Classifying Symptom Change in Eating Disorders: Clinical Significance Metrics for the Change in Eating Disorder Symptoms Scale

Anthony Daniel Hwang

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

Doctor of Philosophy

Diane L. Spangler
Gary M. Burlingame
Bruce N. Carpenter
Bruce L. Brown
Ross Flom

Department of Psychology

Brigham Young University

August 2010

ABSTRACT

Classifying Symptom Change in Eating Disorders: Clinical Significance Metrics for the Change in Eating Disorder Symptoms Scale

Anthony D. Hwang

Department of Psychology

Doctor of Philosophy

Despite well-established diagnostic measures and measures of specific dimensions of eating disorder symptomatology, little work has been done to develop a brief, comprehensive, and valid measure for assessing change in eating disorder symptoms. Further, empirically-supported change indices to assess treatment progression and outcome have not yet been developed. The Change in Eating Disorder Symptoms Scale (CHEDS) is a new comprehensive measure designed to assess progress and change during treatment in persons with diagnoses on the eating disorder spectrum. Previous studies have demonstrated the subscale structure, reliability, and validity of the CHEDS. This study determined clinically significant change criteria for the CHEDS to complement the studies that have supported the CHEDS as a reliable and valid measure of eating disorder symptomatology. The CHEDS was also compared to a life functioning scale, the Clinical Impairment Assessment. A reliable change index (RCI) was developed, which generated an inferential statistic that estimates the magnitude of change in a score necessary for a change score to be considered statistically reliable. A cutscore was also developed, which differentiates between functional and dysfunctional populations, between eating disordered clinical subjects and non-clinical subjects. Trajectories were identified using hierarchical linear modeling methods for use in conjunction with clinical significance criteria to aid in the tracking of symptoms during treatment, treatment decision-making, and tailoring treatment according to expected and observed progress. The clinical significance change criteria were then applied to the clinical sample to determine change patterns descriptive of recovered, reliable improvement, deterioration, and no change. Finally, a scoring program with clinical significance change criteria and trajectory analyses for total and subscale scores was developed.

KEYWORDS: Change in Eating Disorder Symptoms Scale, eating disorders, anorexia nervosa, bulimia nervosa, eating disorder not-otherwise-specified, clinically significant change, hierarchical linear modeling, psychometric change indices, outcome, reliable change index.

ACKNOWLEDGEMENTS

It is hard to find the words to be able to write my acknowledgements of my dissertation as I feel that it is a culmination of a very meaningful chapter of my life. It is amazing how fast the last five years have gone by but the growth that I have experienced as a professional and clinician has been immense. In the last five years, I have experienced the ups and downs of my clinical program; I became an uncle, had the unfortunate losses of family members, made many wonderful and long-lasting friendships, and I met my eternal companion. I would like to thank Dr. Diane Spangler, my mentor, for her continued guidance and patience. I genuinely believe I would not be the clinician I am today without her direction. Also a lot of gratitude to Dr. Gary Burlingame for years of encouragement and helping me refine my edges. I also owe thanks to many other supervisors, professors, and friends for the abundance of unsolicited support.

My parents have worked so hard to support me emotionally as well as financially. I am proud to be their son and to not only receive a Ph.D. for myself but for them as well. My sister Amy, brother-in-law Brian, niece Aubrey, and my dog Rocky also never stopped cheering for me from the sidelines. I will also always cherish the love and care from my extended family, particularly my late uncle Henry Hwang and late aunt Diane Most, whom I miss very much. Adrianna, my wife, has been with me literally every step of the way. From standing by my side while I worked on testing reports, externship work, studying for comps and finals, applying for internship, being apart for most of my internship, applying for jobs, to this dissertation defense. Her support and love has never wavered despite the stresses of a clinical psychology program. I would be gravely remiss if I did not thank my eternal companion and I am so grateful and blessed to have her in my life. Not lastly, all this is possible through the grace of my Heavenly Father and the blessings He has bestowed upon me and upon my family.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

Figures:

Tables:

Classifying Symptom Change in Eating Disorders:

Clinical Significance Metrics for the Change in Eating Disorder Symptoms Scale

There is no comprehensive measure for eating disorders that can be used on a session-by-session basis with empirically-validated change indices to gauge progress. Moreover, recovery from eating disorders is not well-defined in the literature, as noted by Jarman and Walsh (1999), who state that there is "an absence of an agreed upon definition of the term *recovered* within the eating disorder literature…[and that clinicians] often make implicit assumptions of the definition and meaning of recovery" (p. 775).

The current assessment of change during treatment in eating disorder research is primarily categorical, based on whether or not a diagnostic threshold is reached for a disorder (Keel, Mitchell, Miller, et al., 1999). The typical research definition of eating disorder recovery is defined and determined by patients moving from meeting to not meeting established diagnostic criteria for an eating disorder. Numerous efficacy and effectiveness studies of treatments for eating disorders have employed this definition of outcome. The utilization of a diagnostic criterion as an indicator of change is beneficial as it provides a clearer distinction between those who do and do not meet criteria for an eating disorder and can provide a clear determination of outcome in patients. However, this method of defining treatment progress has various shortcomings, outlined below. Unidimensional measures, measures which assess a specific dimension of eating disorders, have also been used to assess change in eating disorder symptomatology during treatment but are, by design, limited to a single or very few symptom domains.

A comprehensive measure of eating disorder symptoms, which goes beyond current methods of assessing eating disorder symptoms, is needed. This study sought to develop change

indices for the Change in Eating Disorder Symptoms Scale (CHEDS), which was designed for session-by-session tracking of eating disorder symptom change. The CHEDS utilizes a different method of determining symptom change and outcome that is based on comprehensive symptoms of eating disorders, linked to tracking changes in symptomatology to circumvent the drawbacks of utilizing a categorical, diagnostic based determination of outcome. The previous absence of a comprehensive tracking and outcome eating disorder measure with change indices results in a clarion call for its development.

Existing Measures of Eating Disorders

Various types of eating disorder measures are currently used. Each type was developed for various purposes. Diagnostic measures are instruments designed to assess the presence or absence of an eating disorder based upon diagnostic criteria typically derived from the Diagnostic and Statistical Manual for Mental Disorders (DSM: American Psychological Association, 2000) or the International Classification of Disease (ICD: Medical Management Institute, 2008). These instruments classify individuals into mutually exclusive categories of either meeting criteria for a specific of type of eating disorder (e.g., anorexia nervosa (AN), bulimia nervosa (BN), or eating disorder not otherwise specified (EDNOS)). In contrast, unidimensional measures are designed to assess one or a few specific symptoms of eating disorders and tend to be continuous measures, although cutpoints are occasionally established to indicate when a score falls in a categorical clinical versus non-clinical range. The following will examine these two types of measures used for assessing change in eating disorder symptomatology. The uses and limitations of each type of measure for tracking symptoms and assessing recovery will also be explicated.

Diagnostic measures. Both interviewer-based and self-report measures have been developed to assess the presence or absence of eating disorders.

Interview based diagnostic measures. Some eating disorder symptoms are considered to be complex and/or ambiguous. Concerns that self-report instruments may not be capable of adequately assessing eating disorder symptoms led to the development of the interview diagnostic format (Cooper & Fairburn, 1987). For example, it was assumed that individuals completing a self-report diagnostic instrument may have an arbitrary definition of some symptoms, such as binge eating or restrictive eating, and those definitions would qualitatively vary among individuals. It was also assumed that the severity of symptoms might be difficult to determine by respondents. Thus, interview-based diagnostic measures, in which a trained individual interviews respondents and then rates their responses based on operationalized criteria, were developed to enable more standardized symptom assessment. This class of diagnostic instruments has considerable inter-rater reliability, ranging between .75 and .99, and convergent and discriminant validities in the moderate ranges (Clinton & Norring, 1999; Cooper & Fairburn, 1987; Fairburn & Cooper, 1993; Gharderi & Scott, 2002; Rizvi, Peterson, Crow, & Agras, 2000; Rosen, Vara, Wednt, & Leitenberg, 1990; Sysko, Walsh, & Fairburn, 2005).

Of the interview-based assessments, the most widely used and researched is the Eating Disorder Examination (EDE). The EDE is semi-structured in format, typically requires nearly an hour to administer, and has undergone several revisions (Guest, 2000, Cooper & Fairburn, 1987, Fairburn & Cooper, 1993). It is designed to be sensitive to the presence or absence of an eating disorder, assessing the occurrence of symptoms on 28-day intervals, consistent with current diagnostic criteria for defining the presence of an eating disorder (Sysko, et al., 2005). Several other instruments have been designed for the similar purpose of diagnosing eating disorders,

including the Rating of Anorexia and Bulimia interview (RAB; Clinton & Norring, 1999) and the Interview for Diagnosis of Eating Disorders (IDED-IV; Kutlesic, Williamson, Gleaves, et al., 1998). These measures have also been shown to be reliable and valid measures of diagnosis, although not as comprehensively empirically evaluated as is the EDE.

Despite the strengths of these diagnostic interviews in detecting the presence or absence of eating disorders and their use as reliable outcome measures, they are ill suited for assessing change *during* the course of treatment. One primary reason is that diagnostic interviews are not temporally sensitive. The interview format categorically assesses symptoms of eating disorders in long time intervals, which prevents clinicians from tracking symptoms on a weekly basis (Binford, Le Grange, & Jellar, 2005). For example, the EDE, as previously noted, by design inquires about 28-day intervals. Furthermore, while interviewer-based diagnostic instruments, of which the EDE is the gold-standard for assessing eating disorders, are valid and reliable measures of presence or absence of specific categories of eating disorders, they are time and labor intensive, requiring specialized training to administer and are not feasible to administer on a repeated basis during treatment due the lack of brevity of the assessment.

The use of a nominal level of symptom analysis, typical of interviewer-based diagnostic measures, does not allow for measurement of incremental individual progress or change. The EDE, for example, focuses and is more concerned with detecting the presence or absence of an eating disorder rather than the severity of the eating disorder or specific symptoms displaying elevations. When tracking symptom change during treatment, information on the changes in specific symptoms is of particular interest. Clinicians can use this information to inform treatment, such as tailoring treatment, by assessing which symptoms are elevated. While diagnosis as a criterion for outcome indicates whether an individual has a diagnosable level of an

eating disorder, it cannot inform a clinician if an individual is improving beyond the presence or absence of a disorder or on specific dimensions of eating disorder symptoms. Categorically-based diagnostic information is useful in ascertaining the extent to which the presence of an eating disorder exists, but this information does not help a clinician on the symptom level, nor can this be done on a session-by-session basis using diagnostic measures.

Also problematic for current diagnostic instruments for eating disorders is the EDNOS diagnosis found in the DSM-IV-TR (APA, 2000). Most diagnostic measures do not distinguish very well between subthreshold eating disorders and the absence of an eating disorder diagnosis (Mitnz, et al., 1997; Vetrone, Cuzzolaro, Antonozzi, & Garfinkel, 2006). Those persons who do not quite meet the diagnostic criteria for AN or BN but are displaying some clinically significant eating and/or body image symptoms are aggregated in the EDNOS category, yet this category has no specific definition and vague criteria for inclusion. Furthermore, change in this category is particularly difficult to ascertain using categorical measures since individuals can change significantly on some symptoms yet still meet criteria for EDNOS based on other symptom sets. Categorical measures can also miss significant symptoms in those who may recover categorically from AN or BN. For example, an individual who at pretreatment met the diagnosis of anorexia or bulimia and improves in therapy and no longer meets the full diagnostic criteria for an eating disorder may continue to show significant symptoms and signs of eating disorders and impairment, yet be considered recovered based on categorical diagnostic measures. Consequently, utilizing a categorical diagnostic measure as an assessment of outcome with the EDNOS and subthreshold diagnoses is problematic.

Given the vagueness of the criteria used for diagnosing EDNOS, few if any current diagnostic measures even classify EDNOS. The method of how recovery is defined for EDNOS

has also not been developed. The EDNOS diagnosis is considered to be severe in symptomatology and psychological dysfunction (Nollett & Button, 2005), yet is often considered a less severe eating disorder despite being a viable diagnosis indicating no significant differences from other eating disorders on various measures of impairment in functioning (le Grange, Binford, Peterson, et al., 2006). The EDNOS diagnosis is the most commonly diagnosed category of eating disorders, is severe and persistent, and should not be merely considered a subthreshold of AN or BN (Crow, Agras, Halmi, et al., 2002). Recent work also suggests that there are fundamentally different characteristics between AN, BN, and EDNOS (Fairburn, Cooper, Bohn, et al., 2007). There was no significant difference in the "restraint" subscales between groups, indicating similarities, while differences were found in body mass index (BMI), preoccupation with food, and fear of losing control. Significant differences were found in relation to "importance of shape," "fear of weight gain," and "desire to lose weight," signifying that EDNOS is not just a mild or subthreshold form of anorexia or bulimia (Turner & Bryant-Waugh, 2004). The difficulty of operationalizing the EDNOS diagnostic category by current categorical diagnostic measures suggests the potential utility of measures using a more dimensional approach, rather than categorical, to assessing eating disorder symptoms.

By design, diagnostic instruments are not constructed to be used as symptom tracking and change measures. When used to track symptom change and determining outcome, which extends beyond the construct validity of such measures, the definitions of recovery (or sufficient change) are inconsistent. Keel, Mitchell, Davis, Fieselman, and Crow (1999) conducted a meta-analysis examining definitions and predictions of eating disorder recovery and found varying definitions. One study defined recovery as 9 consecutive weeks without eating disorder symptoms while another study used 12-month criteria. There was no statistical or clinically

significant index that an individual would be required to meet to be considered recovered. Without such a change index, it is difficult to determine if a client is "on track" in regard to progress during treatment, to determine prognosis given response patterns, or to determine if and when treatment should be modified or ended. Consequently, clinicians are unable to address whether the symptoms a patient is experiencing are improving even when meeting criteria for an eating disorder as there are no psychometric indices to determine if changes in severity of symptoms are significant. Current diagnostic measures lack psychometric change indices for a clinician to utilize to interpret change scores, which limits their use in tracking eating disorder symptoms or determining if individuals are approaching recovery.

In sum, the interview-based, diagnostic measures are inherently limited in assessing change in eating disorder symptoms. Though comprehensive interview-based diagnostic measures assess the spectrum of eating disorder symptoms and can be utilized as a diagnostic criteria-based outcome measure, they cannot be used at each session as they are not designed in this manner due to time and resource limitations. Interview diagnostic measures are able to discriminate well between individuals with and without a diagnosis of an eating disorder and determine outcome categorically, but the presence of qualitative changes in symptoms are not examined (Jarman & Walsh, 1999) and no valid psychometric indices of change are used.

