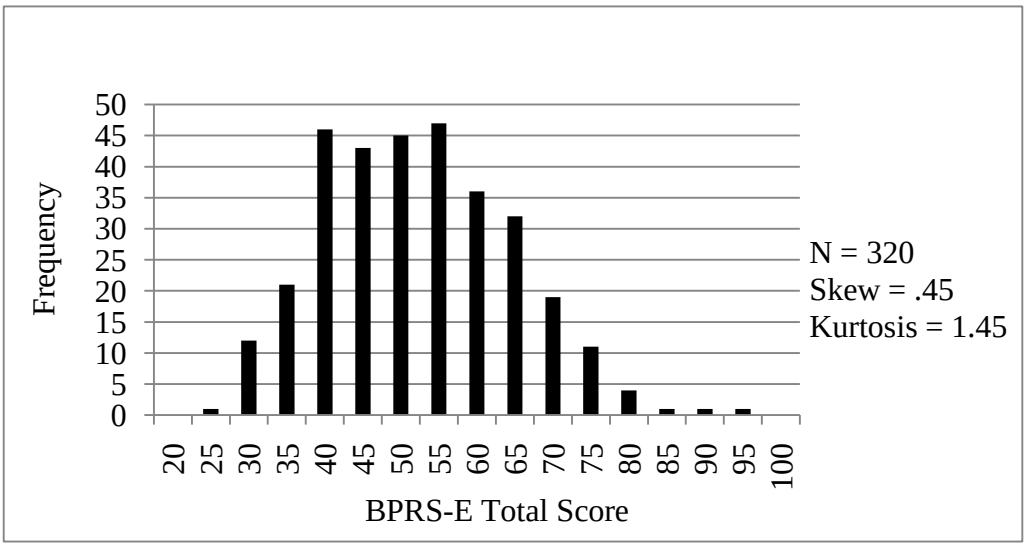


Figure 11. *Distribution of Admission Scores for the BPRS-E*

As shown in Table 14, the range, *M*, and *SD* of each item and total score of the admission scores for the SOQ contained adequate variability. Every individual item was represented at the full range of its scale and the total score reflected 85.56% of its total range, suggesting adequate variability for subsequent analyses. Furthermore, as shown in Figure 12, the distribution of the admission scores was slightly positively skewed and leptokurtic. However, these values are within reasonable means to meet the assumptions of the statistical procedures.

Table 14

SOQ Item Descriptive Statistics

Item	Min	Max	*M*	*SD*	Skew	Kurtosis
Total	3	154	64.73	29.89	0.01	-0.65
1	0	4	1.74	1.33	0.32	-1.12
2	0	4	1.63	1.24	0.41	-0.92
3	0	4	1.84	1.29	0.24	-1.09
4	0	4	1.93	1.37	0.21	-1.23
5	0	4	2.02	1.40	0.14	-1.33
6	0	4	1.71	1.11	0.23	-0.78
7	0	4	0.92	1.21	1.29	0.60
8	0	4	1.60	1.22	0.41	-0.89
9	0	4	1.90	1.35	0.22	-1.21
10	0	4	1.48	1.25	0.54	-0.79
11	0	4	0.38	0.89	2.67	6.54
12	0	4	1.38	1.35	0.69	-0.80
13	0	4	2.06	1.39	0.07	-1.31
14	0	4	1.98	1.34	0.07	-1.24
15	0	4	1.12	1.11	1.00	0.31
16	0	4	1.74	1.29	0.31	-1.00
17	0	4	1.56	1.40	0.53	-1.02
18	0	4	2.08	1.32	0.03	-1.22
19	0	4	1.42	1.29	0.67	-0.64
20	0	4	0.45	0.97	2.41	5.15
21	0	4	1.31	1.18	0.73	-0.28
22	0	4	1.72	1.32	0.36	-1.03
23	0	4	1.18	1.17	0.89	-0.05
24	0	4	0.38	0.92	2.76	7.15
25	0	4	1.45	1.31	0.61	-0.79
26	0	4	1.87	1.26	0.18	-1.04
27	0	4	1.55	1.35	0.49	-0.99
28	0	4	1.51	1.44	0.59	-1.05
29	0	4	1.80	1.34	0.24	-1.14
30	0	4	2.02	1.32	0.06	-1.23
31	0	4	1.15	1.39	0.99	-0.39
32	0	4	1.23	1.27	0.85	-0.37
33	0	4	1.13	1.20	0.90	-0.22
34	0	4	0.67	1.32	1.75	2.04
35	0	4	1.38	1.28	0.64	-0.72
36	0	4	1.20	1.22	0.90	-0.20
37	0	4	1.69	1.45	0.42	-1.18
38	0	4	1.74	1.22	0.20	-1.04
39	0	4	0.74	0.98	1.38	1.24
40	0	4	1.45	1.37	0.63	-0.86
41	0	4	1.60	1.28	0.46	-0.92
42	0	4	1.05	1.26	1.01	-0.19
43	0	4	1.26	1.28	0.79	-0.49
44	0	4	1.33	1.25	0.74	-0.46
45	0	4	1.43	1.40	0.65	-0.91

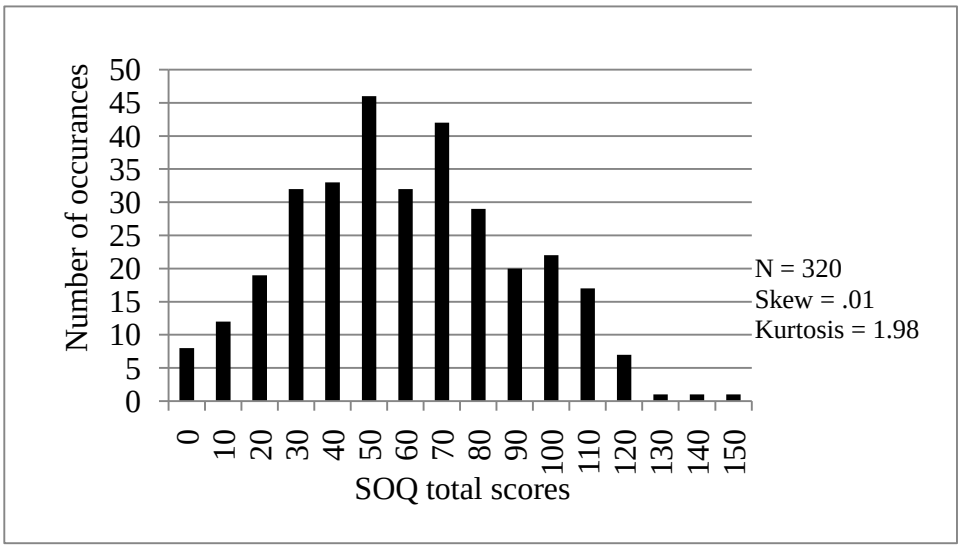


Figure 12. *Distribution of Admission Scores for the SOQ*

Table 15 shows a comparison of the admission and discharge means of both the SOQ and BPRS-E. A one-tailed, t-test analysis indicated the admission scores are significantly higher than the discharge scores. Overall, the BPRS-E and SOQ contain the necessary psychometrics to meet the assumptions of the statistical procedures used to test the hypotheses. Each measure has relatively good variability within appropriate skew and kurtosis metrics.

Table 15

Admission and Discharge Means

	Admission			Discharge			*t*
Measure	*N*	*M*	*SD*	*N*	*M*	*SD*	
SOQ	320	64.73	29.89	320	49.30	23.07	-9.78*
BPRS	320	54.19	12.31	320	39.67	11.56	-18.47*

* $p < .01$

Hypothesis 1

The first hypothesis addressed the correlation between the BPRS-E and SOQ. The initial analysis utilized all the data, irrespective of cutoff scores. Because both measures are outcome measures and developed to track change in symptoms over time, it was necessary to correlate the measures at the different assessment time periods to adequately assess their relationship.

As shown in Table 16, the BPRS-E and SOQ are positively correlated at all points. To further support these findings, additional correlations were calculated at the clinical and subclinical levels on the SOQ with their corresponding BPRS-E. As shown in Table 17, SOQ scores in the clinical range correlated significantly with the BPRS-E scores at all assessment periods. In contrast, the SOQ scores in the subclinical range only correlated with BPRS-E scores at 180- and 270-days.

Table 16

Correlations between the BPRS and SOQ

Iteration	*n*	*r*
All data	1435	.33**
Admission	320	.26**
90 day	326	.29**
180 day	313	.37**
270 day	171	.35**
360+ days	305	.20**

**$p < .001$.

Table 17

Correlations between the BPRS and SOQ above and below a Rational Cutscore

	Clinical		Subclinical	
Iteration	*N*	*r*	*N*	*r*
Initial	227	.19**	93	-.05
90 days	197	.16*	129	.10
180 days	171	.35***	142	.22**
270 days	103	.34***	68	.33**
360+ days	189	.16*	116	.03

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

Hypothesis 2

The purpose of hypothesis two was to identify items or subscales on the BPRS-E that are associated with reality impairment that may predict subclinical SOQ scores. The results of the stepwise logistical regression resulted in the following statistically significant (*LRT* < .01*; Wald* < .01) formula to predict subclinical scores:

$$= 1 - \frac{\text{exponent } (-2.9302 + MD(0.3576) + H(0.2842) + U(-0.4203) + CD(0.3115))}{1 + (\text{numerator})}$$

where MD is the score of the subscale Mood Disturbance, H is the score of the item Hallucinations, U is the score of the item Uncooperativeness, and CD is the score of the item Conceptual Disorganization. Although several different combinations of scores can result in different overall probabilities, four examples are shown in Figure 13. As noted in Table 13, all of the BPRS-E items, with the exception of Uncooperativeness, have properties of a normal

distribution. In these examples, each line represents one item or subscale that is being manipulated while the other variables remain at their lowest value (Mood Disturbance subscale was modeled by multiples of four; thus, the numbers on the horizontal axis are 4, 8, 12, 16, 20,

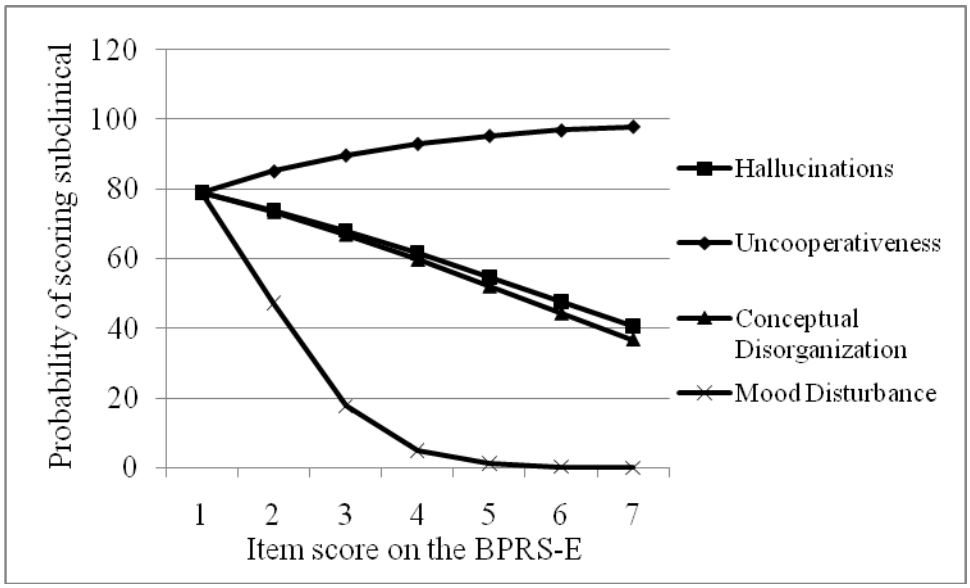


Figure 13. *Examples of Probability Estimates*

24,and 28 for this subscale). For example, as the score for Uncooperativeness increases and the score for Hallucinations and Conceptual Disorganization remain at one and Mood Disturbance remains at four (their relative lowest values), the probability of a patient scoring in the subclinical level on the SOQ increases from 78.99% (when Uncooperativeness is at one) to 97.91% (when Uncooperativeness is at seven), as represented by the line with a diamond marker.

Hypothesis 3

Figure 14 depicts the change trajectories of both the BPRS-E and SOQ. The change trajectory for the BPRS-E was -10.10 and for the SOQ it was -10.08. Although there was a significant difference between the y-intercepts ($p < 0.01$), there was no difference between the

change trajectories (p = 0.98). These results support hypothesis three, in that the two change trajectories are essentially identical.

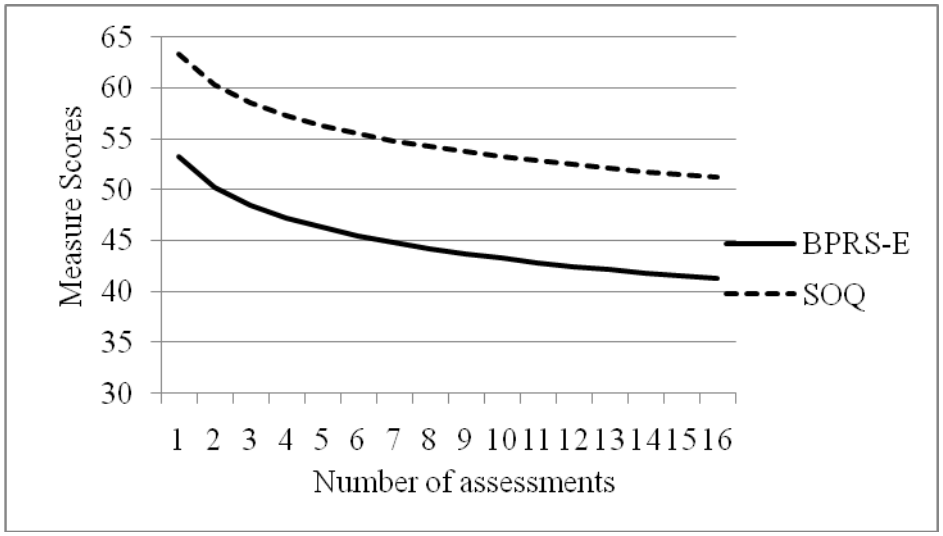
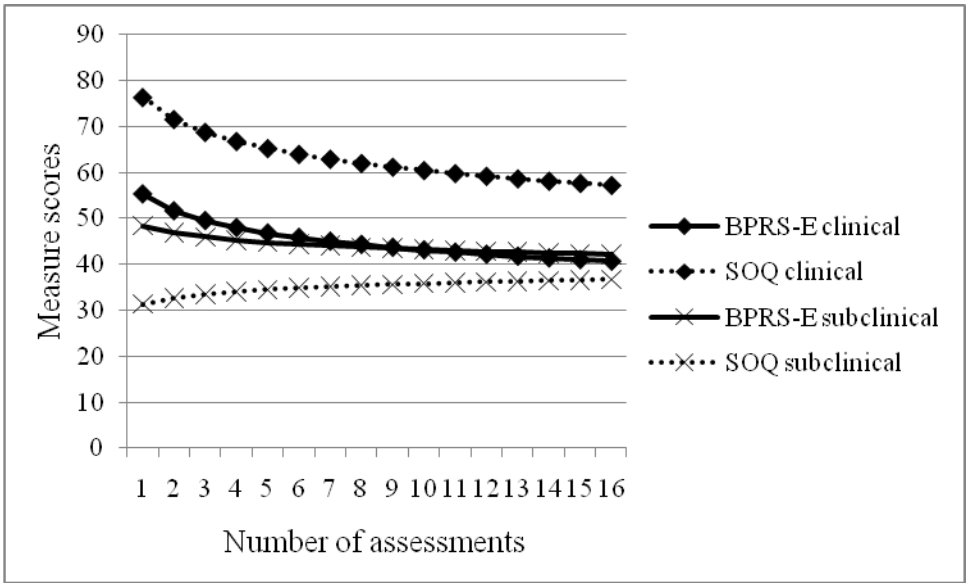


Figure 14. *Change Trajectory Comparison of the BPRS-E and SOQ*

Hypothesis 4

As shown in Figure 15, when the trajectory comparisons are separated by the cutoff score they become significantly different. The change trajectory for the BPRS-E scores and SOQ scores within the clinical range was -12.15 and -15.96, respectively. The change trajectory for the BPRS-E and SOQ in the subclinical range was -5.15 and 4.46, respectively. Furthermore, each change trajectory was significantly different from all other change trajectories ($p < 0.01$).