Self-report based diagnostic measures. The need for easily administered assessment of eating disorder diagnoses resulted in the emergence of several self-report eating disorder assessment instruments, such as the Eating Disorder Inventory (EDI; Cumella, 2006), Eating Disorder Evaluation Questionnaire (EDE-Q; Mond, Hay, Rodgers, & Owen, 2006; Peterson, Crosby, Wonderlich, et al., 2007; Sysko et al., 2005), Bulimia Test-Revised (BULIT-R; McCarthy, Simmons, Smith, et al., 2002), Eating Disorder Diagnostic Scale (EDDS; Stice,

Telch, & Rizvi, 2000), Questionnaire for Eating Disorder diagnosis (Q-EDD; Mintz et al., 1997), and Eating Attitude Test 40 (EAT-40; Mintz & O'Halloran, 2000). Most of these instruments have high internal consistencies, and convergent validity with interview format diagnostic measures. Studies of these measures have indicated coefficient alpha reliabilities of .90 for full scales (Bennett, 1997; Clinton & Norring, 1999; Cooley & Toray, 2001; Eberenze & Gleaves, 1994; Peterson, Crosby, Wonderlich, et al., 2007; Reas, Grilo, & Masheb, 2006) and also high construct validity (Espelage, Mazzeo, Aggen, et al., 2003). The Gharderi Survey for Eating Disorders (G-SEDS) also emerged with high positive predictive value with the EDE, concurrent and discriminant validity with the EDI, and high test-retest reliability, meeting the need for an "easily administered and cost-effective instrument for screening and establishing the diagnoses of eating disorders" (Gharderi & Scott, 2002, pg. 61).

The two most commonly used self-report diagnostic measures are the EDI and EDE-Q. The EDI's internal structure consists of eight factors, such as bulimia, drive for thinness, and perfectionism, similar to the factor structure of the EDE (Kordy, Percevic, & Martinovich, 2001). These self-report diagnostic measures were initially criticized for their inability to objectively assess binges, though comparisons of objectively and subjectively assessed binges indicated no significant differences in symptoms of psychopathology (Pratt, Niego, & Agras, 1998). There were no significant differences found between the two types of assessment of binges on measures of BMI, restriction, and psychological functioning, supporting the utility of the self-report format and assessment of subjective binges. The EDI is now in its third revision. The latest version includes 91 items, 12 scales, 6 composite scores, and 3 response style or profile validity indicators (Cumella, 2006). As with the interviewer-based measures, the EDI-3 is designed for use as a measure of diagnostic status. The EDE-Q is the shortened self-report

questionnaire of the EDE, which has concurrent validity with the EDE (Binford, et al., 2005), though showing some discrepant results on factors such as binge eating and shape concerns. Despite these discrepancies between the interview and self-report versions of the EDE, the concurrent and discriminant validities are not greatly affected.

Though several self-report diagnostic measures have high reliabilities and convergent validities, there are several limitations of these measures. For example, measures such as the EAT-40 have been shown to have high false-negative diagnoses for EDNOS (Mintz & O'Halloran, 2000). This category of measures also does not address several aspects of eating disorders included in the EDE, such as dietary restriction, binge eating, and vomiting. The EDI-2 was further shown to have scales that did not contain discriminant validity between BN patients and the general psychiatric population (Schoemaker, Verbraak, Breteler, & Van Der Staak, 2003).

Moreover, the use of self-report diagnostic measures for tracking symptom change and outcome is problematic for various reasons. Mintz et al. (1997) examined existing self-report diagnostic measures. The authors examined two types of *self-report* diagnostic measures—pre-existing inventories and questionnaires designed de novo for specific studies. The use of such measures for tracking of symptoms was criticized for various reasons. First, the examination revealed that many of the self-report diagnostic measures were outdated and based on the Diagnostic and Statistical Manual third revision (DSM-III). Second, the measures often did not capture all of the eating disorders and generally only assessed one eating disorder, such as AN, excluding BN and EDNOS diagnoses. Measures designed to assess BN also identified those with AN and EDNOS diagnoses as having BN, supporting evidence that a dimensional approach to eating disorders is needed as many symptoms overlap between eating disorders (Tylka &

Subich, 1999). Further, Mintz, et al., (1997) found that if a measure did assess more than one eating disorder, the differential diagnoses were occasionally inadequately reliable, such as the EAT-40 had high false positives for AN. Third, the de novo measures did not indicate generalizability across studies as many measures were only used once. The development of many of the operationalized criteria of eating disorder diagnoses in the de novo measures were also not well laid out or explained, contributing to the decreased generalizability. Finally, many de novo measures had arbitrary decision rules used to arrive at diagnoses. These variable decision rules applied to criterion variables such as binge eating, which were not accurate operationalizations of DSM criteria and were inconsistent.

There are several additional methodological flaws when using self-report diagnostic measures for symptom tracking. Logistically, these measures are administered over longer time intervals and not on a session-by-session basis. Though more brief than the interview format diagnostic measures, most self-report measures also still require approximately 20 to 30 minutes to complete, which can vary with greater severity of symptoms. Thus, it is impractical to administer such measures on a session-by-session basis and change cannot be assessed in the midst of treatment. More importantly, like the interview format, self-report diagnostic measures are used for assessing the presence or absence of an eating disorder. The self-report measures also do not provide the sort of specific information about specific dimensions of eating disorder symptomatology at each session, such as the amount of body dissatisfaction an individual experiences or the frequency of body checking, that would be useful for a clinician to know during therapy.

Overall, despite the availability of fairly brief, valid, and reliable self-report diagnostic measures, specific measures of eating disorder symptom tracking are not available. There are

weaknesses with diagnostic instruments, interview and self-report, but they are generally reliable and valid measures of the presence or absence of eating disorders and determining outcome based upon diagnostic status. However, despite which diagnostic measure is used, diagnostic instruments are not designed nor intended to be administered on a session-by-session basis and do not provide indices which could: (a) indicate progress during treatment (or lack thereof), (b) provide detailed feedback about client severity on particular dimensions of eating disorder symptomatology, nor (c) provide information as to when the client is no longer in a clinical range on the measure during treatment.

Unidimensional, domain-specific measures. A number of self-report measures have been developed to assess the presence and severity of specific dimensions of eating disorder pathology. These include measures such as the Beliefs About Appearance Scale (BAAS; Spangler & Stice, 2001); Body Checking Questionnaire (BCQ; Calugi & Grave, 2006), Body Parts Satisfaction Scale (BPSC; Petrie, Tripp, & Harvey, 2002), Body Image Avoidance Questionnaire (BIAQ; Rosen, Srebnik, Saltzberg, & Wendt, 1991), Dieting Beliefs Scale (DBS; Stotland & Zuroff, 1990), Goldfarb Fear of Fat Scale (GFFS; Goldfarb, Dykens, & Gerrand, 1985), Three-Factor Eating Questionnaire (TFEQ; Stunkard & Messick, 1985), Fear of Gaining Weight (FGW; Rushford, 2006), Body Uneasiness Test (BUT; Cuzzolaro, Vetrone, Marano, & Garfinkel, 2006), and Testable Assumptions Questionnaire-Eating Disorders (TAQ-ED; Hinrichsen, Garry, & Waller, 2006). These measures have adequate to good test-retest reliability ranging from .69 to .90 and internal consistencies between .69 and .92.

Self-report measures attempt to assess one or a very few specific aspects of eating disorders singly, such as the single constructs of binge eating, food restriction, body dissatisfaction, body checking, or body avoidance. The BIAQ, for example, attempts to assess

only body image avoidance, or how individuals avoid viewing their own bodies. Rosen et al. (1990) stated that clinical diagnostic interviews such as the EDE, though longer, have superior breadth, depth, objectivity, and favorable psychometric characteristics than unidimensional measures of eating disorders. Comparing the BET, BSQ, and BIAQ, there was only moderate concurrent validity with measures of dietary restraint and overeating from eating records and several behavioral measures. Compared to the EDE, these measures were not indicated to add discriminant validity (Rosen, et al., 1990).

These domain-specific measures are often found to reliably distinguish between controls and eating disorder patients, and many of them are brief enough to be administered on a session-by-session basis. However, few criteria have been developed to establish clinically significant change on such measures, nor do they assess a comprehensive range of eating disorder symptomatology (Ametller, Castro, Serrano, et al., 2005; Benninghove, Jurgens, Mohr, et al., 2006; Rodriquez-Cano & Beato-Fernandex, 2005) and are therefore limited in scope. Adequately assessing change in eating disorder symptomatology would therefore require administering a number of such measures simultaneously, requiring a significant time commitment that would be unwieldy on a session-by-session basis. Lack of psychometric change indices for these measures also results in limited ability to judge whether or not significant change is occurring.

Existing eating disorder treatment tracking measures. There are currently two published measures which are designed to be used on a session-by-session basis to assess eating disorder symptom change in patients, namely the Multiaxial Assessment of Eating Disorders Symptoms (MAEDS; Anderson, Williamson, Duchmann, Gleaves, & Barbin, 1999; Martin,

Williamson, & Thaw, 2000) and Short Evaluation of Eating Disorders (SEED; Bauer, Winn, Schmidt, & Kordy, 2005).

The MAEDS is a 56-item, self-report measure for the assessment of eating disorder symptoms. It assesses six factors associated with eating disorders: binge eating, purgative behavior, avoidance of forbidden foods, restrictive eating, fear of fatness, and depression. The internal consistencies for the MAEDS subscales range between .80 and .92. The MAEDS purports that it is able to assess and screen for the presence and severity of eating disorder symptoms, as well as to be used for differentially diagnosing eating disorders as particular patterns of subscale scores are associated with specific eating disorders. Further, it is used to assess treatment progress (Martin, et al., 2000). The MAEDS is intended to specifically assess eating disorder symptoms, but does not contain items to comprehensively assess the eating disorder spectrum. The MAEDS does not include relevant eating disorder symptoms such as body checking and. Additionally, the authors of the MAEDS acknowledged there are overlaps between factors which may be assessing similar rather than distinct features of eating disorders, such as restrictive eating and avoidance of forbidden foods. The MAEDS also includes factors not directly relevant to eating disorders. For example, while depression is at times comorbid with an eating disorder and is sufficiently common that the inclusion of such a factor may be useful, it is not a recognized feature of eating disorder symptomatology, yet depression is one of the six main factors on the MAEDS and accounts for 11.1% of the variance of the MAEDS (Anderson et al., 2000). Further, the MAEDS does not provide psychometric indices that would indicate clinically significant change nor linear models to aid in projecting recovery.

The SEED was developed as a very brief, monitoring measure of treatment progress (Bauer, et al., 2005) and contains only six items. The six questions assess weight and height, the

degree to which an individual fears becoming fat or gaining weight, how they perceive their body, the frequency of purging behaviors, if amenorrhea is experienced, and ascertaining birth control pill consumption. The SEED was intended to be sensitive to change in eating disorder symptomatology (Bauer et al., 2005), though is designed to assess some symptoms, such as binge eating, over a four week period, rather than weekly. The SEED is also indicated to have weak construct validity as it only correlated between .25 and .40 with the EDI for BN and AN subjects, respectively. Concurrent reliability of clinician and patient ratings ranged from .19 to .78 for items on the SEED. It is also unclear how the questions were developed and the rationale for assessing limited domains was not given. For instance, no information on the amount of restriction is provided from the limited questions, nor areas assessing the degree to which individuals are engaging in body checking behaviors or the extent of binge eating. In general, the SEED does not provide enough specific information about symptoms of eating disorders. Symptom tracking measures have limited utility if specific information on symptomatology is not available and clinicians are not given a comprehensive understanding of the degree to which individuals are experiencing various symptoms and if any change is observed.

The MAEDS and SEED are quickly administered tracking tools that can be used on a session-by-session basis. These instruments are a starting point for tracking symptom change, though they have problematic areas. Both are not adequately comprehensive in their assessment of eating disorder symptoms. One of the most pressing limits, which have not been addressed for tracking tools, is the lack of utilization of psychometric indices of change. These existing tracking measures that have been developed to be given on a session-by-session basis in eating disorders have not provided empirically supported psychometric indices which would signify change and relation to a functional or more functional population. Without psychometric indices

for eating disorder tracking tools, there are no evidence-based options for clinicians to apply in daily practice. An inferential statistic is needed that estimates the magnitude of change in a score necessary for a change score to be considered statistically reliable. Cut points which differentiate membership between two or more populations would also be useful indicators for use in treatment decisions, which led to the development of the Changes in Eating Disorder Symptoms Scale. The internal and temporal reliability of the CHEDS and its construct validity have been recently demonstrated (Spangler, 2010). However, there remains a need to establish psychometric change indices and trajectory analysis for the CHEDS.

Assessment of Symptom Change in Eating Disorders

The need for a psychometrically sound tracking and outcome measurement tool that is both concise enough to be given at each session and comprehensive, assessing the range of eating disorder symptoms and features, is important in assessing the occurrence of change in individuals with eating disorders to inform and aid clinicians in treatment decision making. Brown, Burlingame, Lambert, et al. (2001) state that "outcome measurement involves assessing the clinical outcome of treatment through the use of standardized measures of clinical severity" (p. 925). Further, Brown et al. add that "outcome management is an effort to improve the effectiveness of treatment services…by evaluating outcomes data" (p. 925). This indicates a need for "continuous monitoring of patient progress [which] has been recognized as a core component of evidence based treatment" (Burlingame, Hwang, Lee, et al., in progress, p. 3). Without a validated measure of change in the eating disorder literature, clinicians were unable to reliably assess treatment progress (or lack thereof). Further, tailoring treatment interventions for those without adequate progress or determining decisions concerning termination or proper

treatment setting placement for individuals approaching recovery is currently not based on empirically derived indices for those with eating disorders.

Weekly assessment of patient progress accompanied by the use of clinical significance markers to aid in the calibration of treatment decision processes has been advocated by several researchers (e.g., Burns, 1995; Jacobson & Truax, 1991; Lambert, Okiishi, Finch & Johnson, 1998). A dimensional measure of main features of the entire eating disorder spectrum circumvents the problems of diagnostic categorizing of individuals, such that all persons with eating disorder spectrum disorders can be tracked during treatment and evaluated on severity of major dimensions of eating disorder symptomatology. With specific feedback for clinicians on major dimensions of symptoms of eating disorders, relevant and meaningful information can be used to assess intervention effectiveness, future planning, and current functioning.