The results of hypothesis three, at first glance, suggest that the BPRS-E and SOQ have nearly identical change trajectories. However, when considering the self-completed measurement patterns of the patients, the change trajectories are quite different. Indeed, the subclinical SOQ scores actually had a positive change trajectory (the patients get worse), while all other change trajectories were in the theoretically predicted (improved) direction.



Note: ♦ = clinical group; × = subclinical group; •••= SOQ; — = BPRS-E

Figure 15. *Change Trajectories for the BPRS-E and SOQ by Clinical Status*

Hypothesis 5

Hypothesis five was intended to better understand subclinical SOQ scores. As shown in Table 18, three primary groups were created to test this hypothesis: (1) admission scores that had a positive reliable change and exceeded the cutoff score by 90 days, (2) admission scores that had a positive reliable change and exceeded the cutoff score after 90 days, and (3) admission scores that did not exceed the cutoff score. Figure 16 depicts the change trajectories of each group. The change trajectory for group three was -0.92 (p = .51), group two was 11.76 (p < .01), and group one was 10.02 (p < .05). A post hoc, chi-squared (χ^2) analysis revealed that there was not a significant difference between groups one and two [$\chi^2(2, n = 35) = -0.50$, p = .62], but both groups were significantly different from group 3 [$\chi^2(2, n = 75) = 6.99$, p <.01; $\chi^2(2, n = 72) = 4.72$, p <.01]. Thus, groups one and two were combined into a single group for subsequent analyses. Table 19 depicts the frequency of patients when the SOQ score exceeded the cutoff

Table 18

Frequency Groupings of Subclinical SOQ Scores

Categories	Frequency
Exceeded cutoff score after 90 days and had reliable positive change	19
Exceeded cutoff score by 90 days and had reliable positive change	16
Never went above the cutoff score	58

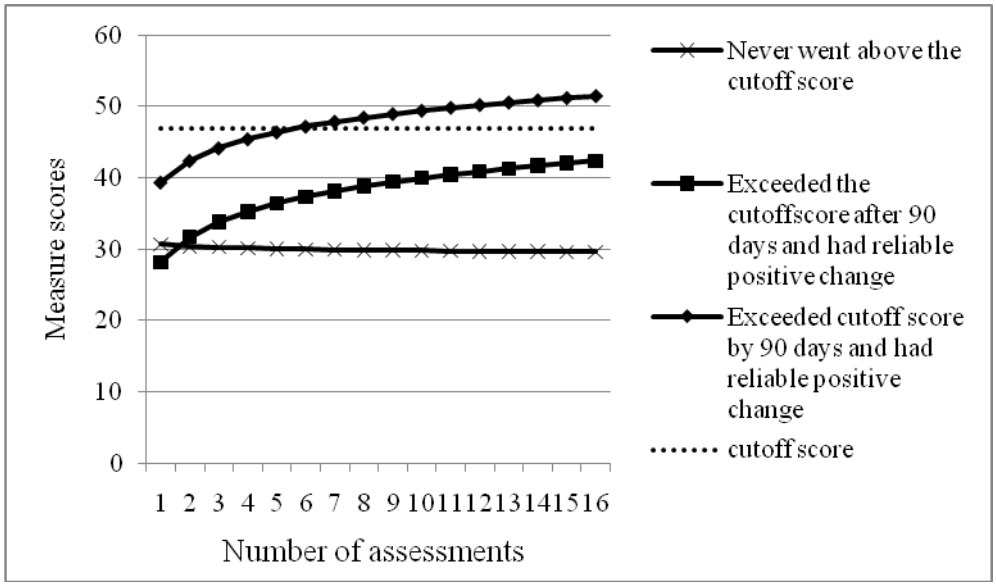


Figure 16. *Change Trajectories of the Three SOQ Subclinical Groups*

score. Figure 17 depicts the change trajectories using the time period when the cutoff score was exceeded [(90 days, 10.02, $p < .05$); (180 days, 9.50, $p < .05$); (270 days, 19.44, $p < .01$); (360+

days, 5.17, p = .20)]. Of note, the change trajectory for the group that meet the criteria at 360 or more days was not significant, which is likely due to the low sample size.

Table 19
Frequency of Subclinical SOQ Admission Scores with Positive Change

	90 days	180 days	270 days	360+ days	Total
Scores in the subclinical range at admission and *eventually* went above the cutoff score and had a reliable positive change	16	10	6	3	35

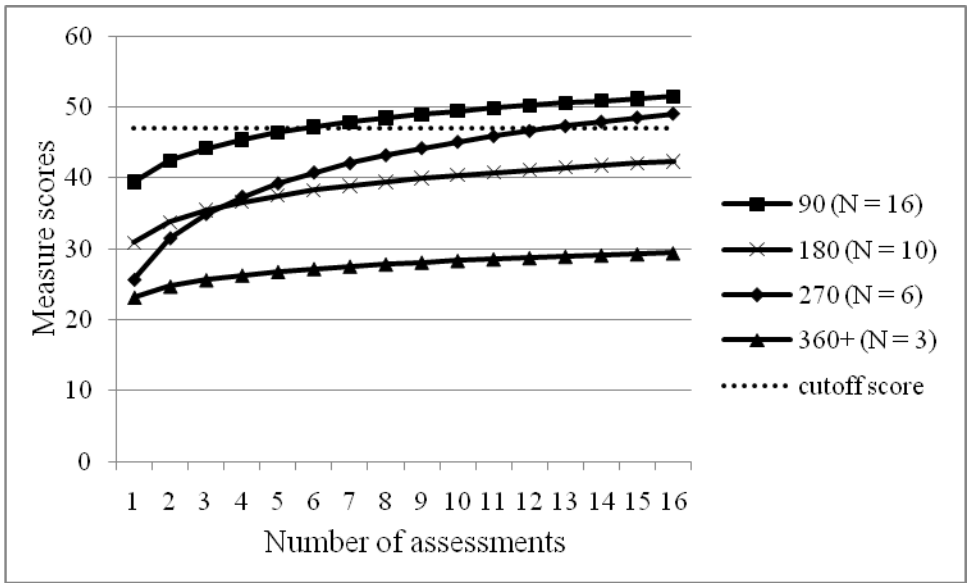


Figure 17. *Change Trajectories by when Subclinical Scores Exceeded the SOQ Cutoff Score*

As shown in Table 20, the clinical group and subclinical group three were examined by their change condition at termination using the RCI. As can be seen, their rates of change are

near mirror images of one another. Furthermore, there were five participants (1.40%) that exceeded the cutoff score but did not have a reliable change at any time during treatment.

Table 20

Frequencies of subcategories of clinical and subclinical SOQ scores

Primary categories	Subcategory Frequencies			Total
	Reliable negative change	*Reliable positive change*	*No reliable change*	
SOQ scores in the clinical range at admission	154 (60%)	33 (12%)	71 (28%)	258
SOQ scores in the subclinical range at admission that *never* went above the cutoff score	14 (24%)	9 (16%)	30 (52%)	53
Total	168 (53%)	42 (13%)	101 (34%)	311

Hypothesis 6

The goal of hypothesis six was to identify predictors from the BPRS-E that would explain different subclinical SOQ change trajectories. The results of the stepwise logistical regression resulted in the following statistically significant (*LRT* < .01*; Wald* < .01) formula to predict subclinical scores:

$$= 1 - \frac{\text{exponent } (2.4704 + S(-0.3283) + SC(-0.3450))}{1 + (\text{numerator})}$$

where S is the score of the Suspiciousness item and SC is the score of the Somatic Concern item. Although there are a number of different score combinations, Figure 4.7 provides three examples of the rate of probability increase relative to the increase in score on a particular item. In these examples, the lines with markers represent one item that is being manipulated while the other variable remains at the lowest value. The line without a marker is a model of the both items

increasing at the same rate. For example, as the score for Suspiciousness increases and the score for Somatic Concerns remains at one, the probability of a patient scoring in the subclinical level on the SOQ increases from 10.51% (when Suspiciousness is at one) to 45.70% (when Suspiciousness is at seven), as represented by the line with the square marker. In contrast, when both scales increase at the same rate the probability of a patient scoring in the subclinical level on the SOQ increases from 14.22% (when Suspiciousness and Somatic Concerns are at one) to 90.40% (when Suspiciousness and Somatic Concerns are at seven), as represented by the line without a marker.

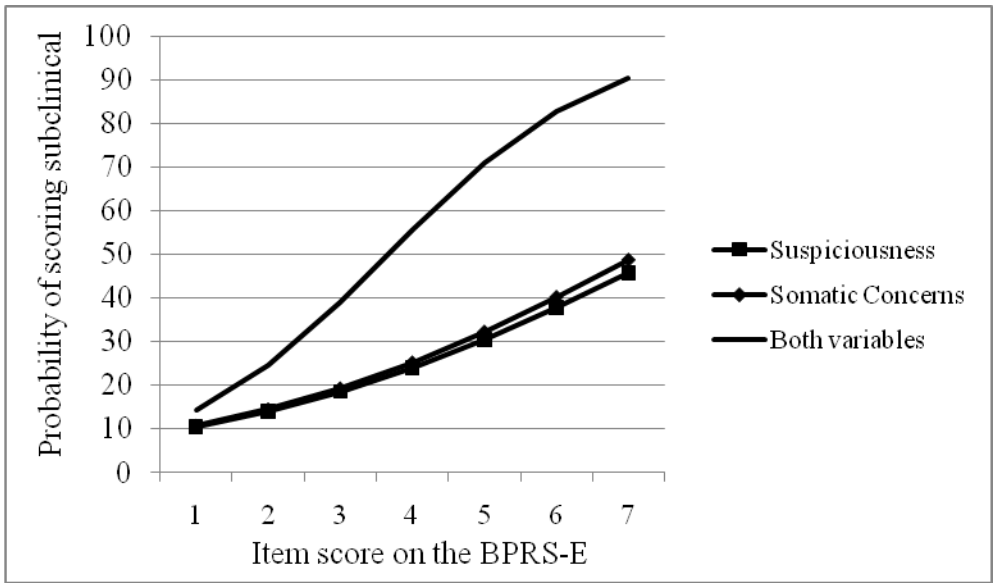


Figure 18. *Examples of Probability Estimates for having a Subclinical SOQ Admission Score*

Discussion

Conclusions

This study sought to examine the relationship between a clinician- and self-completed outcome measures applied to an SMI population. The BPRS-E, a clinician-completed measure, has been recognized as a "gold standard" outcome instrument for the SMI population. Despite such support, the BPRS is not without limitations. Indeed, clinician-completed measures require extensive training and ongoing time for recalibration to insure proper assessment. Unlike clinician-completed measures, self-completed measures, such as the SOQ, are easily implemented and require minimal clinician resources, although they too are not without limitations. In particular, the SMI population often present with cognitive and reality-based impairments which call into question the reliability of self-completed instruments with this population.

This study examined the relationship between the BPRS-E and SOQ at point-in-time assessments as well as longitudinal change trajectories. Two-thirds of the participants' inpatient treatment lasted 360 days or less and just under half of the participants' treatment lasted less than 270 days. Thus, longitudinal change should be calibrated with this time frame in mind. Finally, we explored the subclinical self-completed patterns on the SOQ defined as patients who produced SOQ admission scores that fell below what would be expected from those entering a state psychiatric hospital. We tested BPRS items and subscales as predictors of subclinical SOQ admission scores.

Low to moderate correlations between the BPRS-E and SOQ were found, where approximately 10% of the variance in one measure is explained by the variance in the other. When the SOQ total score fell within the clinical range it was correlated with the BPRS total

score at all assessment points. However, this relationship stability was absent with subclinical SOQ admission scores. Specifically, there was no relationship between the subclinical BPRS-E and SOQ scores until the 180-day assessment.

These findings suggest a reliable low-to-moderate relationship between clinician and self-completed outcome measures representing a medium effect size (Cohen, 1988). The size of the relationship suggests that each measure captures unique features of the symptom distress picture in the SMI population. Thus, the two measures cannot be used in an interchangeable fashion. The results for subclinical SOQ group at 180 days suggests that at this point in treatment, the SOQ may reflect a more accurate reading of patient distress in that the two measures begin to agree. However, as was evidenced by subsequent analyses, subclinical SOQ admission scores may reflect a complex clinical picture.

Membership in the two groups of SOQ admission scores (clinical and subclinical) were reliably predicted by a combination of one BPRS-E subscale (Mood Disturbance) and three BPRS-E items (Hallucinations, Uncooperativeness, and Conceptual Disorganization). Uncooperativeness reliably predicted membership in the *subclinical* group with higher ratings on uncooperativeness associated with membership in the subclinical SOQ group. Mood Disturbance, Hallucinations, and Conceptual Disorganization reliably predicted *clinical* SOQ scores at admission. Patients who scored higher on these items and subscale were more likely to produce SOQ admission scores in the clinical range.

These findings suggest that patients with severe psychological distress can provide self-completed assessments that "match" clinical reality, especially when the patient is sufficiently cooperative. Our original thought was that the presence of hallucination and thought disturbance (as measured by the BPRS) would be related to subclinical SOQs admission scores. However,

the logistic regression produced the opposite pattern. Higher levels of disturbance on these items were related to higher overall SOQ and BPRS-E admission scores. Nonetheless, it is important to interpret these findings in light of previous research (and the present study) that shows the average distress levels for SMI inpatients at intake being substantially lower than comparable SMI outpatients (Burlingame, et al., 2008). Thus, inpatient SMI self-completed distress levels are clearly attenuated, undoubtedly due in part to the severity of the impairment that brings them to the hospital.