The eating disorder literature is mature and advanced with regard to diagnosis and treatment content. The next step in eating disorder assessment is to develop a measure that is comprehensive and brief, while being reliable, valid, and sensitive to change and possessing empirically derived change indices. The Changes in Eating Disorder Symptoms Scale (CHEDS; Spangler, 2010) is a tracking and outcome measure of eating disorder symptomatology with breadth, brevity, and high reliability and construct validity. The content domains for the CHEDS were derived utilizing eating disorder diagnostic measures, DSM-IV diagnostic criteria, existing measures related to eating disorder symptomatology, clinical experience, and the theoretical literature regarding the primary dimensions of eating disorder symptoms. The CHEDS loads significantly onto seven factors, which accounts for 73% of the variance. Only two of the 35 items load onto more than one factor; these two items were retained as they loaded more onto one factor over the other both theoretically and quantitatively. Intercorrelations among the

factors were in the low to moderate ranges (range .09 to .43). The CHEDS also has strong discriminant validity, being able to discriminate between eating disordered and non-eating disordered groups with a sensitivity and specificity of 80%. The CHEDS was compared to several full-scale unidimensional measures which were expected to be differentially related to the CHEDS subscales and demonstrated expected correlational patterns with such unidimensional measures. Although there is strong support for the psychometric properties of the CHEDS, it lacks psychometric change indices such as empirically-derived reliable change indices, cutscores, and hierarchical linear modeling (HLM) benchmarked trajectories to allow for the determination of clinically significant change during treatment and to provide an index of recovery.

Clinical Significance

Utilizing evidence-based change indices, as opposed to qualitative judgments, is vital in assessing change (Wise, 2004). Current instruments assessing eating disorders, whether diagnostic or dimensional, do not provide psychometric indices to define and judge whether reliable and clinically significant change has occurred. Clinical significance refers to meeting the standards of efficacy set by consumers, clinicians, and researchers (Jacobson & Truax, 1991). In some cases, these standards are set as changing one, two, or three standard deviations on a measure (Jacobson, Follette, & Revenstorf, 1984). In recent literature, the criteria which should be used for clinical significance has been actively debated. For example, debates have focused on what the appropriate magnitude of standard deviation a patient has to change on a measure to be considered to show clinically significant change, and on the use of normative cutscores between populations (Wise, 2004; Burlingame, in progress). The importance of determining clinical significance is first defining what it means for a patient to be recovered. This is to say,

how much a patient has to change to be considered as moving from a dysfunctional category to a functional category (Jacobson & Truax, 1991; Wise, 2004).

In the analysis of clinical significance, Jacobson et al. (1984) addressed whether a patient has changed two standard deviations towards the direction of functionality and whether that patient's post-test score falls under a normal population. This suggests that normal distributions of functional and dysfunctional populations are necessary for comparison to ascertain which population a patient's score is indicative of, exhibiting clinically significant change as a patient's post-test score is representative of what would be considered a functional or more functional population (Wise, 2004). Tingey, Lambert, Burlingame, and Hansen (1996a) proposed using multiple samples in a continuum such as functional and dysfunctional. Recent studies have begun to utilize multiple relevant samples, such as functional, less functional, and dysfunctional samples (i.e., community normal, outpatient, and inpatient, respectively; Burlingame, in progress). When there are no normative samples to compare to, changes of standard deviations on a measure noted above from pretest to posttest has been used. This method of assessing clinical significance, though, is problematic as it does not take into consideration measurement error and uses an arbitrary magnitude of clinically significant change (Wise, 2004).

In contrast, a reliable change index (RCI) and normative cutscores are psychometric indices that are useful tools to compare an individual's score on a measure to a respective population. The RCI indicates a statistically significant value that clinicians and administrators can use to make empirically based decisions in their assessments of patient progress. A cutscore is a change index indicating a point in which marks the difference between populations and are based on specific norms that reflect the typical distress experienced across the spectrum of

psychiatric populations. For instance, one would assume increasing levels of symptom distresses as one would move from a functional sample to a less functional, or dysfunctional, sample.

Reliable change index. A RCI is the difference between scores on a measure that a patient must achieve to be considered to have made a statistically significant change. It also takes into consideration measurement error (i.e., the more unreliable the measure, the greater change is required). The formula is as follows:

$$\text{RCI} = (time_1) - (time_2) = S_{diff} \times 1.96(a = .05)$$

$$S_{diff} = \sqrt{2S_E{}^2}$$

$$S_E = SD\sqrt{1 - r_{xx}}$$

Alpha levels for the RCI can be tailored to the specific aims of a study or practice. An alpha level of .05 provides a strict measure of change as the RCI will be a greater value than with a .10 alpha level; subsequently, a patient measured using a strict alpha level will have to show greater amounts of change or a greater difference score between scores.

Lunnen and Ogles (1998) found the RCI was able to distinguish individuals who exhibited a positive reliable change but found that the RCI was less effective in differentiating those who showed no change or deteriorated, or had a negative reliable change. The RCI has also been criticized by some as not accounting for regression to the mean, or the "phenomenon observed that if a specific variable is to be predicted from another variable, each of the values pertaining to the variable which it is predicting (having some distance from its mean), corresponds with a less extreme (i.e., closer to its mean) predicted value" (Hageman & Arrindell, 1993, pg. 695). Though when an RCI is calculated taking into account measurement error, it provides a more accurate index than arbitrarily set statistical significant measures. When the

RCI is calculated with a measure of reliability (r_{xx}), it makes the RCI more stringent. The more unreliable the assessment, the greater RCI is needed for a subject to exhibit a reliable change.

Cutscores. A normative cutscore is a statistically significant cut off value between two populations. As previously mentioned, relevant normative populations are needed to identify how much change a patient must exhibit to be considered to have reliably changed on a measure while also moving from a dysfunctional to a functional population. Cutscores are calculated using the following formula:

$$cutscore = \frac{(SD_1)(mean_2) + (SD_2)(mean_1)}{SD_1 + SD_2}$$

SD= standard deviation

The critics of clinical significance indices emphasize that for a patient to exhibit true change, they must have both a significant difference between their scores and also change from a dysfunctional to functional population (Wise, 2004). In response, an RCI and cutscores must be used in conjunction to identify patients who are considered recovered and also to identify relevant comparative normative populations. These methods have yet to be applied to eating disorder measures, though would add significantly to the utility of measures in assessing client change.

Eating disorder clinical significance testing. A relatively recent meta-analysis assessed clinically significant change in eating disorder symptoms in various studies of cognitive-behavioral therapy (CBT) efficacy, utilizing the RCI and equivalency testing for group level analysis. The study indicated that CBT for eating disorder yields clinically significant change as assessed by the EDE (Lundgren, Dannoff-Burg, & Anderson, 2004). This analysis of clinical significance used a diagnostic measure on an aggregate level, providing support for the efficacies of treatment on the group level by comparing the means of diagnostic criteria items (e.g.,

frequency of binge eating) among various studies at pre and post intervention. Though this study provided relevant feedback on the aggregate level, outcome was defined using diagnostic criteria comparing group means, which does not provide utility in assessing individual change or the ability track change in patients during treatment. The meta-analysis did not utilize an RCI using a comprehensive assessment of eating disorder symptoms as reliable change was only assessed on binge and purge frequencies in a given week. Clinically significant change as defined by Lundgren, et al. (2004) was primarily based on the dichotomous method of meeting an eating disorder diagnosis, despite a reliable change on specific dimensions of binge eating and purging frequencies. There was also no evidence that clients were more representative of a functional population by utilizing normative cutscores on the dimensional measures, though diagnostic criteria for the EDE was used.

The EDE-Q was examined in a study by Sysko et al. (2005) as a diagnostic outcome measure. This outcome, though, was only defined as achieving concurrent validities of pre and post treatment assessment of diagnosis compared to the EDE. Though the use of a pre and post treatment diagnosis using the self-report diagnostic measures can show change in an individual, it goes beyond its intended measure, is not time-sensitive, and does not indicate the amount of change needed from pre and post scores to be considered clinically significantly changed. There have also been uses of statistically significant change in studies of eating disorders, such as changes of standard deviation on an eating disorder measure or effect sizes (Safer, Agras, Lowe, & Bryson, 2004). Criteria of eating disorder outcome using statistical significance are problematic as they do not indicate if an individual actually experiences clinically significant change. Utilizing solely a statistically significant change criterion does not give enough information supporting change, as it is an arbitrary statistical criterion based on a percentage of

change (e.g., standard deviation). This method does not account for extraneous variables such as error in the measure. Using statistical criteria for clinically significant change does not account for low reliability of a measure, or the degree of confidence of the probability of clinically significant change (such as a reliability of .95 or .90).

Change trajectory analyses. There are some proposed predictors of change in eating disorder treatment using various measures in the eating disorder research (Satandar-Pinnock, Woodside, Carter, et al., 2003; Miller, Schmidt, Vaillancourt, et al., 2006; Peake, Limbert, & Whitehead, 2005; Wonderlich, Crosby, Joiner, et al., 2005), but the only consistently empirically supported predictor of outcome is initial severity of symptoms and early changes in binge eating and purging symptoms (Halmi, Agras, Crow, et al., 2005; Fairburn, Agras, Walsh, et al., 2004; Agras, Crow, Halmi, et al. 2000). With the various assessments of predictors of change providing a breadth of knowledge concerning initial severity scores, there has been no analysis of trajectory of clinically significant change using a repeated, comprehensive, and psychometrically sound eating disorder tracking and outcome measure.

Hierarchical linear modeling (HLM) is a method to classify or arrange groups that share the same qualities, and the relationship in each level can be studied. This method allows for trajectories of groups of people to be projected with respect to outcome. A linear model that accounts for variations in each level of initial severity scores on an outcome measure can be developed. This is possible in using a longitudinal analysis as repeated measurements of outcome can be nested within levels of initial scores (Laurenceau, Hayes & Feldman, 2007; Singer & Willett, 2003; Weinfurt, 2000; Harrison & Raudenbush, 2006; Lambert et al., 1998; Wells, Burlingame, Lambert, et al., 1996). Trajectories of individual client progress can be developed to determine if patients are exhibiting on-track improvements or off-track

deteriorations as clinicians utilizing the aforementioned clinically significance criteria can also use these trajectories. Amalgamating the RCI, cutscores, and HLM trajectories will allow for empirically based decisions of treatment placement, intervention strategies, and monitoring of week-to-week progress.

Quality of life assessment. Clinically significant change has also been analyzed through the assessment of quality of life assessments. Health surveys have been used as indications of disorders having adversely and significantly interfering or impairing impact with overall life functioning. While broad measures exist, disorder-specific measures are necessary as general health related quality of life measures do not capture the intricacies of specific disorders (Engel, Wittrock, Crosby, et al., 2006). The Eating Disorders Quality of Life (EDQOL) scale was developed by Engel et al. (2006) and was tested and validated against the Structured Clinical Interview of DSM. Compared to other health related quality of life questionnaires, it was found to be more sensitive to change than general quality of life measures. The EDQOL, though, was not compared with the EDE or any other specific measures of eating disorders.

The Centre for Research on Eating Disorders at Oxford (CREDO) recently developed the Clinical Impairment Assessment (CIA) as an assessment of the effect of having an eating disorder on the level of life functioning and quality of life (Bohn, Doll, Cooper, et al., 2008; Bohn & Fairburn, 2008). The CIA is designed for use in clinical subjects only, as it assesses how eating disorder symptoms are functionally impairing in major life domains such as relationships, work, etc. Its goal is to provide "a simple single index of the severity of psychosocial impairment secondary to eating disorder features" (Bohn & Fairburn, 2008, appendix iii). These life functioning measures have utility in providing convergent data of clinical significance, but also can be used to measure eating disorder specific changes in

functioning that would be able to indicate progress. The CIA can be used as a valid measure of change in life functioning associated with an eating disorder with a test-retest reliability of .86 (95% CI .75–.92; $p < .001$) and a construct validity with the EDE-Q of .89 ($p < .001$; Bohn & Fairburn, 2008).

Eating Disorder Tracking and Outcome Measure Status

The current use of diagnostic criteria for determining outcome does not provide enough information on specific symptoms. This definition of outcome is not based on symptom reduction but a categorical determination of meeting a diagnosis. The spectrum of eating disorders is wide, and individuals can vary greatly in severity of pathology and in presentation. Defining recovery should not only assess the key aspects of eating disorders but also track the changes in symptoms. The subscale structure of a comprehensive measure should assess the primary dimensions of eating disorder symptoms, operationalized from diagnostic criteria as well as key features indicated in the literature. These aspects include assessing behavioral components such as binge eating and purging, as well as psychological domains such as fear of becoming fat, body dissatisfaction, and preoccupations with food.

It is also preferable to use empirically calibrated methods to assess progress itself in individuals; without empirically supported psychometric indices, tracking of symptoms becomes arbitrary. Assessing clinically significant change on an individual basis is needed to determine progress or lack thereof. Being able to develop trajectories to assess if clients are on or off the expected trajectory in regards to treatment progress is also useful, but this methodology has yet to be applied to eating disorder assessment. An eating disorder specific, comprehensive measure with these indices has not been developed up to this point. Providing change metrics to assess

clinically significant change while assessing trajectories of progress would provide clinicians with useful and relevant information to guide treatment decisions.

Specific Aims

This study utilized two datasets: (a) an archival data set of community non-eating disordered participants, and (b) a data set of clinically diagnosed eating disordered patients receiving treatment. These two datasets were used to determine change indices of clinical significance for the CHEDS.

The goal of the study was to develop clinical significance criteria for interpreting eating disorder symptom change during treatment and determining progress trajectories. An RCI for CHEDS total and subscales scores was calculated as an indication of an amount of change needed to be considered reliably changed, while cutscores were used to distinguish between disordered and non-disordered populations. The trajectory analysis using HLM provide expected trajectories of change for patients based on initial severity scores in order to tailor treatment as necessary or to assess projected linearity to clinically significant change (Harrison & Raudenbush, 2006). These psychometric indices were not only calculated on the total score level but also the subscale (or subscale) level.

There were five specific aims of the study:

1. Establish normative data for functional and dysfunctional samples for the CHEDS utilizing the Tingey, Lambert, Burlingame, and Hansen (1996a) method. Compare the cutscore derived to the existing cutscore derived by Spangler (2010) using ROC analysis. .
2. Determine a reliable change index to indicate the amount of change necessary for a subject to be considered reliably changed for total and subscale scores.