Surprisingly, the change trajectories on the SOQ and BPRS-E were virtually identical with a *p*-value of 0.98. The remarkable similarity between the change trajectories suggests that both measures are tracking change on the same construct (distress on severe psychiatric symptoms) adding confidence to the construct validity of the clinician- and self-rated assessment battery. This similarity balances the low-to-moderate correlations found at the point-in-time assessments. It may be that the BPRS-E and SOQ measure different aspects of severe psychiatric symptoms producing the low-to-moderate point-in-time correlations, but that change on these symptoms is sufficiently distinct that both the clinician and patient can reliably detect such. It also suggests that even though SMI inpatients under report distress relative to SMI outpatients, reliable change trajectories can be produced by the majority of SMI inpatients using a self-completed outcome instrument.

A less favorable explanation for the surprisingly high relationship between the SOQ and BPRS-E may be rater bias. It may be that the amount of time a patient spends in the hospital equally affects both self-completed and clinician-completed assessments (e.g., "I've (S/he has) been here for so long I (they) must be doing better"). The USH attempts to attenuate such affects by ongoing rater calibration but we cannot entirely rule out this explanation.

A finding that partially addresses the rater bias explanation lies in the change trajectories for the patients who produced subclinical admission SOQs. Individuals who produced a subclinical SOQ score showed symptom worsening. These same patients showed a reliably shallower change trajectory on the BPRS-E. Admittedly, the change trajectories on the two measures were still in the opposite direction. However, the fact that more treatment was not necessarily associated with more change for a subgroup of patients on both measures attenuates our confidence in rater bias as the primary explanation for the high level of agreement for change on the two measures. There were findings that further qualify this explanation. For instance, the reduced change on the BPRS for the subclinical group may also be due to their lower admission scores. Moreover, attenuation in change for both measures may be related to the fact that uncooperativeness predicted subclinical SOQ admission scores. In support of the uncooperativeness explanation, Burlingame et al. (2006) found that change on the Uncooperativeness BPRS-E item was shallower than for other BPRS-E items (Hallucinations, Conceptual Disorganization) and subscales (Mood Disturbance) using a similar population of SMI inpatients. It's fair to say that the jury is still out and that further study of this population using both clinician- and self-completed instruments is warranted.

The authors found that subclinical SOQ scores at admission can be reliably classified into two primary groups; subclinical SOQ admission scores that eventually surpassed the clinical cutoff score during treatment and ultimately led to reliable change on the SOQ versus subclinical SOQ admission scores that never exceeded the cut score (i.e., remained subclinical throughout the hospital stay) with most patients showing no reliable change. Two BPRS-E items (Suspiciousness and Somatic Concerns) reliably predicted membership in these two groups. Patients who were rated as highly suspicious on the BPRS-E at admission apparently provide

very little information about their symptom distress on the SOQ. This denial of symptoms leads to admission scores that fall in the "normal" range. This finding suggests that those who score high on the uncooperative and suspiciousness may not be good candidates for self-completed assessment of symptoms.

In contrast to highly suspicious patients, those who scored high on the somatic BPRS item may simply lack insight into their symptoms making them poor candidates for self-completed outcome assessment For instance, research on the MMPI-2 suggests that those who lack insight often somaticize their psychological distress (Graham, 2000). Indeed, ratings above 5 on the BPRS-E somatic item essentially describe severe impairment with delusions that corresponds to an absence of insight.

These findings suggest that it is possible to distinguish between SMI patients who eventually provide a reliable self-completed assessment and those who do not. Indeed, in our sample 40% of subclinical admission SOQ scores eventually had a reliable positive change and exceeded the clinical cutoff score. If replicated by future research, these findings may have clinical import. For example, the somatic and suspicious items combined predicted membership in the subclinical group at an accuracy of 90% when both items were rated a seven. Thus, these two items could be used to eliminate self-completed outcome assessment at admission since these individuals ultimately produce useless self-completed assessments. Using a process such as this, however, should be approached with caution and consideration for the patient. For example, subclinical scores on the SOQ may also represent understandable suspiciousness as patients go through the hospital admission process (e.g., involuntary commitment) rather than a more complex psychological dynamic. Indeed, careful interpretation of outcome assessment

should be maintained throughout treatment, particularly with those who suffering from severe mental illness.

Limitations

This study utilized archival data from USH, an inpatient mental health hospital. This naturalistic setting provides several benefits to the current study. Both measures were developed and are intended to be used with the SMI population. Patients at an inpatient mental health hospital, by definition, represent the most severe of the mentally ill. Furthermore, because this study was archival eight years of data was easily incorporated into this study, which is often difficult in prospective studies. However, this study was not immune to limitations that are often associated with archival research designs and outcome assessment.

The absence of a control group is a significant limitation. Although much can be said about the relationship between the two measures and the changes, without a control group interpretations must remain tentative.

The USH administers the BPRS-E on 90-day intervals based on length of stay and resource availability. Ideally more frequent assessments (i.e., on a bi-weekly basis) might reveal more information on the change patterns of the inpatient SMI population. Indeed, nearly one-third of those in the original data were excluded because the participant had less than three assessments and over half were excluded because of incomplete BPRS-E ratings. There are undoubtedly a host of clinical (e.g. early discharge) and non-clinical (e.g., staff oversight) reasons for this incomplete data. Nonetheless, our confidence in their ecological validity is bolstered by their agreement with previous change trajectory findings (Burlingame et al., 2006).

As a note of caution, the inpatient SMI population is probably the most vulnerable population of society. This study suggests there are limits in our ability to fully capture the

patient's functioning when only using self-completed outcome measures. The overlap of clinician-completed and self-completed is only 10%, suggesting there is much to be learned with symptom distress assessment and change in this population. Thus, caution is recommended when decision makers use such data to make treatment decisions.

Given the limitations of this study, it behooves future researchers to address these challenges. A prospective study would offer controls (e.g., assessment intervals, training, and data interpretation) that is absent in archival studies. Future research could also address the predictive properties of the BPRS-E. For example, if treatment were to focus on patients who score high in the uncooperativeness item, might this produce more reliable and self-completed SOQs? As noted by Earnshaw et al. (2005), clinician "buy-in" to the applicability and utility of outcome assessment also influences the patient's compliance. The findings of this study, accompanied by future research, may help clinician's understanding of self-completed outcome assessment with the SMI population, providing guidelines on when and how to use them and more importantly interpretation thereof.

References

Abel, K. M., O'Keane, V., Murray, R. M., & Cleare, A. J., (1997). Serotonegeric function and negative and depressive symptomatology in schizophrenia and major depression. *Psychoneuroendocrinology, 22,* 539-548.

Aiken, L. R. (1977). Note on sensitivity: A neglected psychometric concept. *Perceptual and Motor Skills, 45,* 1330.

Aleman, A., Hijman, R., de Haan, E. H. F., & Kahn, R. S. (1999). Memory impairment in schizophrenia: A meta-analysis. *American Journal of Psychiatry, 156*(9), 1358-1366.

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory.* Monterey, CA: Brooks/Cole Publishing Company.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Arnold, C. L. (1992). Methods, plainly speaking: An introduction to hierarchical linear models. *Measurement and Evaluation in Counseling and Development, 25,* 58-91.

Averill, P.M., Hopko, D.R., Small, D.R., Greenlee, H.B., & Varner, R.V. (2001). The role of psychometric data in predicting inpatient mental health service utilization. *Psychiatric Quarterly, 72,* 215-235.

Bachrach, H., Mintz, J., & Luborsky, L. (1971). On rating empathy and other psychotherapy variables: An experience with the effects of training. *Journal of Consulting and Clinical Psychology, 36,* 445.

Bech, P., Larsen, J. K., & Andersen, J. (1988). The BPRS: Psychometric developments. *Psychopharmacology Bulletin, 24,* 118-121.

Begtsson-Tops, A., & Hansson, L. (1999). Subjective quality of life in schizophrenic patients living in the community: Relationship to clinical and social characteristics. *European Psychiatry, 14*, 256-263.

Beller, S. A., & Overall, J. E. (1984). The Brief Psychiatric Rating Scale (BPRS) in geropsychiatric research: II. Representative profile patterns. *Journal of Gerontology, 39*, 194-200.

Blanchard, J. J, Bellack, A. S., & Mueser, K. T. (1994). Affective and social-behavioral correlates of physical and social anhedonia in schizophrenia. *Journal of Abnormal Psychology, 103*, 719-728.

Bordin, E. S., Cutler, R. I., Dittmann, A. T., Harway, N. I., Rausch, H. L. & Rigler, D. (1954). Measurement problems in process research on psychotherapy. *Journal of Consulting Psychology, 18,* 79-82.

Borison, R. L., Sinha, D., Haverstock, S., McLarnon, M. C., & Diamond, B. I. (1989). Efficacy and safety of tiosprine vs. haloperidol and thioridazine in a double-blind, placebo-controlled trial. *Psychopharmacology Bulletin, 25*, 190-193.

Bow-Thomas, C. C., Velligan, D. I., Miller, A. L., & Olsen, J. (1999). Predicting quality of life from symptomatology in schizophrenia at exacerbation and stabilization. *Psychiatry Research, 86*, 131-142.

Breier, A., Schreiber, J. L., Dyer, J., & Pickar, D. (1991). National Institute of Mental Health longitudinal study of chronic schizophrenia: Prognosis and predictors of outcome. *Archives of General Psychiatry, 48*, 239-246.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of Hierarchical Linear Models to assessing change. *Psychological Bulletin, 101*, 147-158.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and data analysis methods*. Newbury Park: SAGE publications.

Bryk, A. S., & Weisberg, H. I. (1977). Use of the nonequivalent control group design when subjects are growing. *Psychological Bulletin, 84,* 950-962.

Burlingame, G. M., Dunn, T. W., Chen, S., Lehmann, A., Axman, R., Earnshaw, D., Rees, F. (2005) Selection of outcome assessment instruments for inpatients with severe and persistent mental illness. *Psychiatric Services, 56(4),* 444-451.

Burlingame, G. M., Lambert, M. J., Reisinger, C. W., Neff, W. L., & Mosier, J. I. (1995). Pragmatics of tracking mental health outcomes in a managed care setting. *Journal of Mental Health Administration, 22,* 226-236.

Burlingame, G. & Lee, J. L. (2004). The Brief Psychiatric Rating Scale. In Strauss & J. Schumacher (Eds.), Klinische Interviews and ratings (Clinical Interviews and Ratings). Goettingen: Hogrefe.

Burlingame, G. M., Mosier, J. I., Wells, M. G., Atkin, Q. G., Lambert, M. J., Whoolery, M., & Latkowski, M. (2001). Tracking the influence of mental health treatment: The development of the Youth Outcome Questionnaire. *Clinical Psychology and Psychotherapy, 8,* 361-379.

Burlingame, G. M., Nelson, P., Lee, J. L., Thayer, S., & Lambert M. J. (2008). *Administration and scoring manual Severe-Outcome Questionnaire.* American Publishing Credentialing Services L.L.C.

Burlingame, G. M., Seaman, S., & Johnson, J. (2006). Sensitivity to change of the Brief Psychiatric Rating Scale (BPRS): An item and subscale analysis. *Psychological Services, 3(2),* 77-87.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. *Psychological bulletin, 56*(2), 81-105.

Caracena, P. F., & Vicory, J. R. (1969). Correlates of phenomenological and judged empathy. *Journal of Counseling Psychology, 16,* 510-515.

Carey, J. (2000). The Severe Outcome Questionnaire: A preliminary study of reliability and validity. *Dissertation Abstracts International Section B. 61,* 5554.

Carey, M. P., & Carey, K. B. (1999). Behavioral research on severe and persistent mental illnesses. *Behavioral Therapy, 30,* 345-353.

Centers for Medicaid and Medicare Services (2009). *National Health Expenditures 2008 Highlights.* Retrieved from http://www.cms.hhs.gov/NationalHealthExpendData/downloads/highlights.pdf and *National Health Expenditures Projections for 2009-2019.* Retrieved from http://www.cms.hhs.gov/NationalHealthExpendData/downloads/proj2009.pdf

Ciarlo, J. A., Brown, T. R., Edwards, D. W., Kiresuk, T. J., & Newman, F. L. (1986). *Assessing Mental Health Treatment Outcome Measurement Techniques.* Series FN, No. 9, DHHS Publication No. (ADM) 86-1301. Washington, DC: US Government Printing Office.

Coffey, R. M., Mark, T., King, E., Harwood, H., McKusick, D., Genuardi, J., Dilonardo, J., & Buck, J. A. (1997). *National Estimates for Expenditures for Mental Health the Substance Abuse Treatment.* SAMHSA Publication No. SMA-00-3499. Rockville, MD: Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration, July 2000.

Cohen, J. (1988). *Statistical power analysis in the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, L. M., & Cliff, N. (1990). Use of longitudinal Guttman simplex as a basis for measuring growth. *Psychological Bullietin, 108*, 128-134.

Corrigan, P. W., & Addis, I. B. (1995). The effects of cognitive complexity on a social sequencing tasks in schizophrenia. *Schizophrenia Research, 16*, 137-144.

Corrigan, P. W., Hirschbeck, J. N., & Wolfe, M. (1995). Memory and vigilance training to improve social perceptions in schizophrenia. *Schizophrenia Research, 17*, 257-265.

Craig, T. J., Richardson, M. A., Pass, R., & Bregmann, Z. (1985). Measurement of mood and affect in schizophrenic inpatients. *American Journal of Psychiatry, 142*, 1272-1277.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed). New York: Harper & Row.

Dawes, R. M. (1996). *House of cards: Psychology and psychotherapy built on myth.* New York, NY: Free Press.

den Boer, J. A., Ravelli, D. P., Huisman, J., Ohrivk, J., Verhoeven, W. M. A. M, & Westenberg, H. G. M. (1990). A double-blind comparitive study of remoxipride and haloperidol in acute schizophrenia. *Acta Psychiatrica Scandinavica, 82*(Suppl. 358), 108-110.

Deyo, R. A., Diehr, P., Patrick, D. L. (1991). Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. *Controlled Clinical Trials, 12*, 142S-158S.

Dibben, C. R. M., Rice, C., Laws, K., & McKenna, P. J. (2009). Is executive impairment associated with schizophrenic syndromes? A meta-analysis. *Psychological Medicine 39*, 381-392.

Du Bois, P. H. (1970). *A history of psychological testing*. Boston: Allyn & Bacon.

Durlak, J. A., Wells, A. M., Cotton, K. J., & Johnson, S. (1995). Analysis of selected methodological issues in child psychotherapy research. *Journal of Clinical Child Psychology, 24,* 141-148.

Earnshaw, D., Rees, F., Dunn, T., Burlingame, G., & Chen, S (2005). Implementing a multi-source outcome assessment battery in a state psychiatric hospital: A case from the public sector. *Psychiatric Services, 56,* 411-413.

Faustman, W. O. (1994). Brief Psychiatric Rating Scale. In Maruish, M.E. (Ed.) *The use of psychological testing for treatment planning and outcome assessment* (pp. 371 401). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Feigher, J. P., Merideth, C. H., & Claghorn, J. L. (1984). Multicenter placebo-controlled evaluation of nomifensine treatment in depressed outpatients. *Journal of Clinical Psychiatry, 45*, 47-51.

Finn, J. D. (1974). *A general model for multivariate analysis.* New York: Holt, Rinehart, & Winston.

Fiske, D. W. (1977). Methodological issues in research on the psychotherapist. In A. S. Gurman & A. M. Razin (Eds.), *Effective psychotherapy* (pp. 23-43). New York: Pergamon Press.