3. Create a series of expected trajectories of patient progress based on initial level of severity on the total score and subscale level, providing empirically based guidelines of a subject's path to clinically significant change, and a separate trajectory analysis based on initial scores.
4. Determine the relationship between the CIA, a measure of impairment of life functioning, and the CHEDS to corroborate clinically significant change. It was hypothesized that individuals exhibiting clinically significant change on the CHEDS would have greater reductions in life impairment as measured by the CIA.
5. Categorize patient outcome at the end of treatment into level of clinically significantly change in four categories: recovered (clinically significant change), reliable improvement, no change, and deterioration.

It was hypothesized that the clinical sample at intake would be significantly higher than the non-clinical sample on all subscales of the CHEDS. It was also hypothesized that using the clinically significant change criteria using the RCI and cutscores in the clinical sample would result in significant differences between pre and post treatment scores on the CHEDS, with more patients having clinically significant change and reliable improvement than exhibiting no change or deterioration. It was predicted that very few clinical subjects' initial CHEDS scores would fall in the non-clinical range as all subjects were diagnosed with an eating disorder using the EDE. It was also expected that the clinical subjects' post-treatment CHEDS score would exhibit clinically significant change after completion of treatment. It was also anticipated that a proportion of subjects exhibiting no change or deterioration might also be observed. The subjects' change scores from pre and post-treatment on the CIA were expected to correlate with scores indicating clinically significant change on the CHEDS.

Methods

Participants

Participants were drawn from two populations: (a) a community non-eating disordered group and, (b) an eating disordered group undergoing cognitive-behavioral treatment for their eating disorder. A power analysis for the clinical sample RCI analysis revealed that for a moderate effect size (d = .50), a sample size of approximately 50 would be needed for the clinical or dysfunctional sample, which resulted in a power of .86 at the .05 significance level. The unequal sample sizes of the clinical and non-clinical samples, 58 and 95, revealed a power size of .95, also at the .05 significance level, with an effect size estimate of .50. This indicates that the sample sizes were large enough to ensure adequate power and to be able to compare the two samples. The power analysis supported that a test of the distinctiveness of the two samples is possible given the sample sizes. These power indices indicate that with the two sample sizes, the probabilities of rejecting the null hypotheses, if warranted, are high.

The non-eating disordered sample consisted of 95 participants. Participants were recruited from universities in Colorado and Utah. Participants were recruited through informational flyers distributed in their general education courses and voluntarily contacted experimenters themselves. They scheduled times to meet with experimenters where they consented to participate and were paid $15.00 upon completion of the questionnaires. The control sample was composed of participants who were 80% Caucasian, 9% Asian, 5% Hispanic, 1% Native American, and 5% who described their ethnicity as “other.” Age ranged from 18 to 46 (M = 22.4, SD = 4.6). The non-clinical subjects signed an informed consent for participation in the study and could receive extra credit if they were recruited from an undergraduate course.

The eating disordered sample was recruited through fliers, newspaper ads, and referrals to an outpatient mental health clinic. The *Eating Disorder Examination* (EDE; Cooper & Fairburn, 1987; Fairburn & Cooper, 1993) was used to diagnose clients for admission into the study. Participants were excluded from the study if they had comorbid psychosis, bipolar disorder, a medical condition that significantly impacts weight (e.g., thyroid conditions), or if they had a history of bariatric surgery. Participants were not allowed to participate in any other psychotherapeutic treatment, although concurrent medication use was allowed. All participants received therapy at no-cost. Fifty-eight eating disordered patients were recruited. The majority of the participants were female (93%) and Caucasian (98%) with 2% reporting a Hispanic ethnicity. Age ranged from 18 to 65 (M = 25.24, SD = 9.97). The diagnoses of the clinical sample primarily consisted of BN and EDNOS, consistent with the literature and prevalence rates of eating disorders (APA, 2000), and included six subjects with a diagnosis of AN. The patients signed an informed consent for participation in the study, videotaping for supervision use, and completion of the measures.

Occasionally items were missing from total or subscale scores due to participants not completing items. Participants were retained in analyses for the total score analyses if at least 90% of items were completed. For subscale scores, subjects were dropped from the subscale level analyses if any items were missing. For total scores, five subjects were not used in the analyses. On the subscale level, three, five, five, four, five, two, and six subjects dropped, respectively, for the seven subscale analyses.

Treatment

The therapy consisted of cognitive-behavior therapy (CBT) according to the treatment protocol described by Fairburn (2008). Treatment was of a 40-session duration for patients

considered underweight (BMI < 17.5), and 20 sessions for normal and above normal weight (approximating anorexia nervosa versus bulimia nervosa and EDNOS). The therapists in the study were Ph.D. candidates in clinical psychology. The therapists received training prior to beginning treatment with study patients, including seeing pilot patients until competency in delivering the treatment protocol was reached. Therapists also received weekly supervision with videotaped session review and completed fidelity checks with the treatment protocol at each session. Each session in the treatment protocol had a specified agenda with assessments and clinical forms to be used. At session 7, the subject's progress was examined and barriers to change were identified to tailor their treatment plan accordingly. While the treatment protocol was manualized with session-by-session agendas and goals, crucial elements of CBT such as tailoring for specificity of patients' symptoms and flexibility to adapt to barriers to change were included.

Measures

Before each session, the clinical sample was administered all measures by clinic receptionists. The patients arrived early before each therapy session and completed the appropriate measures in the lobby area and then returned them to the receptionists. The results of these measures were not utilized by the therapists and were not used to guide and direct treatment; the results were also not shared with the clients. Only members of the research team were permitted access to the results of these measures.

Eating disorder symptoms. The CHEDS was administered at each session for the clinical group and at a single time point for the non-clinical group. As described in Spangler (2010), CHEDS items were generated using several methods and then analyzed with respect to item discrimination (Wilks' Lambda), item reliability, and endorsement patterns. The CHEDS

consists of 35 items indicated to have high discriminant ability between eating disordered and non-eating disordered groups. The CHEDS scale is composed of seven subscales. The subscales of the CHEDS are: eating concerns/preoccupation, restriction, body preoccupation, body dissatisfaction, body checking, vomiting, and binge eating (Spangler, 2010). The seven subscales accounted for 72% of the variance. The reliability coefficients of the subscales range from .85 to .93, with the exception of one subscale at .73, while the overall internal reliability coefficient alpha was .96. The subscale with the lowest internal reliability, vomiting, only contained two items.

The construct validity of the CHEDS was also confirmed as the subscales correlated in expected patterns with other measures. Means comparisons between non-eating disordered and eating disordered groups also significantly differed in expected patterns (Spangler, 2010). CHEDS items are also sensitive to change as the items change in the theoretically proposed direction during treatment, which is indicated as a useful analysis in measure development (Vermeersch, et al., 2000; Burlingame, Seaman, Johnson, et al., 2006). CHEDS scores have also been shown to be significantly higher in eating disordered versus non-ED groups (Spangler, 2010), signifying that the clinical and community sample groups are significantly different. The ROC analysis yielded a cutscore of 60.

Eating disorder diagnosis. The eating disordered sample was diagnosed using the Eating Disorder Examination, an interviewer-based diagnostic interview (Cooper & Fairburn, 1987). The interrater reliability coefficients for the five subscales of the EDE range from .83 to .99 (Rosen et al., 1990). The internal consistency coefficients ranged from .67 to .90 and the discriminant and concurrent validities are high (Guest, 2000). The control sample was diagnosed using the Eating Disorder Diagnostic Scale (EDDS; Stice, Telch, & Rizvi, 2000), a self-report

diagnostic questionnaire. The sensitivity and specificity of the EDDS are .88 and .98, respectively (Stice, Fisher, & Martinez, 2004). The internal consistency of the EDDS yields a Cronbach's alpha of .89.

The eating disordered patients were assessed with the EDE at two time points, treatment intake and post-treatment. The non-clinical participants completed the EDDS once; at the same time they completed the CHEDS. Undergraduate-level research assistants who were trained by a licensed PhD clinical psychologist administered the EDE interviews. Fidelity checks of the EDE interviews were achieved through audiotape feedback of interview sessions.

Quality of life. The Clinical Impairment Assessment (CIA) was used to assess the extent to which eating disorder symptoms affect the clients' level of life functioning (Bohn & Fairburn, 2008) and was administered to the clinical sample only. The CIA was used as a supplemental indication of clinical significance by assessing level of functioning and obtaining concurrent validity of clinically significant change on the CHEDS from pre-treatment to post-treatment. The CIA was administered at intake, mid-treatment (session 7), and post-treatment. The psychometric properties of the CIA are strong, with test-retest reliability of .86 and a concurrent validity with the EDE-Q of .89 (Bohn, et al., 2008).

Data Analyses

Reliable change index. Reliable change index (RCI) values were derived for the CHEDS on the total and subscale score levels. The original formula proposed by Jacobson, Follette, and Revenstorf (1984) has been criticized as only taking into consideration inherent measurement error, not a true pre-test score (Wise, 2005). Jacobson and Truax (1991) and Tingey et al. (1996a, 1996b) proposed use of S_{diff} for the measurement of standard error (SE) to

reflect the amount of difference one could expect between scores (Wise, 2004). This would make the RCI more accurate in assessing between repeated scores. The formula is as follows:

$$\text{RCI} = (time_1) - (time_2) = S_{diff} \times 1.96(a = .05)$$

$$S_{diff} = \sqrt{2S_E{}^2}$$

$$S_E = SD\sqrt{1 - r_{xx}}$$

An RCI was derived for various alpha levels for the total CHEDS score, as well as for the seven subscales of the CHEDS. In this study, the standard error of measure (S_E) was computed using the internal reliability coefficient of the CHEDS. The S_E for the total score was computed using the clinical and non-clinical samples and a pooled standard deviation value (*SD*). The resulting S_E value was inserted into the standard error of difference formula (S_{diff}). This value was multiplied by the z-value of the significance level desired, such as 1.96 ($p < 0.05$). The resulting value represented the size of the difference needed to achieve reliably significant change given the error of the instrument and the standard deviations of the eating disordered and non eating-disordered samples. Change indices for the seven subscales of the CHEDS were also computed to enable evaluation of change on a subscale level. The CHEDS-RCI scores were also calculated with different alpha levels ($p < .05$ & .10) for total and subscale scores to provide for a range of confidence intervals.

Cutscores. A cutscore refers to a statistically derived point that divides the scores of two groups that have been created in reference to some criterion. It is a cutoff between adjacent samples which defines the point where it is statistically more likely for a score to be in one population as opposed to the adjacent overlapping distribution (Tingey, Lambert, Burlingame, & Hansen, 1996a; Jacobson & Truax, 1991). In the context of this study, the CHEDS is a measure designed for a clinical population and as such, the cutscore is the cutoff value on the CHEDS of

an individual suffering from an eating disorder that moves from a less functional to more functional. When representative scores for each population have been formed and there is a statistically and clinically meaningful difference between the populations, a cutscore can be established that separates the populations from one another (See Figure 1).

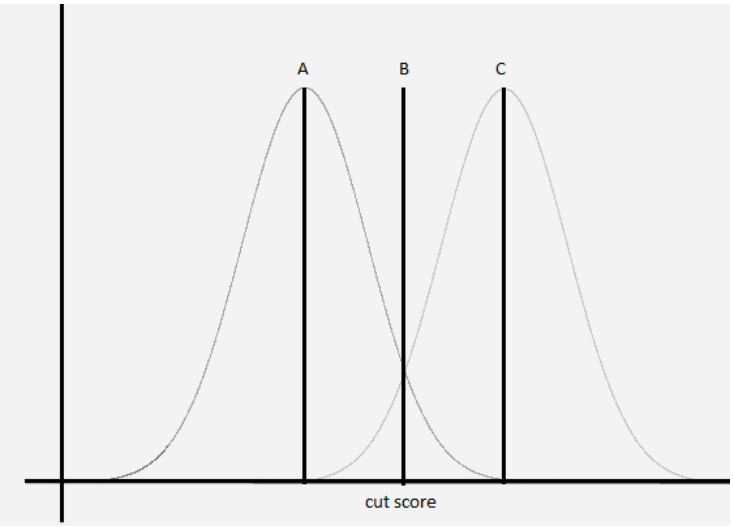

Figure 1. Normal Sample Distribution (example). Illustration of two sample distributions with means (A & C) and cutscore (B).

This study utilized two samples that are close on the continuum of dysfunction. The non-eating disordered community sample and the outpatient sample utilized in this study are closer in comparison than a community normal and an inpatient sample, and thus are expected to be closer bimodal distributions. Using samples that are close in level of dysfunction allows a clear explication of the differences between these two populations. It was hypothesized that there would be overlap between the samples used in this study but a statistically significant difference indicated by the normative data analysis aforementioned. A psychometric calculation used by Tingey, Lambert, Burlingame, and Hansen (1996a) was employed to determine the clinical significance cutscore for the CHEDS. The two sample distributions were used to establish the

cutscores. Cutscores were calculated between non-clinical and clinical samples. The following formula was used to calculate the normative cutscores:

$$cutscore = \frac{(SD_1)(mean_2) + (SD_2)(mean_1)}{SD_1 + SD_2}$$

Spangler (2010) determined a cutscore on the total CHEDS scale score using a different method (i.e., ROC analysis) which indicated a score of 60 to be "the best balance of specificity (i.e., percent of those in the ED group who are correctly classified) and sensitivity (i.e., percent of those in the non-ED group who are correctly classified) … yielding a specificity and sensitivity of 80% and an AUC of .86" (pg. 136). The ROC cutscore obtained by Spangler (2010) was compared to the cutscore obtained in this study using an alternate method.

Clinically significant change classification. The RCI and cutscores were applied to the clinical sample resulting in four categories of change: recovery, reliable improvement, no change, and deterioration. Recovery is defined as meeting the RCI requirement and crossing the cutscore into the community normal sample. Reliable improvement is defined as those who have met the RCI magnitude but did not cross the cutscore. This level of change may contain a small percentage of individuals who may have fallen under the cutscore at admission. No change is defined as not meeting the RCI magnitude of change nor crossing the cutscore. Deterioration is defined as those who met the RCI level in the negative direction. A chi-square analysis of the frequency of patients classified into these four categories was calculated for CHEDS total scores.