Flemenbaum, A. & Zimmerman, R.L. (1973). Inter- and Intra-rater reliability of the Brief Psychiatric Rating Scale. *Psychological Reports, 36,* 783-792.

Froyd, J. E., Lambert, M. J., & Froyd, J D. (1996). A survey and critique of psychotherapy outcome measurement. *Journal of Mental Health, 5,* 11-15.

Gabbard, G. O., Coyne, L., Kennedy, L. L. Beasley, C., Deering, C. D., Schroder, P., Larson, J., & Cerney, M. S. (1987). Interrater reliability in the use of the Brief Psychiatric Rating Scale. *Bulletin of the Menninger Clinic, 51*, 519-531.

Galderisi, S., Davidson, M., Kahn, R. S., Mucci, A., Boter, H., Gheorghe, M. D., … EUREST group (2009). Correlates of cognitive impairment in first episode schizophrenia: The EUFEST study. *Schizophrenia Research, 115*(2-3), 104-114.

Garfield, S. L. (1990). Issues and methods of psychotherapy process research. *Journal of Consulting and Clinical Psychology, 58*, 273-280.

Gaultieri, C. T., & Powell, S. F. (1978). Psychoactive drug interactions. *Journal of Clinical Psychiatry, 39*(9), 720-729.

Gold, J. M., & Harvey, P. D. (1993). Cognitive deficits in schizophrenia. *Psychiatric Clinics of North America, 16*(2), 295-312.

Goldman, R. S., Tandon, R., Liberzon, I., & Greden, J. F. (1992). Measurement of depression and negative symptoms in schizophrenia. *Psychopathology, 25*, 49-56.

Graham, J. R. (2000). *MMPI-2 Assessing Personality and Psychopathology, 3rd edition*. New York: Oxford University Press.

Guyatt, G. (1988). Measuring health status in chronic airflow limitation. *European Respiratory Journal, 1*, 560-564.

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

Hannan, C., Lambert, M., Harmon, C., Nielson, S., Smart, D., Shimokawa, K., & Sutton, S. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology, 61*(2), 155-163.

Hansson, L., Middelboe, T., Merinder, L., Bjarnson, O., Bengtsson-Tops, A., Nilsson, L., …, Vinding, H. (1999). Predictors of subjective quality of life in schizophrenic patients

living in the community. A Nordic multicenter study. *International Journal of Social Psychiatry, 45,* 247-258.

Hatfield D. R., & Ogles, B. M. (2004). The use of outcome measures by psychologists in clinical practice. *Professional Psychology: Research and Practice, 35*(5), 485-491.

Hatfield D. R., & Ogles, B. M. (2006). The influence of outcome measures in assessing client change and treatment decisions. *Journal of Clinical Psychology, 62*(3), 325-337.

Hatfield D. R., & Ogles, B. M. (2007). Why some clinicians use outcome measures and others do not. *Administration and Policy in Mental Health and Mental Health Services Research, 34*(3), 283-291.

Hedlund, J. L., & Vieweg, B. W. (1980). The Brief Psychiatric Rating Scale (BPRS): A comprehensive review. *Journal of Operational Psychiatry, 11,* 49-65.

Heppner, P. P., Kivlighan, D. M., Jr., & Wampold, B. E. (1999). *Research design in counseling* (2nd ed.). New York: Brooks/Cole.

Hill, C. E., & Lambert, M. J. (2004). Methodological issues in studying psychotherapy processes and outcomes. In M. J. Lambert (ed.) *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (pp. 84-135). New York: John Wiley & Sons, Inc.

Hill, C. E., Nutt, E., & Jackson, S. (1994). Trends in psychotherapy process research: Samples, measures, researchers, and classic publications. *Journal of Counseling Psychology, 41,* 364-377.

Hobbs, C., Tennant, C., Rosen, A., Newton, L., Lapsley, H. M., Tribe, K., & Brown, J. E. (2000). Deinstitutionalisation for long-term mental illness: A 2-year clinical evaluation. *Australian and New Zealand Journal of Psychiatry, 34,* 476-483.

Hogan, T. P., Awad, A. G., & Eastwood, R. (1983). A self-report scale predictive of drug compliance in schizophrenia: Reliability and discriminative validity. *Psychological Medicine, 13,* 177-183.

Hopko, D. R., Lachar, D., Bailley, S. E., & Varner, R. V. (2001). Assessing predictive factors for extended hospitalization at acute psychiatric admission. *Psychiatric Services, 52,* 1367-1373.

Howell, D. C. (2002). *Statistical Methods for Psychology, 5th edition*. Duxbury: United States.

Huttenlocher , J. E., Haight, W., Bryk, A. S., & Selter, M. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27*(2), 236-249.

Jacobson, N. S., Follette, W. C., Ravenstorf, D., Baucom, D. H., Hahlweg, K., & Margolin, G. (1984). Variability in outcome and clinical significance of behavioral marital therapy: A reanalysis of outcome data. *Journal of Consulting and Clinical Psychology, 52,* 497-504.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12-19.

Jovanovski, D., Zakzanis, K. K., & Atia, M. (2007). A comparison between a researcher-rated and self-report method of insight assessment in chronic schizophrenia revisited: A replication study using the SUMD and SAIQ. *Journal of nervous and mental disease, 195*(2), 165-169.

Kapur, S., & Kambhampati, R. K. (1992). Drug interactions in psychopharmacology. In M. S. Keshavan, & J. S. Kennedy (Eds.) *Drug-induced dysfunction in psychiatry* (pp. 21-37). Washington, DC, US: Hemisphere Publishing Corp.

Katz, R., Shaw, B. F., & Vallis, T. M. (1995). The assessment of severity and symptom patterns in depression. In E. E. Beckham, & W. R. Leber (Eds.) *Handbook of depression (2nd ed.)* (pp. 61-85). New York, NY: Guilford Press.

Kazdin, A. E. (1992). *Research Design in Clinical Psychology*. Needham Heights, Massachusetts: Simon & Schuster.

Kim, K., Kim, J., Kim, J., Park, D., Jang, H. J., Ku, J., … Kim, S. I. (2007). Characteristics of social perception assessed in schizophrenia using virtual reality. *Cyberpsychology & Behavior 10*(2), 215-219.

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Disease, 38*(1), 27-36.

Kiesler, D. J. (1973). *The process of psychotherapy*. Chicago: Aldine.

Klein, M. H., & Cleary, T. A. (1967). Platonic true scores and error in psychiatric rating scales. *Psychological Bulletin, 68*, 77-80.

Kontaxakis, V. P., Havaki-Kontaxaki, B. J., Stamouli, S. S., Margariti, M. M., Collias, C. T., & Christodoulou, G. N. (2000). Comparison of four scales measuring depression in schizophrenic inpatients. *European Psychiatry, 15*, 274-277.

Lachar, D., Bailley, S. E., Rhoades, H. M., Espadas, A., Aponte, M., Cowan, K. A., Gummatira, P., Kopecky, C. R., & Wassef, A. (2001). New subscales for an anchored version of the Brief Psychiatric Rating Scale: Construction, reliability, and validity in acute psychiatric admissions. *Psychological Assessment, 13(3)*, 384-395.

Lachar, D., Espadas, A., & Bailley, S. E., (2004). The Brief Psychiatric Rating Scale: Contemporary applications. In M.E. Maruish, (Ed.) *The use of Psychological Testing for*

Treatment Planning and Outcomes Assessment: Volume 3: Instruments for Adults (3rd ed.) (pp. 153-190). New Jersey: Lawrence Erlbaum Associates.

Lambert, M. J. (1983). Introduction to assessment of psychotherapy outcome: Historical perspective and current issues. In M. J. Lambert, E. R. Christensen, and S. S. DeJulio (Eds.), *The Assessment of Psychotherapy Outcome*. New York: John Wiley and Sons.

Lambert, M. J. (2001). Psychotherapy outcome and quality improvement: Introduction to the special section on patient-focused research. *Journal of Consulting and Clinical Psychology, 69,* 147-149.

Lambert M. J., Bergin, A. E., & Garfield, S. L. (2004). Introduction and Historical Overview. In M. J. Lambert (ed.) *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (pp. 3-15). New York: John Wiley & Sons, Inc.

Lambert, M. J., Finch, A. M., Okiishi, J., Burlingame, G. M. (2005). *Administration and scoring manual for the OQ-30.2.* OQ Measures.

Lambert, M. J., Hatfield, D. R., Vermeersch, D. A., Burlingame, G. M., Reisinger, C. W., & Brown, G. S. (2001). *Administration and Scoring Manual for the Life Status Questionnaire.* Stevenson, Maryland: American Professional Credentialing Services LLC.

Lambert, M. J., & Hill, C. E. (1994). Assessing psychotherapy outcomes and processes. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (pp. 72-113). New York: Wiley.

Lambert, M. J., Whipple, J. L., Smart, D. W., Vermeersch, D. A., Nielsen, S. L., & Hawkins, E. J. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research, 11*(1), 49-68.

Lee, J. L., Rees, F., Burlingame, G. M., Hwang, T. C., & O'Neal, B. (2005). *Clinically Significant change in the Brief Psychiatric Rating Scale: Establishing normative, outpatient and inpatient cut scores and a Reliable Change Index.* Paper presentation at Society of Psychotherapy Research, Montreal, Canada.

Leipzig, R. M., & Mendelowitz, A. (1992). Adverse psychotropic drug-drug interactions. In J. M. Kane, & J. A. Lieberman (Eds.) *Adverse effects of psychotropic drugs* (pp.13-76). New York, NY: Guilford Press.

Lipsey, M. W. (1990). *Design sensitivity.* Newbury Park, CA: Sage.

Lorr, M., Jenkins, R. L., & Holsopple, J. Q. (1953). *Multidimensional scale for rating psychiatric patients.* V.A. Technical Bulletin No. 10-507, Veterans Administration.

Lorr, M., McNair, D. M., Klett, C. J., & Lasky, J. J. (1960). A confirmation of nine postulated psychotic syndromes. *American Psychologist, 15*, 495.

Lukoff, D., Neuchterlein, K.H., & Ventura, J. (1986). Manual for the Expanded BPRS. *Schizophrenia Bulletin*, Vol 12, p. 594-602.

Lyons, J. S., Howard, K. I., O'Mahoney, M. T., & Lish, J. D. (1997). *The Measurement and Management of Clinical Outcomes in Mental Health.* New York: John Wiley & Sons.

Mahurin, R. K., Vellingan, D. I., & Miller, A. L. (1998). Executive-frontal lobe cognitive dysfunction in schizophrenia: A symptom subtype analysis. *Psychiatry Research, 79*, 139-149.

Mark, T., McKusick, D., King, E., Harwood, H., & Genuardi, J. (1998). *National expenditures for mental health, alcohol, and other drug abuse treatment, 1996*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Mason, W. M., Wong, G. M., & Entwistle, B. (1983). Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.), *Sociological methodology* (pp.72-103). San Francisco: Jossey-Bass.

Mathalon, D. H., Ford, J. M., & Pfefferbaum, A. (2000). Trait and state aspects of P300 amplitude reduction in schizophrenia: A retrospective longitudinal study. *Biological Psychiatry, 47*, 434-449.

McAdams, L. A., Harris, M. J., Bailey, A., Fell, R., & Jeste, D. V., (1996). Validating specific psychopathology scales in older outpatients with schizophrenia. *Journal of Nervous and Mental Disease, 184*, 246-251.

McCay, E. A., & Seeman, M. V. (1998). A scale to measure the impact of a schizophrenic illness on an individual's self-concept. *Archives of Psychiatric Nursing, 12*, 41-49.

Meltzer, H. Y., Burnett, S., Bastani, B., & Ramire, L. F., (1990). Effects of six months of clozapine treatment of the quality of life of chronic schizophrenic patients. *Hospital and Community Psychiatry, 41*, 892-897.

Merrin, E. L., & Floyd, T. C. (1997). Clinical symptoms of schizophrenia affect reference-independent measures of task-induced EEG alpha asymmetry. *Psychiatry Research: Neuroimaging, 74*, 47-62.

Meehl, P. E. (1996). *Clinical versus statistical prediction: A theoretical analysis and review of the evidence (New preface).* Lanham, MD: Rowan and Littlefield/Jason Aronson.

Mercer, R. C., & Loesch, L. C. (1979). Audio tape ratings: Comments and guidelines. *Psychotherapy: Theory, Research and Practice, 16*, 79-85.

Mirin, S., & Namerow, M. (1991). Why study treatment outcome? *Hospital and Community Psychiatry, 42*, 1007-1013.

Mitchell, S. K. (1979). Inter-observer agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin, 86,* 376-390.

Monti, P. M., Wallander, J. L., Ahern, D. K., Abrams, D. B., & Monroe, S. M. (1983). Multi-modal measurement of anxiety and social skills in a behavioral role-play test: Generalizability and discriminate validity. *Behavioral Assessment, 6,* 15-25.

Nair, N. P. V., Suranyi-Cadotte, B., Schwartz, G., Thavundayil, J. X., Achim, A., Lizondo, E., & Nayak, R. (1986). A clinical trial comparing intramuscular haloperidol deaconate and oral haloperidol in chronic schizophrenic patients: Efficacy, safety, and dosage equivalence. *Journal of Clinical Psychopharmacology, 6*(Suppl.), 30S-37S.

National Institute of Mental Health (1987). *Towards a Model for a Comprehensive Community-Based Mental Health System.* Washington, DC: NIMH.

Nelson, E.A., Maruish, M.E., & Axler, J.L. (2000). Effects of discharge planning and compliance with outpatient appointments on readmission rates. *Psychiatric Services, 51(7),* 885-889.

Newcomer, J. W., Faustman, W. O., Yeh, W., & Csernansky, J. G. (1990). Distinguishing depression and negative symptoms in unmedicated patients with schizophrenia. *Psychiatry Research, 31,* 243-250.

Nicholson, R., & Feinstein, A. (1996). Predictors of early psychiatric readmission. *Psychiatric Services, 47,* 199.

Nieznanski, M. (2005). Reality monitoring failure in schizophrenia: Relation to clinical symptoms and impairment of self-concept. In J. E. Pletson (ed.) *Progress in schizophrenia research* (pp. 45-76). Hauppauge, NY: Nova Science Publishers.

Nilsson, E., Wenemark, M., Bendtsen, P., & Kristenson, M. (2007). Respondent satisfaction regarding SF-36 and EQ-5D, and patients' perspectives concerning health outcome assessment with routine health care. *Quality of Life Research, 16,* 1647-1654.

Nunally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

Olfson, M., Mechanic, D., Hansell, S., Boyer, C. A., & Walkup, J. (1999). Prediction of homelessness within three months of discharge among inpatients with schizophrenia. *Psychiatric Services, 50,* 667-673.

Ogles, B. M., Lambert, M. J., Weight, D. G., & Payne, I. R. (1990). Agoraphobia outcome measurement: A review and meta-analysis. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2,* 317-325.

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

Overall, J.E., & Klett, C.J. (1972). *Applied multivariate analysis.* New York: McGraw-Hill.

Ownby, R. L., Koss, E., Smyth, K. A., & Whitehouse, P. J. (1994). The factor structure of the Brief Psychiatric Rating Scale in Alzheimer's disease. *Journal of Geriatric Psychiatry and Neurology, 7,* 245-250.

Packer, S., Husted, J., Cohen, S., & Tomlinson, G. (1997). Psychopathology and quality of life in schizophrenia. *Journal of Psychiatry & Neuroscience, 22*(4), 231-234.

Papassotiropoulos, A., Hawellek, B., Frahnert, C., Rao, G. S, & Rao, M. L. (1999). The risk of acute suicidality in psychiatric inpatients increases with low plasma cholesterol. *Pharmacopsychiatry, 32,* 1-4.

Pearlson, G. D., Kim, W. S., Kubos, K. L., Moberg, P. J., Jayaram, G., Bascom, M. J., … Tune, L. E. (1989). Ventricle-rain ratio, computed tomographic density, and brain area in 50 schizophrenics. *Archives of general psychiatry, 46*(8), 690-697.

Phelps, R., Eisman, E. J., & Kohout, J. (1998). Psychological practice and managed care: Results of the CAPP practitioner survey. *Professional Psychology: Research and Practice, 29,* 31-36.

Poole, J. H., Ober, B. A., Shenaut, G. K., & Vinogradov, S. (1999). Independent frontal-system deficits in schizophrenia: Cognitive, clinical, and adaptive implications. *Psychiatry Research, 85,* 161-176.

Priebe, S., Huxley, P., Knight, S., & Evans, S. (1999). Application and results of the Manchester Short Assessment of Quality of Life (MANSA). *International Journal of Social Psychiatry, 45*, 7-12.

Pull, C. B., & Overall, J. E. (1977). Adequacy of the Brief Psychiatric rating Scale for distinguishing lesser forms of psychopathology. *Psychological Reports, 40,* 167-173.

Ragland, J. D., Censits, D. M., Gur, R. C., Glahn, D. C., Gallacher, F., & Gur, R. E. (1996). Assessing declarative memory in schizophrenia using Wisconsin Card Sorting Test stimuli: The Paired Associate Recognition Test. *Psychiatry Research, 60,* 135-145.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical Linear Models: Applications and data analysis methods: 2nd Edition. Thousand Oaks, CA: Sage Publications, Inc.

Raudenbush, S. W., & Chan, W. (1993). Application of hierarchical linear model to the study of adolescent deviance in an overlapping cohort design. *Journal of Consulting and Clinical Psychology, 61,* 941-951.

Reid, J. B. (1970). Reliability assessment of observation data: A possible methodological problem. *Child Development, 41,* 1143-1150.

Reisinger, C. W., & Burlingame, G. M. (1997). How to evaluate a mental health outcome measure. In K. M. McCoughlin (Ed.), *The 1997 Behavioral outcomes & guidelines sourcebook* (pp. 370-374). New York: Faulker & Gray, Inc.

Rencher, A. C. (2002). *Methods of Multivariate Analysis, 2nd edition*. Wiley-Interscience: Hoboken, New Jersey.

Rhoades, H. M., & Overall, J. E. (1988). The semi-structured BPRS interview and rating guide. *Psychopharmacology Bulletin, 24,* 101-104.

Roitman, S. E. L., Keefe, R. S. E., Harvey, P. D., Siever, L. J., & Mohs, R. C. (1997). Attentional and eye tracking deficits correlate with negative symptoms in schizophrenia. *Schizophrenia Research, 26,* 139-146.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92,* 726-748.

Romanczyk, R. G., Kent, R. N., Diament, C. & O'Leary, K. D. (1973). Measuring the reliability of observational data: A reactive process. *Journal of Applied Behavior Analysis, 6,* 175-184.

Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Meredith.

Rothbard, A. B., Schinnar, A. P., & Goldman, H. (1996). The pursuit of a definition for severe and persistent mental illness. In S. M. Soreff (Ed.), *Handbook for the treatment of the seriously mentally ill* (pp. 9-26). Kirkland, WA: Hogrefe & Huber Publishers.

Rudolf, H., & Priebe, S. (1999). Subjective quality of life in female in-patients with depression: A longitudinal study. *International Journal of Social Psychiatry, 45,* 238-246.

Ruggeri, M., Leese, M., Thornicroft, G., Bisoffi, G., & Tansella, M. (2000). Definition and prevalence of severe and persistent mental illness. *The British Journal of Psychiatry. 177*, 149-155.

Schlegel, S., Frommberger, U., & Buller, R. (1989). Computerized tomography (CT) in affective disorders: Relationship with psychopathology. *Psychiatry Research, 29*, 271-272.

Shagass, C., & Roemer, R. A. (1992). Evoked potential topography in major depression: II. Comparisons between subgroups. *International Journal of Psychophysiology, 13*, 255-261.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Silverstein, M. L., Harrow, M., & Bryson, G. J. (1994). Neuropsychological prognosis and clinical recovery. *Psychiatry Research, 52*, 265-272.

Slade, M., Powell, R., & Strathdee, G. (1997). Current approaches to identifying the severely mentally ill. *Social Psychiatry and Psychiatric Epidemiology, 32*(4), 177-184.

Speer, D. C., & Greenbaum, P. E. (1994). Five methods for computing significant individual client change and improvement rates: Support for an Individual Growth Curve approach. *Journal of Consulting and Clinical Psychology, 63,* 1044-1048.

Swett, C. (1995). Symptom severity and number of previous psychiatric admissions as predictors of readmission. *Psychiatric Services, 46,* 482-485.

Tandon, R., Mann, N. A., Eisner, W. H., & Coppard, N. (1990). Effect of anticholinergic medication on positive and negative symptoms in medication-free schizophrenic patients. *Psychiatry Research, 31*, 235-241.

Taplin, P. S., & Reid, J. B. (1973). Effects of instructional set and experimenter influence on observer reliability. *Child Development, 44,* 547-554.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York: Science Press.

Tryon, W. W. (1991). *Activity measurement in psychology and medicine.* New York: Plenum.

U.S. Census Bureau (2006). Statistical abstract of the United States: Section 3 health and nutrition. Retrieved November 9, 2007 from http://www.census.gov/prod/2005pubs/06statab/health.pdf.

Van Lieshout, R. J., & Goldberg, J. O. (2007). Quantifying self-reports of auditory verbal hallucinations in persons with psychosis. *Canadian Journal of Behavioural Science Revue Canadienne des Sciences du Comportement, 39,* 73-77.

Velligan , D. I., Mahurin, R. K., Diamond, P. L., Hazelton, B. C., Eckert, S. L., & Miller, A. L. (1997). The functional significance of symptomatology and cognitive function in schizophrenia. *Schizophrenia Research, 25,* 21-31.

Ventura, J., Green, M. F., Shaner, A., & Liberman, R. P. (1993). Training and quality assurance with the Brief Psychiatric Rating Scale: 'The drift busters.' *International Journal of Methods in Psychiatric Research, 3,* 221-224.

Ventura, J., Lukoff, D., Nuechterlein, K., Liberman, R. P., Green, M. F., & Shaner, A. (1993). *Brief Psychiatric Rating Scale (BPRS) Expanded Version (version 4.0).* UCLA Department of Psychiatry and Biobehavioral Sciences: West Los Angeles VA Medical Center.

Ventura, J., Nuechterlein, K. H., Subotnik, K. L., Gutkind, D., & Gilbert, E. A. (2000). Symptom dimensions in recent-onset schizophrenia and mania: A principal components analysis of the 24-item Brief Psychiatric Rating Scale. *Psychiatry Research, 97(2-3),* 129-135.

Vermeersch, D. A. (1998). *Sensitivity to Change as an item selection criteria for an outcome measure* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9835469)

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome Questionnaire: Item sensitivity to change. *Journal of Personality Assessment, 74,* 242-261.

Vermillion, J. M., & Pfeiffer, S. I. (1993). Treatment outcome and continuous quality improvement: Two aspects of program evaluation. *Psychiatric Hospital, 24(1-2),* 9-14.

Voruganti, L., Heslegrave, R., & Awad, A. G. (1998). Quality of life measurement in schizophrenia: Reconciling the quest for subjectivity with the question of reliability. *Psychological Medicine, 28,* 165-172.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *American Statistician, 39,* 95-101.

Wells, M. G., Burlingame, G. M., Lambert, M. J., Hoag, M. J., & Hope, C. A. (1996). Conceptualization and measurement of patient change during psychotherapy: Development of the Outcome Questionnaire and Youth Outcome Questionnaire. *Psychtherapy, 33,* 275-283.

Wilkinson, G., Hesdon, B., & Wild, D., (2000). Self-report quality of life measure for people with schizophrenia: The SQLS. *British journal of psychiatry, 177,* 42-46.

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement, 49*, 249-258.

Woerner, M.G., Mannuzza, S., & Kane, J.M. (1988). Anchoring the BPRS: An aid to improved reliability. *Psychopharmacology Bulletin, 24*, 112-117.

Young, A. H., Blackwood, D. H. R., Roxborough, H., McQueen, J. K., Martin, M. J., & Kean, D. (1991). A magnetic resonance imaging study of schizophrenia: Brain structure and clinical symptoms. *British Journal of Psychiatry, 158*, 158-164.

Zakzanis, K. K. (1998). Neuropsychological correlates of positive s. negative schizophrenia. *Psychiatry Research, 66*, 111-120.

ZumBrunnen, T. L., & Jann, M. W. (1998). Drug interactions with antipsychotic agents: Incidence and therapeutic implications. *CNS Drugs, 9*(5), 381-401.

Appendix A: Brief Psychiatric Rating Scale – Expanded Version

Version 4.0
BRIEF PSYCHIATRIC RATING SCALE
(BPRS)

Expanded Version

Scales, Anchor Points, and Administration Manual adapted by

Joseph Ventura, PhD., David Lukoff, Ph.D., Keith H. Nuechterlein, Ph.D, Robert P. Liberman, M.D., Michael F.Green, Ph.D., and Andrew Shaner, M.D.

Clinical Research Center for Schizophrenia and Psychiatric Rehabilitation

UCLA Department of Psychiatry and Biobehavioral Sciences

West Los Angeles VA Medical Center

February 19, 1993

Please use the following references for citation of the Expanded BPRS:

Current Version (4.0):

Ventura, J., Lukoff, D., Nuechterlein, K.H., Green, M.F., Shaner, A. (1993) Manual for the Expanded Brief Psychiatric Rating Scale. International Journal of Methods in Psychiatric Research 3: 227-243.

Initial Version:

Lukoff, D., Nuechterlein, K.H., Ventura, J. (1986) Manual for the Expanded Brief Psychiatric Rating Scale. Schizophrenia Bulletin 12: 594-602.

For Training and Quality Assurance program:

Ventura, J., Green, M.F., Shaner, A., Liberman, R.P. (1993) Training and Quality Assurance with the Brief Psychiatric Rating Scale: "The Drift Busters". International Journal of Method in Psychiatric Research 3: 221-226.

For Symptom Monitoring:

Lukoff, D., Liberman, R.P., and Nuechterlein, K.H. (1986) Symptom monitoring in the rehabilitation of schizophrenic patients. Schizophrenia Bulletin 12: 578-602.

DESCRIPTION AND ADMINISTRATION OF THE BPRS

The Brief Psychiatric Rating Scale (BPRS) provides a highly efficient, rapid evaluation procedure for assessing symptom change in psychiatric patients. It yields a comprehensive description of major symptom characteristics. Factor analyses of the original 18-item BPRS typically yields four or five solutions. The Clinical Research Center's Diagnosis and Psychopathology Unit has developed a 24-item version of the BPRS.

This manual contains interview questions, symptom definitions, specific anchor points for rating symptoms, and a "how to" section for problems that arise in rating psychopathology. The purpose of the manual is to assist clinicians and researchers to sensitively elicit psychiatric symptoms and to reliably rate the severity of symptoms. The expanded BPRS includes six new scales added to the original BPRS (Overall Gorham, 1962) for the purpose of a more comprehensive assessment of a wider range of individuals with serious mental disorders, especially outpatients living in the community (Lukoff, Nuechterlein, and Ventura, 1986).

This manual will enable the clinician or researcher to conduct a high quality interview adequate to the task of eliciting and rating the severity of symptoms in individuals who are often inarticulate or who deny their illness. The following guidelines are provided to standardize assessment. Please familiarize yourself with these methods for assessing psychopathology.

(1) Using all sources of information on symptoms.

(2) Selecting an appropriate period or interval for rating symptoms.

(3) Integrating frequency and severity in symptom rating: the hierarchical criterion.

(4) Rating the severity of past delusions for which the patient lacks insight.

(5) Rating Symptoms when the patient denies them.

(6) Using a standardized reference group in making ratings.

(7) Rating symptoms that overlap two or more categories or scales on the BPRS.

(8) Rating a symptom that has no specified anchor point congruent with its severity level.

(9) "Blending" ratings made in different evaluation situations.

(10) Resolving apparently contradictory symptoms.

1. USING ALL SOURCES OF INFORMATION ON SYMPTOMS

The rating of psychopathology should be made on the basis of all available sources of information about the patient. These sources include behavioral observations and interviews made by treatment staff, family members, or other caregivers in contact with the patient, available medical and psychiatric case records, and the present interview of the patient. The interviewer/rater is encouraged to seek additional sources of information about the patient's psychopathology from others to supplement the present interview—this is particularly important when the patient denies symptoms.

2. SELECTING AN APPROPRIATE PERIOD OR INTERVAL FOR RATING SYMPTOMS

The duration of the time frame for assessment depends upon the purpose for the rating. For example, if the rater is interested in determining the degree of change in psychopathology during a one month period between pharmacotherapy visits, the rating period should be one month. If a research protocol aims to evaluate the emergence of prodromal symptoms or exacerbation of psychotic symptoms, it may be advisable to select a one week interval since longer periods may lose accuracy in retrospective recall. When a study demands completeness in identifying criteria for relapse or exacerbation during a one or two year period, frequent BPRS assessments will be necessary.

Rating periods typically range from one day to one month. Retrospective reporting by patients beyond one month may suffer from response bias, retrospective distortions, and memory problems (which are common in persons with psychotic and affective disorders). When resources and personnel do not permit frequent assessments, important information can still be captured if the frequency of assessments can be temporarily increased when (1) prodromal symptoms or stress are reported; (2) medication titration and dosing questions are paramount; a (3) before and after major changes in treatment programs.

3. INTEGRATING FREQUENCY AND SEVERITY IN SYMPTOM RATING: THE HIERARCHICAL CRITERION

Most of the BPRS scales are scored in terms of the frequency and/or severity of the symptom. It is sometimes the case that the frequency and severity do not match. A hierarchical principle should be followed that requires the rater to select the highest scale level that applies to either frequency or severity. Thus, when the anchor point definitions contain an "OR," the patient should be assigned the highest rating that applies. For example, if a patient has hallucinations persistently throughout the day (a rating of "7"), but the hallucinations only interfere with the patient's functioning to a limited extent (a rating of "5"), the rater should score this scale "7."