Hierarchical linear modeling. Hierarchical linear modeling (HLM), a multilevel analysis (Weinfurt, 2000; Harrison & Raudenbush, 2006), was used to assess trajectories based on the initial level of disturbance for the clinical sample. The main rationale of HLM is to assess how people change over time and how it is related to other variables. The multilevel analysis allows variance in outcome variables to be analyzed at multiple hierarchical levels using

repeated measures data and within-subjects variables. Hierarchical linear modeling is advantageous as it is a type of mixed model analysis used with hierarchical data to view inter-individual variability of change, examining predictors or covariates of interest that will affect the trajectories, such as initial severity of scores.

The HLM analyses were used to create individual growth trajectories using maximum likelihood estimation on the CHEDS using an unstructured model. All session-by-session CHEDS data points available were used in the clinical sample to develop the models. Tukey's Ladder of Transformation was used to determine if any time variable transformations of the data using unconditional growth models was needed. The initial models were decided using theory to examine the fit statistics (Singer & Willet, 2003). Two log-likelihood (2LL) was examined as a measure of deviance with each time variable transformation. Essentially, models were created with variable time transformations, such as natural log, square root, squared, and cubed, to ascertain which model better matches the shape of the projected paths.

After the shape of the projected paths was ascertained with the least deviance, change trajectories were generated based on initial scores which would indicate a patient's projected path with clinically significant change criteria mapped onto the trajectories. A clinician would be able to utilize these trajectories to create idiosyncratic paths based on initial CHEDS scores or strata level. Two main analyses were generated for the trajectories: a path based on initial scores which a patient followed to achieve clinically significant change, and separate paths according to classifications of high, moderate, and low severity scores based on divisions of three ranges of initial CHEDS scores (Lambert, Okiishi, Finch, & Johnson, 1998, Wells, Burlingame, Lambert, et al., 1996). These analyses were done on the total score levels as well as on the subscale levels.

The HLM analysis was used to create score bands by stratifying the data according to initial severity and creating separate models per stratum.

The trajectory models were built starting with all theorized covariates. The main effects of the intercept, time (sessions), diagnosis (20 session vs. 40 session), and severity strata or initial score were used in the model as well and interactions between these main effects (e.g., session x diagnosis and session x severity strata). The models were built centered on the middle severity for the strata HLM model and the initial score HLM model was centered on the CHEDS mean. Each model underwent stepwise deletion of non-significant parameters starting with deleting the worst significance values (*p*-values over .05) and then non-significant interactions. The final models for each analysis then comprised of parameters that yielded all significant values.

Based on the client's initial level of disturbance, RCI and cutscore levels were projected along the HLM trajectory graph so that a clinician could see at any point how a case compares to the RCI and cutoff points and the various outcome classes (e.g., clinically significant change, reliable improvement, no change, & deterioration). The clinical significance change metrics were used along with the upper and lower levels of the trajectories (using a .05 alpha level). The upper and lower tolerance levels are used like confidence intervals in which a trajectory is likely to follow. All HLM calculations were computed using the Statistical Package for Social Sciences (SPSS).

Correlation with life functioning. The CIA was used as a convergent validity measure of change in life functioning. Correlations were calculated between total scores between the CHEDS and CIA at pre, mid, and post treatment which provided additional information of

clinically significant change. High correlations supported that the CHEDS is related to life functioning and impairment.

Results

Distribution Analysis

Tests of normal distribution of all data points of the eating disordered sample revealed a normal distribution using a normality plots test, indicating a Shaprio-Wilk statistic value of larger than .8 (.99, $df = 982$, $p < .01$).

Descriptive Statistics

Descriptive statistics (i.e., sample size, range, min, max, mean, and standard deviation) were compiled for the two samples for CHEDS total and subscale scores (body preoccupation, body dissatisfaction, body checking, binge eating, restrictive eating, food preoccupation, & vomit; see Table 1). Mean comparisons indicated that the eating disordered sample was significantly higher on the total CHEDS score (t (146)= -14.70, p <.01) and all subscales scores than non-eating disordered groups consistent with previous findings (Spangler, 2010). Chi-squared analysis also revealed that the CHEDS is able to discriminate between eating disordered and non-eating disordered groups (χ^2 (1) = 69.08, $p < .01$).

Table 1

Community Normal and Clinical Sample CHEDS Total and Subscale Means

	Community Normal			Clinical			
Subscale	n	M	SD	n	M	SD	*t*
Body Preoccupation	89	9.39	5.74	55	20.35	5.05	12.13*
Body Dissatisfaction	91	9.55	5.26	53	16.74	2.83	9.28*
Body Checking	93	6.83	4.61	53	12.64	4.18	7.52*
Binge Eating	92	3.36	2.88	54	13.26	6.43	13.00*
Restriction	92	3.23	2.61	53	8.58	4.36	9.34*
Food Preoccupation	94	3.74	3.81	56	14.07	4.13	15.65*
Vomit	94	0.41	0.84	52	2.88	2.37	9.20*
Total	95	37.66	21.48	53	91.02	20.60	14.70*

Note. Clinical scores at session 0. * $p < .01$.

The CHEDS total score minimums for community normal sample was 4 while the clinical sample minimum was 38. The maximum score observed for the community normal and clinical samples were 106 and 139, respectively. This indicated that those in the community normal sample who did not qualify for a diagnosis of an eating disorder could still display a high score on the CHEDS. Total score means observed for the two samples were about two pooled standard deviations ($SD_p = 21.17$) apart at 37.66 and 91.02, for the community normal and clinical samples, respectively. Subscale level descriptive statistics also yielded proportionately similar means and standard deviations.

Cutscores

Cutscores were calculated between the two samples on both total and subscale score levels (Table 2). The total cutscore indicates a CHEDS score that is the cut off between the

community normal and clinical samples. Comparisons of the total and subscale cutscores and descriptive statistics of the two samples showed that the cutscores were between the respective means.

Table 2

Total and Subscale CHEDS Cutscores

Subscale	SD_1	SD_2	M_1	M_2	Cutscore	% of Patients Below Cutscore*
Body Preoccupation	5.05	5.74	20.35	9.39	15.22	18%
Body Dissatisfaction	2.83	5.26	16.74	9.55	14.22	21%
Body Checking	4.18	4.61	12.64	6.83	9.87	23%
Binge Eating	6.43	2.88	13.26	3.36	6.41	19%
Restriction	4.36	2.61	8.58	3.23	5.23	28%
Food Preoccupation	4.13	3.81	14.07	3.74	8.69	13%
Vomit	2.37	0.85	2.88	0.41	1.07	33%
Total	20.60	21.48	91.02	37.66	64.89	9%

Note. SD1 and SD2 are clinical and community normal standard deviations, respectively.
* at session 0 for clinical sample only

The total cutscore created utilizes sample sizes, standard deviations, and means of each sample to derive the value whereas a ROC analysis assesses true positive rates and false positive rates, generating a specificity and sensitivity rate. The ROC analysis conducted by Spangler (2010) identified a cutscore of 60, compared to the cutscore of about 65 in this study.

Reliable Change Indices

The internal consistency coefficient was used for the respective total and subscale RCI calculations (see Table 3). The total CHEDS internal consistency reliability coefficient is .96

(Spangler, 2010), which is used for the standard error of difference calculation used in the RCI. A lower reliability would implicate that a greater improvement would be required to signify reliable change. The calculation of the RCI utilizes the pooled standard deviation and the respective pooled standard deviations were used for the total and subscale score RCI calculations. The RCI was derived using all data points from the community normal sample and the clinical sample at baseline. The RCI for total and subscale scores are indicated in Table 3 with respective alpha levels. Table 3 also displays the sample sizes for respective groups as well as the standard deviations and pooled standard deviations used to calculate the RCI's.

Table 3

Reliable Change Indices

	SD			RCI (alpha level)	
Subscale Description	Clinical	Community	Pooled	.05	.10
Body Preoccupation	5.05	5.74	5.49	4.02	3.38
Body Dissatisfaction	2.83	5.26	4.53	3.32	2.79
Body Checking	4.18	4.61	4.46	3.91	3.28
Binge Eating	6.43	2.88	4.54	3.98	3.34
Restriction	4.36	2.61	3.35	3.60	3.02
Food Preoccupation	4.13	3.81	3.93	3.08	2.59
Vomit	2.37	0.85	1.56	2.25	1.89
Total	20.60	21.48	21.17	11.74	9.85

Hierarchical Linear Modeling

Hierarchical linear modeling analyses were completed for the clinical sample utilizing all data points. Time variable transformations using Tukey's Ladder of Transformations indicated that the regular time intervals yielded the lowest two log-likelihood (2 LL) statistics, which is a maximum likelihood estimate measuring goodness of fit (deviance of the model of observed and expected values). Time variable transformations of squared and square root were conducted to ascertain the best shape of the models (see Table 4). These transformations yielded more deviance and no further time transformations were needed as the trends of 2 LL indicated values in the diminishing direction (greater deviance).

Table 4

Tukey's Ladder of Transformations

Time Transformation	2 Log Likelihood
Session - Square Root	8049.99
Session (time)*	7959.14
Session - Squared	8060.95

Note. *Original time variable

Two models were generated, a model based on severity strata and a model based on initial scores. The original HLM models included all theorized variables of the intercept, severity strata level or initial score, diagnosis, session and strata severity or initial score interaction, and session and diagnosis interaction. A centered method was utilized in order to produce a more intuitive model, using middle severity and average initial scores. The stepwise-deletion method was used to eliminate variables which yielded significance levels above $p = .05$.

The initial two models exhibited fairly significant results with the severity strata model yielding all significant variables and interactions and the initial score model containing an insignificant variable of diagnosis (p = .17). Deleting the diagnosis variable in the initial score model and retaining variable interactions with diagnosis, the model consisted of all significant variables. The two final models are displayed in Tables 5 and 6, including the estimates of fixed effects values and significance levels. The initial score model produced a stronger model based on the 2-LL estimates (see Tables 5 and 6).

Results from the HLM analyses indicated that if using the severity strata trajectories, a patient would be expected to improve -1.82 points (decrease in score) per session (see Table 5). The diagnosis variable differentiates between the 20 and 40 session subjects. The 40-session patients, who were in the underweight BMI range, displayed a lower mean and the intersect value was consequently modified for these patients by almost 15 points. These non-underweight patients also exhibited a lower rate of change per session as the session by diagnosis interaction modified the per session change by about two points, which resulted in a session change rate of .40 points. Lastly, the session by severity interaction further modified the intersect point; in specific, for higher strata nearly one point per session more change was observed.

Table 5

Severity Strata Hierarchical Linear Model

Estimates of Fixed Effects							
						Confidence Interval	
Parameter	Estimate	SE	*df*	*t*	*p*	Lower	Upper
Intercept	69.80	2.98	58.29	23.42	.01	63.84	75.77
Severity	22.37	2.31	58.59	9.69	.01	17.75	26.98
Session	-1.82	.33	41.11	-5.53	.01	-2.48	-1.16
Diagnosis	-14.89	6.26	55.01	-2.38	.02	-27.44	-2.36
Session*Severity	-.83	.26	40.46	-3.25	.01	-1.34	-.31
Session*Diagnosis	2.22	.65	35.17	3.39	.01	.89	3.55

Note. 2 Log Likelihood = 7893.61; Confidence interval using 95%.

The initial score HLM model indicated a smaller 2-log likelihood variance rate. The initial score HLM model, as previously noted, dropped the diagnosis variable as it was not shown to be significant (Table 6). With this variable dropped, the session change observed is projected to be about -2.50. The rate of change was also affected by the level of initial scores, with higher scores above the mean improving at a higher rate of -.03 and the converse with lower scores. The diagnosis by session interaction was retained, which also indicated a decrease in change rate for 40-session subjects by 1.83 points per session, bringing the change rate to about -.8 points per session.

Table 6

Initial Score Hierarchical Linear Model

Estimates of Fixed Effects							
						Confidence Interval	
Parameter	Estimate	SE	*df*	*t*	*p*	Lower	Upper
Intercept	87.29	1.75	57.24	49.87	.01	83.79	90.79
initial	1.01	.09	57.43	11.62	.01	.83	1.18
session	-2.51	.24	43.17	-10.63	.01	-2.99	-2.04
session*initial	-.03	.01	41.81	-3.06	.01	-.06	-.01
session*diagnosis	1.83	.67	36.49	2.75	.01	.48	3.19

Note. 2 Log Likelihood = 7882.88; Confidence interval using 95%.

A macro scoring program was created to aid clinicians in the use and interpretation of CHEDS scores. As the initial score model was the stronger model, the scoring program was based upon this model. The CHEDS scoring program derives the total scores, subscale scores for all seven subscales, the reliable change needed, respective cutscores, and the expected trajectories on the total and subscale score levels with separate trajectories, each with its own chart and graph. Upper and lower tolerance levels are also calculated using the respective RCI's.

CHEDS and Clinical Impairment Assessment Comparison

The correlation between the CHEDS and CIA, an assessment of life functioning for those suffering from an eating disorder, was calculated. The CIA was administered to clinical subjects at pre-treatment, session seven, and post treatment and the correlations with the CHEDS used those time points. The clinical sample CIA mean at pre-treatment was 38.40 and decreased to 30 at session seven. At post-treatment, the clinical sample yielded a mean score of 13.17 (Table 7).

An ANOVA comparison of the three administrations of the CIA (pre, session 7, and post) with each other demonstrated significant across time decreases for the total CIA scores (F (2, 125) = 92.21, p < .01).

Table 7

CIA Total Score Descriptives

Session	n	M	SD
Pre	57	37.00	12.12
Seven (7)	48	29.83	12.52
Post	24	13.17	10.68

The CHEDS and CIA had high correlations at each time point. At the pretreatment session, the CHEDS and CIA total scores were correlated at .68 and increased in degree of correlation from .78 and .89 at session 7 and post-treatment, respectively (all correlations, p < .01). These correlations indicated that the CHEDS and CIA relate to a high degree. Changes observed on the CHEDS in eating disorder symptomatology vary at a high rate together with the CIA, an assessment of life functioning of those suffering from an eating disorder. Table 8 displays the correlation matrix of the CHEDS and CIA.

Table 8

CIA-CHEDS Correlation Matrix

		1	2	3	4	5	6
1.	CHEDS pre	--					
2.	CIA pre	.68*	--				
3.	CHEDS 7	.55*	.45*	--			
4.	CIA 7	.36*	.56*	.78*	--		
5.	CHEDS post	-.16	.07	.28	.61*	--	
6.	CIA post	-.01	.23	.32	.59*	.89*	--

Note. Pearson correlation table with 2-tailed significance. * $p < .05$.