The BPRS is suited to making frequent assessments of psychopathology covering short periods of time. If, however, an interviewer intends to cover a relatively long period of time (e.g., 6 weeks), then combining ratings for severity and frequency of symptoms must be carefully thought out depending upon the specific project goals. If the goal of a project is to define periods of relapse or exacerbation, the rating should reflect the period of peak symptomatology. For example, if over a six week period the patient experienced a week of persistent hallucinations, but was free of hallucinations the remaining time, the patient should be rated a "6" on hallucinations, reflecting the "worst" period of symptomalogy. Alternatively, if the goal is to obtain a general level of symptomatology, the rating should reflect a "blended" or average score. For extended rating periods

(e.g., 3 months), the interviewer may prefer to make one rating reflecting the worst period of severity/frequency/functioning and another rating reflecting the "average" amount of psychopathology for the entire period.

4. RATING THE SEVERITY OF PAST DELUSIONS FOR WHICH THE SUBJECT LACKS INSIGHT

Patients may often indicate varying degrees of insight or conviction regarding past symptoms, making their symptoms difficult to rate. Experiences that result from psychotic episodes can often appear quite real to patients. For example, the belief that others tried to poison them, or controlled all their thoughts and forced them to walk into traffic, could have created severe anxiety and intense fear. Patients can give vivid accounts of their psychotic experiences that are as real as if the situations actually occurred. It is important in these cases to rate the extent to which these memories of a delusional experience can be separated from current delusions involving the present.

Please note that a patient may be able to describe his or her past or current delusions as part of an illness or even refer to them as "delusions." However, a patient should always be rated as having delusions if he or she has acted on the delusional belief during the rating period.

When a patient describes a delusional belief once firmly held, but that is now seen as irrational, then a "1" should be scored for Unusual Thought Content (and also for Grandiosity, Somatic Concern, Guilt, or Suspiciousness if the idea fell into one of these thematic categories). However, if the individual still believes that the past psychotic experience or event was real, despite not currently harboring the concern, it should be rated a "2" or higher depending on the degree of reality distortion associated with the b

Consider the following scenarios:

Scenario No. 1 The patient gives an account of delusional and/or hallucinatory experience and realizes in retrospect that he was ill. He indicates that he has a chemical imbalance in his brain, or that he has a mental condition.

Rate "1" on Unusual Thought Content.

Scenario No. 2 The patient gives indications that his past psychotic experiences were due to a chemical imbalance and/or an illness, but entertains some degree of doubt. He claims it is possible that people were trying to kill him, but he is doubtful. The memories of what happened are not bizarre and he indicates that currently he is certain no one is trying to hurt him.

Rate "2" or "3" on Unusual Thought Content depending on degree of reality retained.

Scenario No. 3 The patient describes previous psychotic experiences as if they actually occurred. He can give examples of what occurred, e.g., co-workers put drugs in his coffee, or that machines read his thoughts. However, the patient says those circumstances no longer occur. The patient is not currently concerned about co-workers or machines, but he is convinced that the circumstances on which the delusions are based actually occurred in the past.

Rate "3" or "4" on Unusual Thought Content depending on the degree of reality distortion, and a "1" on Suspiciousness.

Scenario No, 4 The patient holds bizarre beliefs regarding the circumstances that occurred in the past and/or his current behavior is influenced by delusional beliefs. For example, the patient believes that thoughts were at one time beamed into his mind from aliens OR the patient will not watch T.V. for fear that the messages will again be directed to him OR that the mafia is located in shopping malls that he should avoid.
Rate "4" or higher on Unusual Thought Content depending on the degree of preoccupation and impairment associated with the belief. Consider rating suspiciousness.

Scenario No. 5 The patient believes that previous psychotic experiences were real and previous delusional beliefs are currently influencing most aspects of daily life causing preoccupation and impairment.

Rate "6" or "7" on Unusual Thought Content depending on the degree of preoccupation and impairment associated with the belief.

5. RATING SYMPTOMS WHEN THE PATIENT DENIES THEM

An all too common phenomenon in clinical practice or research is the denial or minimization of symptoms by patients. Patients deny, hide, dissemble or minimize their symptoms for a variety of reasons, including fear of being committed, restricted to a hospital, or having medication increased. Simply recording a patient's negative response to BPRS symptom items, if denial or distortion is present, will result in invalid and unreliable data. When an interviewer suspects that a patient may be denying symptoms, it is absolutely essential that other sources of information be solicited and utilized in the ratings.

Several situations might suggest that a patient is not entirely forthcoming in reporting his/her symptom experiences. Patients may deny hearing voices, yet be observed whispering under their breath as if in response to a voice. The phrasing that a patient uses in response to a direct question about a delusion or hallucination can alert the interviewer to the potential denial of symptoms. For example, if a patient responds to an inquiry regarding the presence of persecutory ideas by saying, "Not really," this is not the same as saying "No." Subtleties in patient responses communicate a great deal and must be followed-up before the interviewer concludes that the symptom is absent.

There are several ways for the interviewer to obtain more reliable information from a patient who may be denying or minimizing symptoms. In all these approaches, interviewing skills, interpersonal rapport, and sensitivity to the patient are of paramount importance. If the patient is experiencing difficulty disclosing information about psychotic symptoms, the interviewer can shift to inquire about less threatening material such as anxiety/depression or neutral topics. The interviewer should then return to sensitive topics after the patient feels more comfortable and concerns about disclosure have been addressed.

The use of empathy is critical in helping a patient express difficult and possibly embarrassing experiences. A interviewer may say, "I understand that recalling what happened may be unpleasant, but I am very interested in exactly what you experienced." It is advisable to let patients know what you may be sensing clinically; "I have the impression that you are reluctant to tell me more about what happened. Could that be because you are concerned about what I might think or write down about you?" The interviewer should actively engage the patient in discussing any apparent reasons

for denying symptoms. The interviewer can discuss openly in an inviting and non-critical fashion any discrepancies noted between the patients' self-report of symptoms and observations of speech and behavior. For example, "You have said that you are not depressed, yet you seem very sad and you have been moving very slowly." When denial occurs, the BPRS interview becomes a dynamic interplay between the interviewer desire for accurate symptom information and determining the reasons underlying the patient's reluctance to disclose.

Occasionally, at the time of the interview, the interviewer will have information about the symptoms that the patient is denying. It is permissible to use a mild confrontation technique in an attempt to encourage a patient to disclose accurate symptom information. For example, a BPRS interviewer may learn from the patient's therapist or relatives of the presence of auditory hallucinations. The interviewer may state, "I understand from talking with your therapist (or relative) that you have been hearing voices. Could you tell me about that?" Letting the patient know in a sensitive and gentle manner that information about his symptoms are already known may aid willingness to disclose. This approach is most effective when a policy of sharing patient information in a treatment team situation is explained to all entering patients. In may be necessary to inform the patient that not all clinical material is shared, but that symptom information needed to manage treatment can not in all cases be confidential.

When you cannot resolve conflicts or contradictions between patients' self-report and the report of others, you must use your clinical judgment regarding the most reliable informants. Be sure to make notes on the BPRS rating sheet regarding any conflicting sources of information and specify how the final decision was made.

6. USING A STANDARDIZED REFERENCE GROUP IN MAKING RATINGS

The proper reference group for conducting assessments is a group of normal individuals who are j psychiatric patients who are living and working in the community free of symptoms. BPRS interviewers should have in mind a group of individuals who are able to function either at work/school, socially, or as a homemaker, at levels appropriate to thc patient's age and socioeconomic status. Research has shown that normal controls score at "2" or below on most psychotic items of the BPRS. BPRS interviewers should not use other patients previously interviewed, especially those with severe symptoms, as the reference standard, since this will systematically bias ratings toward lower scores.

7. RATING SYMPTOMS THAT OVERLAP TWO OR MORE CATEGORIES OR SCALES ON THE BPRS

Systematized or multiple delusions can be rated on more than one symptom item or scale on the BPRS, depending on the theme of the delusional belief. For example, if a patient has a delusion that certain body parts have been surgically removed against his/her will and replaced with broken mechanical parts, he or she would be rated at the level of "6" or '7" on both Somatic Concern and at the level of "4" to "7" on Unusual Thought Content depending on the frequency and preoccupation with the delusion. Furthermore, if the patient felt guilty because he believed the metal in his body interfered with radio transmissions between air traffic controllers and pilots resulting in several plane crashes, the BPRS item Guilt should also be rated.

The specific ratings for each of the overlapping symptom dimensions may differ depending on the anchor points of the BPRS item(s). Thus, a patient with a clear-cut persecutory delusion involving the neighbors should be rated a "6" on Suspiciousness. Whereas, the same delusion could be rated a "4" on Unusual Thought Content if it is encapsulated and not associated with impairment.

8. RATING A SYMPTOM THAT HAS NO SPECIFIC ANCHOR POINT CONGRUENT WITH ITS SEVERITY LEVEL

The anchor points for a given BPRS item are critical in achieving good reliability across raters and across research settings. However, there are occasions when a particular symptom may not fit any of the anchor point definitions. Anchor point definitions could not be written to cover all possible symptoms exhibited by patients. In general, ratings of 2 or 3 represent non-pathological but observable mild symptomatology: 4 or 5 represents clinically significant moderate symptomatology; and 6 or 7 represents clinically significant and severe symptomatology.

The anchor points in this manual are guidelines to aid in the process of defining the character, frequency, and impairment associated with various types of psychiatric symptoms. When faced with a complicated rating, the interviewer may find it useful to first classify the symptom as mild (2 or 3), moderate (4 or 5), or severe (6 or 7), and second to consult the anchor point definitions to pinpoint the rating.

BPRS symptoms that are classified in the severe range usually represent pathological phenomena. However, it is possible for a patient to report or be observed to exhibit examples of mild psychopathology that should be rated at much higher levels. For example, on the item Tension, if hand wringing is observed on 2-3 occasions, the interviewer would rate a "2" or "3." However, if the patient is observed to be hand wringing constantly, then consider a higher rating such as "5" or "6" on Tension. Similarly, instances of severe psychopathology that are brief, transient, and non-impairing in nature should be rated in the mild range.

9. "BLENDING" RATINGS MADE IN DIFFERENT EVALUATION SITUATIONS

A psychiatric patient can exhibit different levels of the same symptom depending on the setting in which the patient is observed or the time period involved. Consider the patient who is talkative during a rating session with the BPRS interviewer, but is very withdrawn and blunted with other patients. In the interview session the patient may rate a "3" on blunted affect and "2" on emotional withdrawal, but rate "5" on those symptoms when interacting with other patients. The interviewer can consider integrating the two sources of information and make an averaged or "blended" rating.

10. RESOLVING APPARENTLY CONTRADICTORY SYMPTOMS

It is possible to rate two or more symptoms on the BPRS that represent seemingly contradictory dimensions of phenomenology. For example, a patient can exhibit blunted affect and elevated mood in the same interview period. A patient may laugh and joke with the interviewer, but then shift to a blunted, slowed, and emotionally withdrawn state during the same interview. In this case, rating the presence of both elevated mood and negative symptoms may be appropriate reflecting that both mood states were present. Although the simultaneous presence of apparently

contradictory symptoms is rare, if such combinations do appear, the rater should consider rating each symptom lower than if just one had appeared. This conservative approach to rating reflects a cautious orientation to the rating process when there is ambiguity regarding the symptomatology being assessed.

CLINICAL APPLICATIONS OF THE BPRS: GRAPHING SYMPTOMS

A graph is printed at the end of this administration manual to help raters plot and monitor symptoms from the BPRS. Because psychotic and other symptoms often fluctuate over time, graphing them enables the clinician to identify exacerbations, periods of remission, and prodromal periods that precede a relapse. Monitoring and graphing can be the key to early intervention to reduce morbidity, relapses, and re-hospitalizations.

Graphing of symptomatology can provide vivid representations of the relationships between specific types of symptoms (e.g., hallucinations) and other variables of interest, such as (1) medication type and dose, (2) changes in psychosocial treatment and rehabilitation programs, (3) the use of "street" drugs or alcohol, (4) life events, and (5) other environmental or familial stressors. The preprinted graph shown at the end of this manual provides space to write significant life events or treatment changes and permits the "eyeballing" of the influence of these variables on symptoms. Repeated measurement and graphing of symptoms over time can be done for individual items (e.g., anxiety or hallucinations), or for clusters of symptoms (e.g., psychotic index). Such clusters can be chosen from factor analyses of earlier versions of the BPRS (Guy, 1976; Overall, Hollister, and Pichot, 1967; Overall and Porterfield, 1963). The blank graph in this manual allows raters to select and write in specific symptoms of the BPRS based on the needs of individual patients.

REFERENCES

Guy W: ECDEU *Assessment Manual for Psychopharmacology*. DHEW Pub. No. (ADM) 76-335. Rockville, MD: National Institute of Mental Health, 1976.

Lukoff D, Nuechterlein KH, and Ventura J: Manual for the Expanded Brief Psychiatric Rating Scale. *Schizophrenia Bulletin*, 12: 594-602, 1986.

Overall JE and Gorham DR, The Brief Psychiatric Rating Scale. *Psychological Reports*, 10: 799-812, 1962.

Overall JE, Hollister LE, Pichot P: Major psychiatric disorders: A four-dimensional model. *Archives of General Psychiatry*, 16:146-151, 1967.

Overall JE and Porterfield, JL. Powered vector method of factor analysis. *Psychometrika*, 28:415-422, 1963.

SCALE ITEMS AND ANCHOR POINTS

Rate items 1-14 on the basis of the patient's self-report. Note items 7, 12, and 13 are also rated on the basis of observed behavior. Items 15-24 are rated on the basis of observed behavior and speech.

1. SOMATIC CONCERN: Degree of concern over present bodily health. Rate the degree to which physical health is perceived as a problem by the patient, whether complaints have realistic bases or not. Somatic delusions should be rated in the severe range with or without somatic concern. Note: Be sure to assess the degree of impairment due to somatic concerns only and not other symptoms, e.g., depression. In addition, if the subject rates a SI6 or "7" due to somatic delusions, then you must rate Unusual Thought Content at least a U4 or above.

Have you been concerned about your physical health? Have you had any physical illness or seen a medical doctor lately? (What does your doctor say is wrong? How serious is it?
Has anything changed regarding your appearance?
Has it interfered with your ability to perform your usual activities and/or work?
Did you ever feel that parts of your body had changed or stopped working?

[If patient reports any somatic concerns/delusions, ask the following]:

How often are you concerned about (use patient's description)? Have you expressed any of these concerns to others?

2 **Very Mild**
Occasional somatic concerns that tend to be kept to self.

3 **Mild**
Occasional somatic concerns that tend to be voiced to others (e.g., family, physician).

4 **Moderate**
Frequent expressions of somatic concern or exaggerations of existing ills OR some preoccupation, but no impairment in functioning. Not delusional.

5 **Moderately Severe**
Frequent expressions of somatic concern or exaggeration of existing ills OR some preoccupation and moderate impairment of functioning. Not delusional

6 **Severe**
Preoccupation with somatic complaints with much impairment in functioning OR somatic delusions without acting on them or disclosing to others.

7 **Extremely Severe**
Preoccupation with somatic complaints with severe impairment in functioning OR somatic delusions that tend to be acted on or disclosed to others.

2. ANXIETY: Reported apprehension, tension, fear, panic or worry. Rate only the patient's statements, not observed anxiety which is rated under TENSION.