Assessment of CHEDS Clinical Significance

The clinical significance criteria determined by the RCI's and cutscores were used to assess change in the clinical sample during treatment. This was done by utilizing CHEDS total and subscale scores at session 0 and the last session the subject completed, which in most cases is session 20. The cutscore between the community normal and clinical samples for total and subscale levels were used, as well as the respective RCI's. The analyses consisted of calculating the proportion of patients in each of four categories of change: clinical significant change (recovery), reliable improvement, no change, and reliable deterioration and are displayed in Table 9. The recovered category was comprised of only patients who were above the cutscore at pre-treatment (session 0) and thus only those that could exhibit recovery on a given scale.

Table 9

CHEDS Clinical Sample Change Frequencies

Subscale	Recovered	Reliable Improvement	No Change	Reliable Deterioration
Body Preoccupation	62.22%	65.45%	29.09%	5.45%
Body Dissatisfaction	66.67%	66.04%	30.19%	3.77%
Body Checking	63.41%	60.38%	33.96%	5.66%
Binge Eating	54.55%	72.22%	24.07%	3.70%
Restriction	60.53%	62.26%	35.85%	1.89%
Food Preoccupation	57.14%	69.64%	25.00%	5.36%
Vomit	34.29%	26.92%	71.15%	1.92%
Total	75.56%	76.36%	16.36%	7.27%

Note. The reliable improvement category included patients in the recovered category. Recovered category only included patients above the cutscore at session 0.

The change frequency analysis indicated that 8 of the 44 subjects that displayed reliable improvement on the total score level did not meet criteria for recovery. Of these eight subjects, two were below the cutscore at session 0 and were not eligible for the recovered category, while six of these subjects displayed reliable improvement but did not cross over the cutscore between the community normal and clinical samples. At session 0, a total of five of the patients were below the cutscore on the total CHEDS score. Two of these subjects displayed reliable change, two had no change, and one subject indicated deterioration. The subscale level change analysis was similar to the total change score analysis with the majority of patients exhibiting reliable change on the CHEDS subscales. The vomit subscale displayed the lowest percentage of patients exhibiting reliable improvement. However, 33% of patients entering treatment, although meeting criteria for an eating disorder, did not exhibit the symptom of vomiting (see Table 2).

Thus no change would be expected on the vomiting subscale for such patients. The subscale RCI analyses were computed using respective reliability coefficients and thus required higher levels of change as every subscale's reliability was lower than the CHEDS as a whole.

Finally, a chi square analysis was conducted on the four clinical significance categories for the total CHEDS score to determine whether there were significant differences between change categories. The chi square analysis of the CHEDS total score change comparisons was significant (χ^2 (df = 3) = 48.47, $p < .05$). Due to the treatment administered, it was expected that there would be a greater frequency of patients in the reliable improvement and recovery criterion, but the expected frequency could not be approximated. The expected frequencies were divided evenly in the chi square analysis (Table 10).

Table 10

CHEDS Clinical Sample Chi Square Analysis

Change Descriptive	Observed n	Expected n	Expected %	Residual
Recovery	34	23.5	25%	10.5
Reliable Improvement	42	23.5	25%	18.5
No Change	9	23.5	25%	-14.5
Deterioration	4	23.5	25%	-19.5

Discussion

There currently are no comprehensive eating disorder measures that clinicians can use to track symptom change at every session that are reliable, sensitive to change, specific enough to provide information on various dimensions of eating disorder symptoms, and include empirically supported psychometric change indices to determine clinically significant change. Definitions of

clinically significant change for eating disorders have ranged from meeting criteria for an eating disorder to arbitrary criteria of changing one or two standard deviations on a given measure. Current diagnostic measures are also unable to address the limitations of a categorical approach based on diagnosis to assess change and outcome, as exhibited by the difficulties with the eating disordered not-otherwise-specified (EDNOS) category. In addition to utilizing diagnostic measures, which can be time and labor intensive, or relying upon single dimension measures of eating disorder symptoms, clinicians do not have a means by which to assess comprehensive change in eating disorder symptomatology on a session-by-session basis. Diagnostic measures are helpful in distinguishing eating disordered and non-eating disordered groups and unidimensional measure are useful for obtaining information on specific dimensions of symptoms, but do not maximize the effectiveness of symptom tracking or ascertaining therapeutic outcome. The Changes in Eating Disorder Symptoms scale (CHEDS) was developed for this purpose, though lacked empirically-validated change indices to gauge progress.

The current study sought to develop clinically significant change criteria for the Change in Eating Disorder Symptoms scale. Cutscores were established, confirming the receiver operator characteristics (ROC) analysis cutscore found in a previous study, reliable change indices were created to allow for empirically based interpretation of changes in CHEDS scores, and change trajectories were developed utilizing hierarchical linear modeling (HLM) methods. The CHEDS was then compared to an eating disorder life functioning scale, the Clinical Impairment Assessment (CIA), to determine the association between changes in the CHEDS with changes in life impairment. Finally, the psychometric change indices were then used to compare change criteria using the clinical eating disordered sample in this study.

Tests of normal distribution of the clinical sample yielded a normal distribution. Examining the descriptive statistics of the CHEDS obtained in this study, the means and standard deviations appear to be indicative of two populations. The community normal sample had a mean and standard deviation which were all significantly lower than the mean and standard deviation for patients. The clinical sample indeed yielded higher scores than the community normal sample, which is supported by the fact that the samples are based on eating disorder diagnoses. This also coincides with correlations with other eating disorder diagnostic and unidimensional measures (Spangler, 2010). The community normal sample, though, was shown to have some cases with high scores despite not meeting criteria for an eating disorder diagnosis, as indicated by ranges observed. The means, standard deviations, and ranges showed that there is some overlap between the community normal and clinical samples, which were expected. As Tingey et al. (1996) suggests, relevant normative samples should be selected for comparison that are close in proximity to compare samples in a range of dysfunction. It is possible that a more dysfunctional sample, such as moving from the outpatient sample used in this study to an inpatient sample, could be used for further comparison.

Cutscores between community normal and clinical samples were derived for the total scale and subscales of the CHEDS. The total cutscore that was derived in this study (i.e., 65) corroborated well with the ROC analysis cutscore of 60 obtained by Spangler (2010). While the ROC analysis ascertains the best balance of specificity and sensitivity between samples or populations, the cutscore calculation used herein uses a different calculation that produced a quite similar cutscore. The cutscore fell in between the patient and community sample means. In addition to replicating the ROC analysis for total CHEDS score, cutscores for the subscale scores were also generated which can be used to help determine clinically significant change criteria for

each subscale of the CHEDS. The cutscores on the subscales of the CHEDS add to the utility of the measure as scores on the subscales can now also be interpreted in the same manner as the total scores. The cutscores also lay the groundwork to use with RCI's to determine clinically significant change.

Reliable change indices were established for the CHEDS total and subscale scores. The RCI is a difference score needed to be a significant change, or change that is attributed to actual change rather than chance. The RCI can be calibrated to varying levels of significance determined by the stringency that a clinician or researcher prefers. The calculation of the RCI can utilize varying alpha levels according to a clinician's discretion to what confidence level they desire to indicate a reliable change. As the alpha level increases, the amount of change on the CHEDS required to be considered reliable change decreases. Not only are the RCI for total score and subscale scores useful in indicating positive change (i.e., decrease in frequency and/or severity of eating disorder symptoms) but also negative change or deterioration, such as an increase in frequency and severity of eating disorder symptoms. The CHEDS subscale RCI's can be used to determine change similar to the total score RCI. The CHEDS RCI's provide a clinician with a patient's change score that is readily interpretable. Whereas in the past a change or difference score was perhaps compared to something like the standard deviation, the RCI can now be used as an empirically supported change index that takes into account measurement error, utilizing a measure's reliability in the calculation of required change.

As a result of a high internal reliability of the CHEDS, the difference score required to be considered reliably changed is less than the CHEDS standard deviation. As previously noted, the more sensitive and reliable the instrument being used, less error will be observed and change is more reliably measured. The absence of positive or negative change can also be determined if an

individual's initial score and final score did not achieve a positive or negative RCI. Additionally, utilizing subscale level RCI analyses, an individual can exhibit varying levels of change at different rates on different subscales. Change on the subscales can be examined and are not determined by change observed on the total score level. For example, an individual with an eating disorder measured by the CHEDS may change on certain subscales but not on all subscales uniformly and it is possible that total score reliable improvement is achieved without change observed on all CHEDS subscales. Similarly, certain subscales may reliably change but an individual may not exhibit a total score reliable improvement. The amount and significance of changes in specific symptoms can now be examined using the CHEDS. With the CHEDS and psychometric change indices, clinicians will be able to tailor interventions and focus on relevant areas of disturbance in an individual in an idiosyncratic manner by tracking specific symptoms that might not be exhibiting reliable change.

Hierarchical linear modeling (HLM), a multilevel analysis, was established on the clinical sample using the CHEDS to create trajectories based on the initial level of disturbance on the total and subscale levels. The HLM analyses illustrated how CHEDS scores of individuals change over time, determinant on certain variables. Tukey's Ladder of Transformation was completed on a simple model to determine if the time variable was to be modified to match the shape of the trajectories. The results supported that a time variable transformation was not needed. Two theories were used to create the models: utilizing initial scores or initial score severity strata. In these two models, variables of time (session), diagnosis, interaction with initial score or severity strata, and interactions with diagnosis were used. The severity strata model produced a very low rate of change. This was likely a result of too few of subjects per severity strata to create a viable model. Using this model, 40-session clients were

shown to deteriorate unless they were in the high severity strata. This may be explained, though, by the egosyntonic presentation of underweight individuals as many 40-session clients may deny or minimize the severity of symptoms than those classified as a 20-session client. The initial score HLM model culminated into the strongest model as the severity strata model yielded more variance indicated by 2-Log likelihood with the diagnosis variable dropped. The HLM calculations were constructed for all subscales and the total score.

Utilizing the cutscores, RCI's, and the HLM trajectory calculations, a CHEDS scoring macro program was developed. This program can be used to record responses on the CHEDS at every session, score the CHEDS, generate subscale scores, and map out expected trajectories of change. The program creates charts and corresponding graphs with idiosyncratic RCI values and cutscores between the eating disordered and non-eating disordered groups with separate charts and graphs for total and subscale scores. This program can be used to track observed scores compared to expected scores, with the criteria for reliable improvement, recovery, deterioration, and no change displayed in the charts and mapped onto the graphs. Instructions on how to utilize the program are contained in the first worksheet and each other worksheet is appropriately titled. A clinician is also able to select a 20-session model or 40-session model, which will alter the calculations using the session by diagnosis (20 vs. 40 session subjects) interactions. Data entry space is provided for inputting CHEDS scores on up to 40 sessions. Varying alpha levels used for the RCI calculations can also be selected based on the discretion of the user to select a more stringent or lenient criteria for reliable change. This program can be a useful tool to apply the findings of this study to track symptom change and outcome of patients with eating disorders and for clinicians to maximize the uses of these clinically significant change measures. The HLM trajectories are visualized in the graphs and a clinician is then able to estimate when a client

should achieve clinically significant change or if they are deviating from the projected path. When deviations occur or if certain subscales are not changing at an expected rate, interventions can then be tailored in treatment to help target and abate possible explanations of these issues.

Correlations between the CIA and CHEDS were calculated and the CIA and CHEDS indicated a high correlation, strengthening over time. This finding supported the CHEDS as a measure of eating disorder symptoms as it is expected that as symptoms decrease an increase in life functioning, which the CIA measures, should be exhibited. This finding is consistent with other measure comparisons from a previous examination of the CHEDS (Spangler, 2010).

The outcome of the analysis of change frequencies on the clinical sample of the CHEDS was promising. The CBT treatment received by the clinical sample was shown to be effective. On the CHEDS total score level, 76% exhibited reliable improvement and nearly 76% of patients were recovered at their last session. As previously noted, the recovered category was derived excluding patients that were below the cutscore at pretreatment while the reliable improvement category includes patients in the recovered category. About 16% of subjects experienced no change and 7% were seen to have deteriorated. It is to be noted that the data used in the CHEDS analysis included all patients, which was comprised of those who were still continuing treatment or dropped out of treatment, not just full treatment completers. This analysis used an alpha level of .05, which is also very stringent, meaning more change was required to reach clinical significance criteria. All subjects in the clinical sample met criteria for an eating disorder on the EDE, though on the total score level, five patients were below the CHEDS total cutscore. Despite the clinical and community samples being very distinct as indicated by the ANOVA, it was expected that there would be some overlap between the samples as the samples are closest in functioning (i.e., community normal and outpatient sample, opposed to a community normal and

intensive inpatient sample). The overlap is exhibited in the ranges of scores observed. Similarly, the non-eating disorder sample also exhibited elevations on some eating disorder symptoms despite not meeting full criteria for an eating disorder. The five clinical patients who fell below the CHEDS total cutscore may be the result of underreporting of symptoms, which may have been circumvented on the EDE, which is interviewer-based, but not on the self-report CHEDS. Two of these five patients were diagnosed with AN according to the EDE. Persons with AN are likely to endorse fewer categories of symptoms as most will not have elevations on symptoms such as binge eating or purging which may result in lower overall CHEDS scores.

CHEDS subscale level change was shown to be similar to the total score level, with the exception of the vomit subscale. The lower level of recovery and reliable change on the vomit subscale is likely due to several factors. First, although all patients met criteria for an eating disorder at baseline, a large portion of the patients did not meet criteria for bulimia nervosa and thus came into the treatment without showing elevations on the vomit subscale (33%). Therefore, a large segment of patients fell below the vomit subscale cutoff at baseline and would not be expected to change reliably on this subscale during treatment as there was no change to be made. Additionally, the amount of change necessary on the vomit subscale is proportionate to its reliability level as subscale RCI's were calculated using respective subscale reliability levels and are independent from the total score RCI. As a result, the vomit subscale's initial low mean, higher standard deviation, and lower internal reliability compared to other subscales, made the change necessary to be considered significant on the vomit subscale large in comparison to other subscales. Additionally, all subscales had patients who did not exceed the cutscores at baseline. This is expected as not all patients on the eating disorder spectrum would be expected to exhibit

every symptom on the CHEDS, nor have elevations on all subscales nor exceed the cutoff value for all subscales.