Have you been worried a lot during (mention time frame)? Have you been nervous or apprehensive? (What do you worry about?)
Are you concerned about anything? How about finances or the future? When you are feeling nervous, do your palms sweat or does your heart beat fast (or shortness of breath, trembling, choking)?

[If patient reports anxiety or autonomic accompaniment, ask the following]:

How much of the time have you been (use patient's description)? Has it interfered with your ability to perform your usual activities/work?

2 **Very Mild**
Reports some discomfort due to worry OR infrequent worries that occur more than usual for most normal individuals.

3 **Mild**
Worried frequently but can readily turn attention to other things.

4 **Moderate**
Worried most of the time and cannot turn attention to other things easily but no impairment in functioning OR occasional anxiety with autonomic accompaniment but no impairment in functioning.

5 **Moderately Severe**
Frequent, but not daily, periods of anxiety with autonomic accompaniment OR some areas of functioning are disrupted by anxiety or worry.

6 **Severe**
Anxiety with autonomic accompaniment daily but not persisting throughout the day OR many areas of functioning are disrupted by anxiety or constant worry.

7 **Extremely Severe**
Anxiety with autonomic accompaniment persisting throughout the day OR most areas of functioning are disrupted by anxiety or constant worry.

3. DEPRESSION: Include sadness, unhappiness, anhedonia, and preoccupation with depressing topics (can't attend to TV or conversations due to depression), hopelessness, loss of self-esteem (dissatisfied or disgusted with self or feelings of worthlessness). Do not include vegetative symptoms, e.g., motor retardation, early waking, or the amotivation that accompanies the deficit syndrome.

How has your mood been recently? Have you felt depressed (sad, down, unhappy as if you didn't care)?
Are you able to switch your attention to more pleasant topics when you want to? Do you find that you have lost interest in or get less pleasure from things you used to enjoy, like family, friends, hobbies, watching TV, eating?

[If subject reports feelings of depression, ask the following]:

How long do these feelings last? Has it interfered with your ability to perform your usual activities/work?

2 **Very Mild**
Occasionally feels sad, unhappy or depressed.

3 **Mild**
Frequently feels sad or unhappy but can readily turn attention to other things.

4 **Moderate**
Frequent periods of feeling very sad, unhappy, moderately depressed, but able to function with extra effort.

5 **Moderately Severe**
Frequent, but not daily, periods of deep depression OR some areas of functioning are disrupted by depression.

6 **Severe**
Deeply depressed daily but not persisting throughout the day OR many areas of functioning are disrupted by depression.

7 **Extremely. Severe**
Deeply depressed daily OR most areas of functioning are disrupted by depression.

4. SUICIDALITY: Expressed desire, intent or actions to harm or kill self.

Have you felt that life wasn't worth living? Have you thought about harming or killing yourself? Have you felt tired of living or as though you would be better off dead? Have you ever felt like ending it all?

[If patient reports suicidal ideation, ask the following]:

How often have you thought about (use patient's description)? Did you (Do you) have a specific plan?

2 **Very Mild**
Occasional feelings of being tired of living. No overt suicidal thoughts.

3 **Mild**
Occasional suicidal thoughts without intent or specific plan OR he/she feels they would be better off dead.

4 **Moderate**
Suicidal thoughts frequent without intent or plan.

5 **Moderately Severe**
Many fantasies of suicide by various methods. May seriously consider making an attempt with specific time and plan OR impulsive suicide attempt using non-lethal method or in full view of potential saviors.

6 **Severe**
Clearly wants to kill self. Searches for appropriate means and time, OR potentially serious suicide attempt with patient knowledge of possible rescue.

7 **Extremely Severe**
Specific suicidal plan and intent (e.g., "as soon as _____ I will do it by doing X"), OR suicide attempt characterized by plan patient thought was lethal or attempt in secluded environment.

5. GUILT: Overconcern or remorse for past behavior. Rate only patient's statements, do not infer guilt feelings from depression, anxiety, or neurotic defenses. Note: If the subject rates a "6" or "7" due to delusions of guilt, then you must rate Unusual Thought Content at least a "4" or above depending on level of preoccupation and impairment.

Is there anything you feel guilty about? Have you been thinking about past problems? Do you tend to blame yourself for things that have happened? Have you done anything you're still ashamed of?

[If patient reports guilt/remorse/delusions, ask the following]:

How often have you been thinking about (use patient's description)? Have you disclosed your feelings of guilt to others?

2 **Very Mild**
Concerned about having failed someone or at something but not preoccupied. Can shift thoughts to other matters easily.

3 **Mild**
Concerned about having failed someone or at something with some preoccupation. Tends to voice guilt to others.

4 **Moderate**
Disproportionate preoccupation with guilt, having done wrong, injured others by doing or failing to do something, but can readily turn attention to other things.

5 **Moderately Severe**
Preoccupation with guilt, having failed someone or at something, can turn attention to other things, but only with great effort. Not delusional.

6 **Severe**

Delusional guilt OR unreasonable self-reproach very out of proportion to circumstances. Moderate preoccupation present.

7 Extremely Severe
Delusional guilt OR unreasonable self-reproach grossly out of proportion to circumstances. Subject is very preoccupied with guilt and is likely to disclose to others or act on delusions.

6. HOSTILITY: Animosity, contempt, belligerence, threats, arguments, tantrums, property destruction, fights and any other expression of hostile attitudes or actions. Do not infer hostility from neurotic defenses, anxiety or somatic complaints. Do not include incidents of appropriate anger or obvious self-defense.

How have you been getting along with people (family, co etc.)?
Have you been irritable or grumpy lately? (How do you show it? Do you keep it to yourself?)
Were you ever so irritable that you would shout at people or start fights or arguments? (Have you found yourself yelling at people you didn't know?) Have you hit anyone recently?

2 Very Mild
Irritable or grumpy, but not overtly expressed.

3 Mild
Argumentative or sarcastic.

4 Moderate
Overtly angry on several occasions OR yelled at others excessively.

5 Moderately Severe
Has threatened, slammed about or thrown things.

6 Severe
Has assaulted others but with no ham, likely, e.g., slapped or pushed, OR destroyed property, e.g., knocked over furniture, broken windows.

7 Extremely Severe
Has attacked others with definite possibility of harming them or with actual harm, e.g., assault with hammer or weapon.

7. ELEVATED MOOD: A pervasive, sustained and exaggerated feeling of well-being, cheerfulness, euphoria (implying a pathological mood), optimism that is out of proportion to the circumstances. Do not infer elation from increased activity or from grandiose statements alone.

Have you felt so good or high that other people thought that you were not your normal self?
Have you been feeling cheerful and "on top of the world" without any reason?

[If patient reports elevated mood/euphoria, ask the following]:

Did it seem like more than just feeling good? How long did that last?

2 **Very Mild**
Seems to be very happy, cheerful without much reason.

3 **Mild**
Some unaccountable feelings of well-being that persist.

4 **Moderate**
Reports excessive or unrealistic feelings of well-being, cheerfulness, confidence or optimism inappropriate to circumstances, some of the time. May frequently joke, smile, be giddy or overly enthusiastic OR few instances of marked elevated mood with euphoria.

5 **Moderately Severe**
Reports excessive or unrealistic feelings of well-being, confidence or optimism inappropriate to circumstances much of the time. May describe feeling on top of the world," "like everything is falling into place," or "better than ever before," OR several instances of marked elevated mood with euphoria.

6 **Severe**
Reports many instances of marked elevated mood with euphoria OR mood definitely elevated almost constantly throughout interview and inappropriate to content

7 **Extremely Severe**
Patient reports being elated or appears almost intoxicated, laughing, joking, giggling, constantly euphoric, feeling invulnerable, all inappropriate to immediate circumstances.

8. GRANDIOSITY: Exaggerated self-opinion, self-enhancing conviction of special abilities or powers or identity as someone rich or famous. Rate only patient's statements about himself, not his demeanor. Note: If the subject rates a "6" or "7" due to grandiose delusions, you must rate Unusual Thought Content at least a "4" or above.

Is there anything special about you? Do you have any special abilities or powers? Have you thought that you might be somebody rich or famous?

[If the patient reports any grandiose ideas/delusions, ask the following]:

How often have you been thinking about (use patient's description]? Have you told anyone about what you have been thinking? Have you acted on any of these ideas?

2 **Very Mild**
Feels great and denies obvious problems, but not unrealistic.

3 **Mild**
Exaggerated self-opinion beyond abilities and training.

4 **Moderate**

Inappropriate boastfulness, claims to be brilliant, insightful, or gifted beyond realistic proportions, but rarely self-discloses or acts on these inflated self-concepts. Does not claim that grandiose accomplishments have actually occurred.

5 **Moderately Severe**
Same as 4 but often self-discloses and acts on these grandiose ideas. May have doubts about the reality of the grandiose ideas. Not delusional.

6 **Severe**
Delusional—claims to have special powers like ESP, to have millions of dollars, invented new machines, worked at jobs when it is known that he was never employed in these capacities, be Jesus Christ, or the President. Patient may not be very preoccupied.

7 **Extremely Severe**
Delusional—Same as 6 but subject seems very preoccupied and tends to disclose or act on grandiose delusions.

9. SUSPICIOUSNESS: Expressed or apparent belief that other persons have acted maliciously or with discriminatory intent. Include persecution by supernatural or other nonhuman agencies (e.g., the devil). Note: Ratings of "3" or above should also be rated under Unusual Thought Content.

Do you ever feel uncomfortable in public? Does it seem as though others are watching you? Are you concerned about anyone's intentions toward you?
Is anyone going out of their way to give you a hard time, or tiying to hurt you? Do you feel in any danger?

[If patient reports any persecutory ideas/delusions, ask the following]:

How often have you been concerned that [use patient's description]? Have you told anyone about these experiences?

2 **Very Mild**
Seems on guard. Reluctant to respond to some "personal" questions. Reports being overly self-conscious in public.

3 **Mild**
Describes incidents in which others have harmed or wanted to harm him/her that sound plausible. Patient feels as if others are watching. laughing, or criticizing him/her in public, but this occurs only occasionally or rarely. Little or no preoccupation.

4 **Moderate**
Says others are talking about him/her maliciously, have negative intentions, or may harm him/her. Beyond the likelihood of plausibility, but not delusional. Incidents of suspected persecution occur occasionally (less than once per week) with some preoccupation.

5 **Moderately Severe**

Same as 4, but incidents occur frequently, such as more than once per week. Patient is moderately preoccupied with ideas of persecution OR patient reports persecutory delusions expressed with much doubt (e.g., partial delusion).

6 Severe
Delusional -- speaks of Mafia plots, the FBI, or others poisoning his/her food, persecution by supernatural forces.

7 Extremely Severe
Same as 6, but the beliefs are bizarre or more preoccupying. Patient tends to disclose or act on persecutory delusions.

10. HALLUCINATIONS: Reports of perceptual experiences in the absence of relevant external stimuli. When rating degree to which functioning is disrupted by hallucinations, include preoccupation with the content and experience of the hallucinations, as well as functioning disrupted by acting out on the hallucinatory content (e.g., engaging in deviant behavior due to command hallucinations). Include thoughts aloud ("gedankenlautwerden") or pseudohallucinations(e.g., hears a voice inside head) if a voice quality is present.

Do you ever seem to hear your name being called?
Have you heard any sounds or people talking to you or about you when there has been nobody around? (If hears voices): What do the voice/voices say? Did it have a voice quality?
Do you ever have visions or see things that others do not see? What about smell odors that others do not smell?

[If the patient reports hallucinations, ask the following]:

Have these experiences interfered with your ability to perform your usual activities/work? How do you explain them? How often do they occur?

2 Very Mild
While resting or going to sleep, sees visions, smells odors, or hears voices, sounds or whispers in the absence of external stimulation, but no impairment in functioning.

3 Mild
While in a clear state of consciousness, hears a voice calling the subjects name, experiences non-verbal auditory hallucinations (e.g., sounds or whispers), formless visual hallucinations, or has sensory experiences in the presence of a modality-relevant stimulus (e.g., visual illusions) infrequently (e.g., 1-2 times per week) and with no functional impairment.

4 Moderate
Occasional verbal, visual, gustatory, olfactory, or tactile hallucinations with no functional impairment OR non-verbal auditory hallucinations/visual illusions more than infrequently or with impairment.

5 Moderately Severe

Experiences daily hallucinations OR some areas of functioning are disrupted by hallucinations.

6 **Severe**
Experiences verbal or visual hallucinations several times a day OR many areas of function are disrupted by these hallucinations.

7 **Extremely Severe**
Persistent verbal or visual hallucinations throughout the day OR most areas of functioning are disrupted by these hallucinations.

11. UNUSUAL THOUGHT CONTENT: Unusual, odd, strange or bizarre thought content. Rate the degree of unusualness, not the degree of disorganization of speech. Delusions are patently absurd, clearly false or bizarre ideas that are expressed with full conviction. Consider the patient to have full conviction if he/she has acted as though the delusional belief were true. Ideas of reference/persecution can be differentiated from delusions in that ideas are expressed with much doubt and contain more elements of reality. Include thought insertion, withdrawal and broadcast. Include grandiose, somatic and persecutory delusions even if rated elsewhere. Note: if Somatic Concern, Guilt, Suspiciousness, or Grandiosity are rated "6" or "7" due to delusions, then Unusual Thought Content must be rated a "4" or above.

Have you been receiving any special messages from people or from the way things are arranged around you? Have you seen any references to yourself on TV or in the newspapers?
Can anyone read your mind?
Do you have a special relationship with God?
Is anything like electricity, X-rays, or radio waves affecting you?
Are thoughts put into your head that are not your own?
Have you felt that you were under the control of another person or force?

[If patient reports any odd ideas/delusions, ask the following]:

How often do you think about (use patient's description)?
Have you told anyone about these experiences? How do you explain the things that have been happening?

2 **Very Mild**
Ideas of reference (people may stare or may laugh at him), ideas of persecution (people may mistreat him). Unusual beliefs in psychic powers, spirits, UFO's, or unrealistic beliefs in one's own abilities. Not strongly held. Some doubt.

3 **Mild**
Same as 2, but degree of reality distortion is more severe as indicated by highly unusual ideas or greater conviction. Content may be typical of delusions (even bizarre), but without full conviction. The delusion does not seem to have fully formed, but is considered as one possible explanation for an unusual experience.

4 **Moderate**

Delusion present but no preoccupation or functional impairment. May be an encapsulated delusion or a firmly endorsed absurd belief about past delusional circumstances.

5 **Moderately Severe**
Full delusion(s) present with some preoccupation OR some areas of functioning disrupted by delusional thinking.

6 **Severe**
Full delusion(s) present with much preoccupation OR many areas of functioning are disrupted by delusional thinking.

7 **Extremely Severe**
Full delusions present with almost total preoccupation OR most areas of functioning are disrupted by delusional thinking.

Rate items 12-13 on the basis of patient's self-report and observed behavior.

12. BIZARRE BEHAVIOR: Reports of behaviors which are odd, unusual, or psychotically criminal. Not limited to interview period. Include inappropriate sexual behavior and inappropriate affect.