The problematic use of the EDNOS category as a result of diagnostic measures' inability to distinguish well between subthreshold eating disorders and the absence of an eating disorder diagnosis as well as limitations of non-comprehensive unidimensional measures are also addressed by the CHEDS. The development of a dimensional, multidimensional measure, the CHEDS, may help to circumvent the limitations of categorical definitions of change based on diagnoses by assessing a spectrum of eating disorder symptoms with empirically validated change indices. Making use of the CHEDS, what is considered "recovered" can be defined in clinical significance terms of changing reliably and being more representative of a functional than less-functional population rather than using a qualitative or arbitrary assessment of change. This may also eliminate some of the difficulties diagnosing subthreshold eating disorders and an absence of an eating disorder with tracking change and outcome. Tracking change and outcome can now employ a brief but comprehensive eating disorder symptom measure which gauges functioning of patients on a session-by-session basis, independent from diagnoses, using solidified clinical significant change criteria.

This study utilized a clinical sample of a wide range of eating disordered subjects, increasing the versatility of the CHEDS. In addition, the aim of developing cutscores is to have a means of distinguishing between a community normal population and clinical eating disorder population, rather than relying solely on diagnostic criteria. The implication of cutscore development is providing a means for clinicians to determine if an individual's CHEDS score is more representative of a functional versus less functioning population. The RCI's also stand

apart from diagnoses as the calculations are based on empirically validated methods of using measurement error and bearing a change value on the CHEDS that is readily interpretable.

Further, clinicians are provided with an additional promising tool in assessing paths to recovery using HLM trajectories to predict the slope and point in time for recovery, reliable improvement, or even gauge a lack of change or deterioration. These additional change indices can be used on a session-by-session basis using the CHEDS, allowing clinicians treating individuals with eating disorders to better track and gauge progress or lack thereof. The development of the CHEDS tracking and outcome scoring program also has the potential to make a significant impact on how clinicians are able to use this measure in everyday practice.

Study Strengths

Results of this study could have an immediate and significant impact on the utilization of eating disorder patient evaluations of change and progress. This study has several strengths. There is a clear operationalization of the constructs measured by the CHEDS. It has high internal reliability, and a stable and valid subscale structure. The RCI and cutscores calculated provide clear operationalizations of clinically significant change on the CHEDS, used in conjunction with the trajectory analyses. The study also utilizes the CIA as another indicator of clinically significant change, measuring life functioning for those with eating disorders, expanding the exemplars of possible indicators of change. The CHEDS has been calibrated with the convergent validity of the current use of outcome criteria of eating disorders: the assessment of meeting criteria for an eating disorder at pre and post treatment using diagnostic measures. The high concurrent validity with measures such as the EDE further validates the CHEDS as a comprehensive measure of eating disorder symptoms.

The strengths of the study are also rooted in the controls implemented in the study, such as the manualized protocol used for treatment of the eating disordered patients. The reliability of treatment implementation, for example, is indicated as a strength as there are various fidelity checks to the manual, including pre-training, videotaped sessions, weekly supervision, and session-by-session fidelity checks. This supports the assumption that the internal validity of the study is high as the causal relationship between treatment and outcome is clear. Confounds such as varying implementation of treatments and compensatory equalization of treatments are minimized. The homogeneity of the study also increases internal validity of putative cause and effect of the treatment of eating disorders.

Future directions of CHEDS studies should assess a larger sample size of both eating disordered and non-eating disordered populations, also including various data collection sites. An inpatient sample may also be used as another comparison population and cutscores generated between outpatient and inpatient samples. This sample could also be used to determine the trajectories and RCI calculations to increase the generalizability of the measure. Though, it is supported that outpatient treatment is more empirically validated than inpatient treatment. Outpatient samples are very similar to the inpatient sample except for the higher probability of including more individuals that are underweight or meet criteria for AN, given the nature of individuals hospitalized primarily for an eating disorder. Another possible future study could replicate this study in a population that have utilized the scoring program using the RCI's, cutscores, and HLM trajectories produced from this study. This may result in improved rates of change as clinicians would be able to utilize this tracking and outcome measure and tailor treatments accordingly.

Limitations

There were a few limitations of the study. The number of patients used in the study was not large. However, for calculations such as the HLM analyses, there were several thousand data points used and additions of more data points would likely not change the calculations in any substantive manner. As a result of the sample size, the HLM analyses of the strata of severity scores was not successful and can only be used as a preliminary analysis for future research.

The clinical sample consisted of fairly homogeneous individuals in terms of demographics, with only one site for data collection, which collected the sample utilizing convenience sampling. This homogeneity of subjects limits the generalizability of the change metrics created to populations that do not have similar demographic characteristics. Further, while the sample size is more than adequate for the power analysis, the proportion of anorexia nervosa subjects was small. The analyses were based primarily on individuals with bulimia nervosa and EDNOS. Though this is representative of the prevalence rates of eating disorders and of treatment resistance in persons with anorexia nervosa, it also has implications of a lack of generalizability for anorexia nervosa.

The larger treatment study from which the patient sample was drawn utilized various other measures to evaluate treatment. This introduced the possibility of a validity threat of interaction of testing and treatment. The various testing that each individual received may inadvertently prime the individuals in treatment to their symptoms of eating disorders, though this may be minimal since all patients spontaneously present to the study reporting such symptoms.

Overall, with the psychometric change indices and trajectories produced in this study, the eating disordered population can be better served with clinicians that will be able to better utilize

the CHEDS as a measure of change and progress of patients. With the development of this comprehensive, brief measure of eating disorder symptoms that can be given at each session, these psychometric tools help maximize the potential of this measure. The psychometric change indices and scoring program created in this study provide promising tools that have not been applied to eating disorder treatment up to this point.

References

Abraham, S. F., Pettigrew, B., Boyd, C., Russell, J., & Taylor, A. (2005). Usefulness of amenorrhoea in the diagnoses of eating disorder patients. *Journal of Psychosomatic Obstetrics and Gynecology, 26*(3), 211-215.

Achenbach, T. M. (2001). What are norms and why do we need valid ones? *Clinical Psychology Science Practice, 8*, 446-450.

Agras, W. S., Crow, S. J., Halmi, K. A., Mitchell, J. E., Wilson, G. T., & Kraemer, H. C. (2000). Outcome predictors for the cognitive behavior treatment of bulimia nervosa: data from a multisite study. *American Journal of Psychiatry, 157*, 1302-1308.

American Psychiatric Association (2000). (DSM-IV-TR) *Diagnostic and Statistical Manual of Mental disorders, Fourth Edition, Text Revision*. Washington DC: American Psychiatric Press, Inc.

Ametller, L., Castro, J., Serrano, E., Martinez, E., & Toro, J. (2005). Readiness to recover in adolescent anorexia nervosa: prediction of hospital admission. *Journal of Child Psychology and Psychiatry, 46*(4), 394-400.

Anderson, D. A., Williamson, D. A., Duchmann, E. G., Gleaves, D. H., & Barbin, J. M. (1999). Development and validation of a multifactorial treatment outcome measure for eating disorders. *Assessment, 6*(1), 7-20.

Arnow, B., Kenardy, J., & Agras, W.S. (1995). The emotional eating scale: the development of a measure to assess coping with negative affect by eating. *International Journal of Eating Disorders, 18*(1), 79-90.

Bastiani, A. M., Rao, R., Weltzin, T., & Kaye, W. H. (1995). Perfectionism in anorexia nervosa. *International Journal of Eating Disorders, 17*(2), 147-152.

Bauer, S., Winn, S., Schmidt, U., & Kordy, H. (2005). Construction, scoring and validation of the Short Evaluation of Eating Disorders (SEED). *European Eating Disorders Review, 13*(3), 191-200.

Bennett, K. (1997). The internal structure of the eating disorder inventory. *Health Care for Women International, 18*(5), 495-504.

Benninghoven, D., Jurgens, E., Mohr, A., Hberlein, I., Kunzendorf, S., & Jantschek, G. (2006). Different changes of body-images in patients with anorexia or bulimia nervosa during inpatient psychosomatic treatment. *European Eating Disorders Review, 14*, 88-96.

Ben-Tovim, D. I. (2003). Eating disorders: outcome, prevention and treatment of eating disorders. *Current Opinion in Psychiatry, 16*, 66-69.

Beutler, L. E., & Moleiro, C. (2001). Clinical versus reliable and significanct change. *Clinical Psychology Science Practice, 8*, 441-445.

Binford, R. B., Mussell, M. P., Peterson, C. B., Crow, S.J., & Mitchell, J. E. (2004). Relation of binge eating age of onset to functional aspects of binge eating in binge eating disorder. *International Journal of Eating Disorders, 35*, 286-292.

Binford, R. B., Le Grange, D., & Jellar, C. C. (2005). Eating disorders examination versus eating disorders examination-questionnaire in adolescents with full and partial-syndrome bulimia nervosa and anorexia nervosa. *International Journal of Eating Disorders, 37*, 44-49.

Bohn, K. F., & Fairburn, C. G. (2008). The Clinical Impairment Assessment Questionnaire. In C. G. Fairburn (Ed.), *Cognitive-Behavior Therapy and Eating Disorders*. New York: Guilford Press.

Brown, G., Burlingame, G. M., Lambert, M. J., Jones, E., & Vaccaro, J. (2001). Pushing the

quality envelope: a new outcomes management system. *Psychiatric Services, 52,* 925-934.

Burlingame, G. M., Hwang, A. D., Lee, J., Rees, F., & Earnshaw, D. (in progress). Clinical significance and the Brief Psychiatric Rating Scale, Expanded Version.

Burlingame, G. M., Seaman, S., Johnson, J. E., Whipple, J., Richardson, E., Rees, F., Freeman, D., Spencer, R., Payne, M., & O'Neil, B. (2006). Sensitivity to change of the brief psychiatric rating scale-extended (SEDS-E): an item and subscale analysis. *Psychological Services, 3*(2), 77-87.

Burns, D.D. (1995). *The Burns Toolkit.* Author: Los Altos Hills, CA.

Calugi, S., & Grave, R. D. (2006). Validation of the Body Checking Questionnaire (BCQ) in an eating disorders population. *Behavioural and Cognitive Psychotherapy, 34*, 233-242.

Carter, J. C., Aime, A. A., & Mills, J. S. (2001). Assessment of bulimia nervosa: a comparison of interview and self-report questionnaire methods. *International Journal of Eating Disorders, 30*, 187-192.

Cash, T. F., & Deagle, E. A., III. (1997). The nature and extent of body-image disturbances in anorexia nervosa and bulimia nervosa: A meta-analysis. *International Journal of Eating Disorders, 22*(2), 107-125.

Celio, A. A., Wilfley, D. E., Crow, S. J., Mitchell, J., & Walsh, B. T. (2004). A comparison of the binge eating scale, questionnaire for eating and weight patterns-revised, and eating disorder examination questionnaire with instructions with the eating disorder examination in the assessment of binge eating disorder and its symptoms. *International Journal of Eating Disorders, 36*, 434-444.

Clinton, D., & Norring, C. (1999). The rating of anorexia and bulimia (RAB) interview:

development and preliminary validation. *European Eating Disorders Review, 7*, 362-371.

Cooley, E., & Toray, T. (2001). Body image and personality predictors of eating disorder symptoms during the college years. *International Journal of Eating Disorders, 30*, 28-36.

Cooper, Z., & Fairburn, C. (1987). The eating disorder examination: A semi-structured interview for the assessment of the specific psychopathology of eating disorders. *International Journal of Eating Disorders, 6*(1), 1-8.

Corstorphine, E., Mountford, V., Tomlinson, S., Waller, G., & Meyer, C. (2007). Distress tolerance in the eating disorders. *Eating Behaviors, 8*, 91-97.

Crow, S. J., Agras, W.S., Halmi, K., Mitchell, J. E., & Kraemer, H. C. (2002). Full syndromal versus subthreshold anorexia nervosa, bulimia nervosa, and binge eating disorder: A multicenter study. *International Journal of Eating Disorders, 32*, 309-318.

Crow, S. J., Peterson, C. B., Levine, A. S., Thuras, P., & Mitchell, J. E. (2004). A survey of binge eating and obesity treatment practices among primary care providers. *International Journal of Eating Disorders, 35*, 248-353.

Cumella, E. J. (2006). Review of the Eating Disorder Inventory-3. *Journal of Personality Assessment, 87*(1), 116-117.

Cuzzolaro, M., Vetrone, G., Marano, G., & Garfinkel, P. E. (2006). The Body Uneasiness Test (BUT): Development and validation of a new body image assessment scale. *Eating Weight Disorders, 11*, 1-13.

Davis, R., Olmsted, M. P., & Rockert, W. (1990). Brief group psychoeducation for bulimia nervosa: Assessing the clinical significance of change. *Journal of Consulting and Clinical Psychology, 58*(6), 882-885.

Eberenze, K. P., & Gleaves, D. H. (1994). An examination of the internal consistency and factor

structure of the Eating Disorder Inventory-2 in a clinical sample. *International Journal of Eating Disorders, 16*(4), 371-379.

Eklund, K., Paavonen, E. J., & Almqvist, F. (2005). Factor structure of the Eating Disorder Inventory-C. *International Journal of Eating Disorders, 37*, 330-341.

Engel, S. G., Wittrock, D. A., Crosby, R. D., Wonderlich, S. A., Mitchell, J. E., & Kolotkin, R. L. (2006). Development and psychometric validtation of an eating disorder-specific health-related quality of life instrument. *International Journal of Eating Disorders, 39*, 62-71.

Engelsen, B. K., & Laberg, J. C. (2001). A comparison of three questionnaires (EAT-12, EDI, and EDE-Q) for assessment of eating problems in healthy female adolescents. *Nordic Journal of Psychiatry, 55*, 129-135.

Espelage, D. L., Mazzeo, S. E., Aggen, S. H., Quittner, A. L., Sherman, R., & Thompson, R. (2003). Examining the construct validity of the Eating Disorder Inventory. *Psychological Assessment, 15*(1), 71-80.

Fairburn, C. G., & Cooper, Z. (1993). Binge eating: nature, assessment, and treatment. In C. G. Fairburn, & Wilson, G.T. (Eds.), *Approaches to Assessment and Management*. New York: Guilford Press.

Fairburn, C. G., & Beglin, S. J. (1994). Assessment of eating disorders: Interview or self-report questionnaire? *International Journal of Eating Disorders, 16*(4), 363-370.

Fairburn, C. G., Agras, W. S., Walsh, B. T., Wilson, G. T., & Stice, E. (2004). Prediction of outcome in bulimia nervosa by early change in treatment. *American Journal of Psychiatry, 161*(12), 2322-2324.

Fairburn, C. G., Cooper, Z., Bohn, K., O'Conner, M. E., Doll, H. A., & Palmer, R. L. (2007). The

severity and status of eating disorder NOS: Implications for DSM-V. *Behavior Research and Therapy, 45*, 1705-1715.