Have you done anything that has attracted the attention of others? Have you done anything that could have gotten you into trouble with the police?
Have you done anything that seemed unusual or disturbing to others?

2 **Very Mild**
Slightly odd or eccentric public behavior, e.g., occasionally giggles to self, fails to make appropriate eye contact, that does not seem to attract the attention of others OR unusual behavior conducted in private, e.g., innocuous rituals, that would not attract the attention of others.

3 **Mild**
Noticeably peculiar public behavior, e.q., inappropriately loud talking, makes inappropriate eye contact, OR private behavior that occasionally, but not always, attracts the attention of others, e.g., hoards food, conducts unusual rituals, wears gloves indoors.

4 **Moderate**
Clearly bizarre behavior that attracts or would attract (if done privately) the attention or concern of others, but with no corrective intervention necessary. Behavior occurs occasionally, e.g., fixated staring into space for several minutes, talks back to voices once, in appropriate giggling/laughter on 1-2 occasions, talking loudly to self.

5 **Moderately Severe**
Clearly bizarre behavior that attracts or would attract (if done privately) the attention of others or the authorities, e.g., fixated staring in a socially disruptive way, frequent inappropriate giggling/laughter, occasionally responds to voices, or eats non-foods.

6 **Severe**
Bizarre behavior that attracts attention of others and intervention by authorities, e.g., directing traffic, public nudity, staring into space for long periods, carrying on a conversation with hallucinations, frequent inappropriate giggling/laughter.

7 **Extremely Severe**
Serious crimes committed in a bizarre way that attracts the attention of others and the control of authorities e.g., sets fires and stares at flames OR almost constant bizarre behavior, e.g., inappropriate gig gling/laughter, responds only to hallucinations and cannot be engaged in interaction.

13. SELF-NEGLECT: Hygiene, appearance, or eating behavior below usual expectations, below socially acceptable standards, or life-threatening.

How has your grooming been lately? How often do you change your clothes? How often do you take showers? Has anyone (parents/staff) complained about your grooming or dress? Do you eat regular meals?

2 **Very Mild**
Hygiene/appearance slightly below usual community standards, e.g, shirt out of pants, buttons unbuttoned, shoe laces untied, but no social or medical consequences.

3 **Mild**
Hygiene/appearance occasionally below usual community standards, e.g., irregular bathing, clothing is stained, hair uncombed, occasionally skips an important meal. No social or medical consequences

4 **Moderate**
Hygiene/appearance is noticeably below usual community standards, e.q., fails to bathe or change clothes, clothing very soiled, hair unkempt, needs prompting, noticeable by others OR irregular eating and drinking with minimal medical concerns and consequences.

5 **Moderately Severe**
Several areas of hygiene/appearance are below usual community standards OR poor grooming draws criticism by others, and requires regular prompting. Eating or hydration is irregular and poor, causing some medical problems.

6 **Severe**
Many areas of hygiene/appearance are below usual community standards, does not always bathe or change clothes even if prompted. Poor grooming has caused social ostracism at school/residence/work, or required intervention. Eating erratic and poor, may require medical intervention.

7 **Extremely Severe**
Most areas of hygiene/appearance/nutrition are extremely poor and easily noticed as below usual community standards OR hygiene/appearance/nutrition requires urgent and immediate medical intervention.

14. DISORIENTATION: Does not comprehend situations or communications, such as questions asked during the entire BPRS interview. Confusion regarding person, place, or time. Do not rate if incorrect responses are due to delusions.

May I ask you some standard questions we ask everybody?
How old are you? What is the date? (allow + or - 2 days]
What is this place called? What year were you born? Who is the president?

2 **Very Mild**
Seems muddled or mildly confused 1-2 times during interview. Oriented to person, place, and time.

3 **Mild**
Occasionally muddled or mildly confused 3-4 times during interview. Minor inaccuracies in person, place, or time, e.g., date off by more than + or - 2 days, or gives wrong division of hospital.

4 **Moderate**
Frequently confused during interview. Minor inaccuracies in person, place, or time are noted, as in "3" above. In addition, may have difficulty remembering general information, e.g., name of president.

5 **Moderately Severe**
Markedly confused during interview, or to person, place, or time. Significant inaccuracies are noted, e.g., date off by more than one week, or cannot give correct name of hospital. Has difficulty remembering personal information, e.g., where he/she was born, or recognizing familiar people.

6 **Severe**
Disoriented to person, place, or time, e.g., cannot give correct month and year. Disoriented in 2 out of 3 spheres.

7 **Extremely Severe**
Grossly disoriented to person, place, or time, e.g., cannot give name or age. Disoriented in all 3 spheres.

Rate items 15-24 on the basis of observed behavior and speech.

15. CONCEPTUAL DISORGANIZATION: Degree to which speech is confused, disconnected, vague or disorganized. Rate tangentiality, circumstantiality, sudden topic shifts, incoherence, derailment, blocking, neologisms, and other speech disorders. Do not rate content of speech.

2 **Very Mild**
Peculiar use of words or rambling but speech is comprehensible.

3 **Mild**

Speech a bit hard to understand or make sense of due to tangentiality, circumstantiality or sudden topic shifts.

4 Moderate
Speech difficult to understand due to tangentiality, circumstantiality, idiosyncratic speech, or topic shifts on many occasions OR 1-2 in stances of incoherent phrases.

5 Moderately Severe
Speech difficult to understand due to circumstantiality, tangentiality, neologisms, blocking, or topic shifts most of the time OR 3-5 instances of incoherent phrases.

6 Severe
Speech is incomprehensible due to severe impairments most of the time. Many BPRS items cannot be rated by self-report alone.

7 Extremely Severe
Speech is incomprehensible throughout interview.

16. BLUNTED AFFECT: Restricted range in emotional expressiveness of face, voice, and gestures. Marked indifference or flatness even when discussing distressing topics. In the case of euphoric or dysphoric patients, rate Blunted Affect if a flat quality is also clearly present.

Use the following probes at end of interview to assess emotional responsivity:

Have you heard any good jokes lately? Would you like (hear a joke?

2 Very Mild
Emotional range is slightly subdued or reserved but displays appropriate facial expressions and tone of voice that are within normal limits.

3 Mild
Emotional range overall is diminished, subdued, or reserved, without many spontaneous and appropriate emotional responses. Voice tone is slightly monotonous.

4 Moderate
Emotional range is noticeably diminished, patient doesn't show emotion, smile, or react to distressing topics except infrequently. Voice tone is monotonous or there is noticeable decrease in spontaneous movements. Displays of emotion or gestures are usually followed by a return to flattened affect.

5 Moderately Severe
Emotional range very diminished, patient doesn't show emotion, smile or react to distressing topics except minimally, few gestuleb, facial expression does not change very often. Voice tone is monotonous much of the time.

6 Severe

Very little emotional range or expression. Mechanical in speech and gestures most of the time. Unchanging facial expressions. Voice tone is monotonous most of the time.

7 **Extremely Severe**
Virtually no emotional range or expressiveness, stiff movements. Voice tone is monotonous all of the time.

17. EMOTIONAL WITHDRAWAL: Deficiency in patient's ability to relate emotionally during interview situation. Use your own feeling as to the presence of an "invisible barrier" between patient and interviewer. Include withdrawal apparently due to psychotic processes.

2 **Very Mild**
Lack of emotional involvement shown by occasional failure to make reciprocal comments, occasionally appearing preoccupied, or smiling in a stilted manner, but spontaneously engages the interviewer most of the time.

3 **Mild**
Lack of emotional involvement shown by noticeable failure to make reciprocal comments, appearing preoccupied, or lacking in warmth, but responds to interviewer when approached.

4 **Moderate**
Emotional contact not present much of the interview because subject does not elaborate responses, fails to make eye contact, doesn't seem to care if interviewer is listening, or may be preoccupied with psychotic material.

5 **Moderately Severe**
Same as "4" but emotional contact not present most of the interview.

6 **Severe**
Actively avoids emotional participation. Frequently unresponsive or responds with yes/no answers (not solely due to persecutory delusions). Responds with only minimal affect.

7 **Extremely Severe**
Consistently avoids emotional participation. Unresponsive or responds with yes/no answers (not solely due to persecutory delusions). May leave during interview or just not respond at all.

18. MOTOR RETARDATION: Reduction in energy level evidenced by slowed movements and speech, reduced body tone, decreased number of spontaneous body movements. Rate on the basis of observed behavior of the patient only. Do not rate on the basis of patient's subjective impression of his own energy level. Rate regardless of the medication effects.

2 **Very Mild**
Slightly slowed or reduced movements or speech compared to most people.

3 **Mild**
Noticeably slowed or reduced movements or speech compared to most people.

4 **Moderate**
Large reduction or slowness in movements or speech.

5 **Moderately Severe**
Seldom moves or speaks spontaneously OR very mechanical or stiff movements.

6 **Severe**
Does not move or speak unless prodded or urged.

7 **Extremely Severe**
Frozen, catatonic.

19. TENSION: Observable physical and motor manifestations of tension, "nervousness," and agitation. Self-reported experiences of tension should be rated under the item on anxiety. Do not rate if restlessness is solely akathisia, but do rate if akatihsia is exacerbated by tension.

2 **Very Mild**
More fidgety than most but within normal range. A few transient signs of tension, e.g., picking at fingernails, foot wagging, scratching scalp several times, or finger tapping.

3 **Mild**
Same as "2," but with more frequent or exaggerated signs of tension.

4 **Moderate**
Many and frequent signs of motor tension with one or more signs some times occurring simultaneously, e g., wagging one's foot while wringing hands together. There are tirn when no signs of tension are present.

5 **Moderately Severe**
Many and frequent signs of motor tension with one or more signs often occurring simultaneously. There are still rare times when no signs of tension are present.

6 **Severe**
Same as "5", but signs of tension are continuous.

7 **Extremely Severe**
Multiple motor manifestations of tension are continuously present, e.g., continuous pacing and hand wringing.

20. UNCOOPERATIVENESS: Resistance and lack of willingness to cooperate with the interview. The uncooperativeness might result from suspiciousness. Rate only uncooperativeness in relation to the interview, not behaviors involving peers and relatives.

2 **Very Mild**
Shows nonverbal signs of reluctance, but does not complain or argue. 3 Mild
Gripes or tries to avoid complying, but goes ahead without argument.

3 **Mild**

Gripes or tries to avoid complying, but goes ahead without argument

4 Moderate
Verbally resists but eventually complies after questions are rephrased or repeated.

5 Moderately Severe
Same as 4, but some information necessary for accurate ratings is withheld.

6 Severe
Refuses to cooperate with interview, but remains in interview situation.

7 Extremely Severe
Same as 6, with active efforts to escape the interview.

21. EXCITEMENT: Heightened emotional tone, or increased emotional reactivity to interviewer or topics being discussed, as evidenced by increased intensity of facial expressions, voice tone, expressive gestures or increase in speech quantity and speed.

2 Very Mild
Subtle and fleeting or questionable increase in emotional intensity. For example, at times seems keyed-up or overly alert.

3 Mild
Subtle but persistent increase in emotional intensity. For example, lively use of gestures and variation in voice tone.

4 Moderate
Definite but occasional increase in emotional intensity. For example, reacts to interviewer or topics that are discussed with noticeable emotional intensity. Some pressured speech.

5 Moderately Severe
Definite and persistent increase in emotional intensity. For example reacts to many stimuli, whether relevant or not, with considerable emotional intensity. Frequent pressured speech.

6 Severe
Marked increase in emotional intensity. For example reacts to most stimuli with inappropriate emotional intensity. Has difficulty settling down or staying on task. Often restless, impulsive, or speech is often pressured.

7 Extremely Severe
Marked and persistent increase in emotional intensity. Reacts to all stimuli with inappropriate intensity, impulsiveness. Cannot settle down or stay on task. Very restless and impulsive most of the time. Constant pressured speech.

22. DISTRACTIBILITY: Degree to which observed sequences of speech and actions are interrupted by stimuli unrelated to the interview. Distractibility is rated when the patient shows a change in the focus of attention or a marked shift in gaze. Patient's attention may be drawn to noise in adjoining room, books on shelf, interviewer's clothing, etc. Do not rate circumstantiality,

tangentiality, or flight of ideas. Also, do not rate rumination with delusional material. Rate even if the distracting stimulus cannot be identified.

2 **Very Mild**
Generally can focus on interviewer's questions with only 1 distraction or inappropriate shift of attention of brief duration.

3 **Mild**
Patient shifts focus of attention to matters unrelated to the interview 2-3 times.

4 **Moderate**
Often responsive to irrelevant stimuli in the room, e.g., averts gaze from the interviewer.

5 **Moderately Severe**
Same as above, but now distractibility clearly interferes with the flow of the interview.

6 **Severe**
Extremely difficult to conduct interview or pursue a topic due to preoccupation with irrelevant stimuli.

7 **Extremely Severe**
Impossible to conduct interview due to preoccupation with irrelevant stimuli.

23. MOTOR HYPERACTIVITY: Increase in energy level evidenced in more frequent movement and/or rapid speech. Do not rate if restlessness is due to akathisia.

2 **Very Mild**
Some restlessness, difficulty sitting still, lively facial expressions, or somewhat talkative.

3 **Mild**
Occasionally very restless, definite increase in motor activity, lively gestures, 1-3 brief instances of pressured speech.

4 **Moderate**
Very restless, fidgety, excessive facial expressions or nonproductive and repetitious motor movements. Much pressured speech, up to one
third of the interview.

5 **Moderately Severe**
Frequently restless, fidgety. Many instances of excessive non productive and repetitious motor movements. On the move most of the time. Frequent pressured speech, difficult to interrupt. Rises on 1-2 occasions to pace.

6 **Severe**
Excessive motor activity, restlessness, fidgety, loud tapping, noisy, etc. throughout most of the interview. Speech can only be interrupted with much effort. Rises on 3-4 occasions to pace.

7 Extremely Severe
Constant excessive motor activity throughout entire interview, e.g., constant pacing, constant pressured speech with no pauses, interviewee can only be interrupted briefly and only small amounts of relevant information can be obtained.

24. MANNERISMS AND POSTURING: Unusual and bizarre behavior, stylized movements or acts, or any postures which are clearly uncomfortable or inappropriate. Exclude obvious manifestations of medication side-effects. Do not include nervous mannerisms that are not odd or unusual.

2 Very Mild
Eccentric or odd mannerisms or activity that ordinary persons would have difficulty explaining, e.g., grimacing, picking. Observed once for a brief period.

3 Mild
Same as "2," but occurring on two occasions of brief duration.

4 Moderate
Mannerisms or posturing, e g., stylized movements or acts, rocking, nodding, rubbing, or grimacing, observed on several occasions for brief periods or infrequently but very odd. For example, uncomfortable posture maintained for 5 seconds more than twice.

5 Moderately Severe
Same as "4," but occurring often, or several examples of very odd mannerisms or posturing that are idiosyncratic to the patient.

6 Severe
Frequent stereotyped behavior, assumes and maintains uncomfortable or inappropriate postures, intense rocking, smearing, strange rituals, or fetal posturing. Subject can interact with people and the environment for brief periods despite these behaviors.

7 Extremely Severe
Same as "6" but subject cannot interact with people or the environ ment due to these behaviors.