Ferguson, R. J., Robinson, A. B., & Splaine, M. (2002). Use of the reliable change index to evaluate clinical significance in SF-36 outcomes. *Quality of Life Research, 11*, 509-516.

Fichter, M. M., & Quadflieg, N. (2004). Twelve-year course and outcome of bulimia nervosa. *Psychological Medicine, 34*, 1395-1406.

Follette, W. C., & Callaghan, G. M. (2001). The evolution of clinical significance. *Clinical Psychology Science Practice, 8*, 431-435.

Forbush, K., Heatherton, T. F., & Keel, P. K. (2007). Relationships between perfectionism and specific disordered eating behaviors. *International Journal of Eating Disorders, 40*, 37-41.

Ghaderi, A., & Andersson, G. (1999). Meta-analysis of CBT for bulimia nervosa: Investigating the effects using DSM-III-R and DSM-IV criteria. *Scandinavian Journal of Behaviour Therapy, 28*(2), 79-87.

Gharderi, A., & Scott, B. (2002). The preliminary reliability and validity of the survey for eating disorders (SEDS): A self-report questionnaire for diagnosing eating disorders. *European Eating Disorders Review, 10*, 61-76.

Gila, A., Castro, J., Gomez, M. J., & Toro, J. (2005). Social and body self-esteem in adolescents with eating disorders. *International Journal of Psychology & Psychological Therapy, 5*(1), 63-71.

Gilbert, N., & Meyer, C. (2005). Fear of negative evaluation and the development of eating psychopathology: a longitudinal study among nonclinical women. *International Journal of Eating Disorders, 37*, 307-312.

Goldfarb, L. A., Dykens, E. M., & Gerrard, M. (1985). The Goldfarb Fear of Fat Scale. *Journal of Personality Assessment, 49*(3), 329-332.

Grilo, C. M., Masheb, R. M., & Wilson, G. T. (2001). A comparison of different methods for assessing the features of eating disorders in patients with binge eating disorder. *Journal of Consulting and Clinical Psychology, 69*(2), 317-322.

Guest, T. (2000). Using the eating disorder examination in the assessment of bulimia and anorexia: Issues of reliability and validity. *Social Work in Health Care, 31*(4), 71-83.

Hageman, W. J., & Arrindell, W. A. (1993). A further refinement of the Reliable Change (RC) Index by improving the pre-post difference score: Introducing RC-sub(ID). *Behaviour Research and Therapy, 31*(7), 693-700.

Hageman, W. J. J. M., & Arrindell, W. A. (1999). Establishing clinically significant change: Increment of precision and the distinction between individual and group level of analysis. *Behaviour Research and Therapy, 37*(12), 1169-1193.

Halmi, K. A., Agras, W. S., Crow, S., Mitchell, J., Wilson, G. T., Bryson, S. W., & Kraemer, H. C. (2005). Predictors of treatment acceptance and completion in anorexia nervosa. *Archives of General Psychiatry, 62*, 776-781.

Halmi, K. A., Agras, W. S., Mitchell, J., Wilson, G. T., Crow, S., Bryson, S. W., et al. (2002). Relapse predictors of patients with bulimia nervosa who achieved abstinence through cognitive behavioral therapy. *Archives of General Psychiatry, 59*(12), 1105-1109.

Harrison, D. and Raudenbush, S. W. (2006). Linear regression and hierarchical linear models. In Green, J., Camilli, G., & Elmore, P. (Eds). *Complementary Methods in Education Research.* Washington, DC, American Educational Research Association. 24:411-426.

Hinrichsen, H., Garry, J., & Waller, G. (2006). Development and preliminary validation of the

Testable Assumptions Questionnaire-Eating Disorders (TAQ-ED). *Eating Behaviors, 7*, 275-281.

Jacobson, N. S., Follette, W. C. and Revenstorf, D. (1984). Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. *Behavior Therapy 15*, 336-352.

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*(1), 12-19.

Jarman, M., & Walsh, S. (1999). Evaluating recovery from anorexia nervosa and bulimia nervosa: Integrating lessons learned from research and clinical practice. *Clinical Psychology Review, 19*(7), 773-788.

Joiner, T. E., & Heatherton, T. F. (1998). First- and second-order factor structure of five subscales of the Eating Disorders Inventory. *International Journal of Eating Disorders, 23*, 189-198.

Kazdin, A. E. (2001). Almost clinically significant ($p<.10$): current measures may only approach clinical significance. *Clinical Psychology Science Practice, 8*, 455-462.

Keel, P. K., Mitchell, J. E., Davis, T. L., Fieselman, S., & Crow, S. J. (1999). Impact of definitions on the description and prediction of bulimia nervosa outcome. *Journal of Eating Disorders, 28*, 377-386.

Keel, P. K., Mitchell, J. E., Miller, K. B., Davis, T. L., & Crow, S. J. (1999). Long-term outcome of bulimia nervosa. *Archives of General Psychiatry, 56*, 63-69.

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling* (Vol. 2). New York: The Guilford Press.

Kordy, H., Percevic, R., & Martinovich, Z. (2001). Norms, normality, and clinical significant change: Implications for the evaluation of treatment outcomes for eating disorders. *International Journal of Eating Disorders, 30*, 176-186.

Kutlesic, V., Williamson, D. A., Gleaves, D. H., Barbin, J. M., & Murphy-Eberenz, K. P. (1998). The interview for the diagnosis of eating disorders-iv: application to the DSM-IV diagnostic criteria. *Psychological Assessment, 10(1),* 41-48.

Laessle, R. G., Zoettl, C., & Pirke, K. M. (1987). Metaanalysis of treatment studies for bulimia. *International Journal of Eating Disorders, 6*(5), 647-653.

Lambert, M. J., Okiishi, J. C., Finch, A. E., & Johnson, L. D. (1998). Outcome assessment: from conceptualization to implementation. *Professional Psychology: Research and Practice, 29*(1), 63-70.

Laurenceau, J. P., Hayes, A. M., & Feldman, G. C. (2007). Statistical and methodological issues in the study of change in psychotherapy. *Clinical Psychology Review*, *27,* 682-695.

le Grange, D., Binford, R. B., Peterson, C. B., Crow, S. J., Crosby, R. D., Klein, M. H., Bardone-Cone, A. M., Joiner, T. E., Mitchell, J. E., & Wonderlich, S. A. (2006). DSM-IV threshold versus subthreshold bulimia nervosa. *International Journal of Eating Disorders, 39*, 462-467.

Leung, N., & Price, E. (2007). Core beliefs in dieters and eating disordered women. *Eating Behaviors, 8*, 65-72.

Levy, R. K. (1997). The transtheoretical model of change: An application to bulimia nervosa. *Psychotherapy, 34*(3), 278-285.

Lewandowski, L. M., Gebing, T. A., Anthony, J. L., & O'Brien, W. H. (1997). Meta-analysis of cognitive-behavioral treatment studies for bulimia. *Clinical Psychology Review, 17*(7),

703-718.

Limbert, C. (2001). A comparison of female university students from different school backgrounds using the Eating Disorder Inventory. *International Journal of Adolescent Medicine and Health, 13*(2), 145-154.

Limbert, C. (2004). The Eating Disorder Inventory: A test of the factor structure and internal consistency in a nonclinical sample. *Health Care for Women International, 25*, 165-178.

Loeb, K. L., Pike, K. M., Walsh, B. T., & Wilson, G. T. (1994). Assessment of diagnostic features of bulimia nervosa: Interview versus self-report format. *International Journal of Eating Disorders, 16*(1), 75-81.

Luce, K. H., & Crowther, J. H. (1999). The reliability of the eating disorder examination - self-report questionnaire version (EDE-Q). *International Journal of Eating Disorders, 25*, 349-351.

Lundgren, J. D., Danoff-Burg, & Anderson, D. A. (2004). Cognitive-behavioral therapy for bulimia nervosa: An empirical analysis of clinical significance. *International Journal of Eating Disorders, 35*, 262-274.

Lunnen, K. M., & Ogles, B. M. (1998). A multiple, multivariable evaluation of reliable change. *Journal of Consulting and Clinical Psychology, 66*(2), 400-410.

Martin, C. K., Williamson, D. A., & Thaw, J. M. (2000). Criterion validity of the Multiaxial Assessment of Eating Disorders Symptoms. *International Journal of Eating Disorders, 28*(3), 303-310.

McCarthy, D. M., Simmons, J. R., Smith, G. T., Tomlinson, K. L., & Hill, K. K. (2002). Reliability, stability, & factor structure of the bulimia test-revised and Eating Disorder Inventory-2 scales in adolescence. *Assessment, 9*, 382-389.

Miller, J. L., Schmidt, L. A., Vaillancourt, T., McDougall, P., & Laliberte, M. (2006). Neuroticism and introversion: A risky combination for disordered eating among a non-clinical sample of undergraduate women. *Eating Behaviors, 7*, 69-78.

Mintz, L. B., O'Halloran, M. S., Mulbolland, A. M., & Schneider, P. A. (1997). Questionnaire for eating disorder diagnoses: Reliability and validity of operationalizing DSM-IV criteria into a self-report format. *Journal of Counseling Psychology, 44*(1), 63-79.

Mintz, L. B., & O'Halloran, M.S. (2000). The Eating Attitudes Test: Validation with DSM-IV eating disorder criteria. *Journal of Personality Assessment, 73*(3), 489-503.

Mitchell, J. E., Hoberman, H. N., Peterson, C. B., Mussell, M., & Pyle, R. L. (1996). Research on the psychotherapy of bulimia nervosa: Half empty or half full. *International Journal of Eating Disorders, 20*(3), 219-229.

Mond, J. M., Hay, P. J., Rodgers, B., Owen, C., & Beumont, P. J. V. (2004). Temporal stability of the Eating Disorder Examination Questionnaire. *International Journal of Eating Disorders, 36*, 195-203.

Mond, J. M., Hay, P. J., Rodgers, B., Owen, C., & Beumont, P. J. V. (2004). Validity of the Eating Disorder Examination Questionnaire (EDE-Q) in screening for eating disorders in community samples. *Behavior Research and Therapy, 42*, 551-567.

Mond, J. M., Hay, P. J., Rodgers, B., & Owen, C. (2006). Eating Disorder Examination Questionnaire (EDE-Q): Norms for young adult women. *Behavior Research and Therapy, 44*, 53-62.

Nangle, D. W., Johnson, W. G., Carr-Nangle, R. E., & Engler, L. B. (1994). Binge eating disorder and the proposed DSM-IV criteria: Psychometric analysis of the questionnaire of eating and weight patterns. *International Journal of Eating Disorders, 16*(2), 147-157.

Nevonen, L., Clinton, D., & Norring, C. (2006). Validating the EDI-2 in three Swedish female samples: Eating disorders patients, psychiatric outpatients and normal controls. *Nordic Journal of Psychiatry, 60*, 44-50.

Niego, S. H., Pratt, E. M., & Agras, W. S. (1997). Subjective or objective binge: Is the distinction valid? *International Journal of Eating Disorders, 22*, 291-298.

Nollett, C. L., & Button, E. J. (2005). Questionnaire measures of psychopathology in eating disorders: Comparisons between clinical groups. *European Eating Disorders Review, 13*, 211-215.

Openshaw, C., Waller, G., & Sperlinger, D. (2004). Group cognitive-behavior therapy for bulimia nervosa: Statistical versus clinical significance of changes in symptoms across treatment. *International Journal of Eating Disorders, 36*, 363-375.

Peake, K. J., Limbert, C., & Whitehead, L. (2005). Gone, but not forgotten: An examination of the factors associated with dropping out from treatment of eating disorders. *European Eating Disorders Review, 13*, 330-337.

Peterson, C. B., Mitchell, J. E., Engbloom, S., Nugent, S., Mussell, M. P., Crow, S. J., & Miller, J. P. (1998). Binge eating disorder with and without a history of purging symptoms. *International Journal of Eating Disorders, 24*, 251-257.

Peterson, C. B., Crosby, R. D., Wonderlich, S. A., Joiner, T., Crow, S. J., Mitchell, J. E., Bardone-Cone, A. M., Klein, M., & le Grange, D. (2007). Psychometric properties of the Eating Disorder Examination-Questionnaire: Factor structure and internal consistency. *International Journal of Eating Disorders, 40*, 386-389.

Petrie, T. A., Tripp, M. M., & Harvey, P. (2002). Factorial and construct validity of the body parts satisfaction scale-revised: An examination of minority and nonminority women.

Psychology of Women Quarterly, 26, 213-221.

Pratt, E. M., Niego, S. H., & Agras, W. S. (1998). Does the size of a binge matter? *International Journal of Eating Disorders, 24*, 307-312.

Raudenbush, S. W., & Bryk, A. S. (2001). Hierarchical linear models: Applications and data analysis methods. *Advanced Quantitative Techniques in the Social Sciences.* Thousand Oaks, California: Sage Publications, Inc.

Reas, D. L., Grilo, C. M., & Masheb, R. M. (2006). Reliability of the Eating Disorder Examination-Questionnaire in patients with binge eating disorder. *Behavior Research and Therapy, 44*, 43-51.

Richards, P. S., Baldwin, B. M., Frost, H. A., Clark-Sly, J. B., Berrett, M. E., & Hardman, R. K. (2000). What works for treating eating disorders? Conclusions of 28 outcome reviews. *Eating Disorders, 8*, 189-206.

Rizvi, S. L., Peterson, C. B., Crow, S. J., & Agras, W. S. (2000). Test-retest reliability of the Eating Disorder Examination. *International Journal of Eating Disorders, 28*, 311-316.

Rodriguez-Cano, T., Beato-Fernández, L. (2005). Attitudes towards change and treatment outcome in eating disorders. *Eating Weight Disorders, 10*, 59-65.

Rosen, J. C., Srebnik, D., Saltzberg, E., & Wendt, S. (1991). Development of a Body Image Avoidance Questionnaire. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3(1),* 32-37.

Rosen, J. C., Vara, L., Wednt, S., & Leitenberg, H. (1990). Validity studies of the Eating Disorder Examination. *International Journal of Eating Disorders, 9*(5), 519-528.

Rushford, N. (2006). Fear of gaining weight: Its validity as a visual analogue scale in anorexia nervosa. *European Eating Disorders Review, 14*, 104-110.

Ruuska, J., Kaltiala-Heino, R., Rantanen, P., & Koivisto, A. M. (2005). Are there differences in the attitudinal body image between adolescent anorexia nervosa and bulimia nervosa. *Eating Weight Disorders, 10*, 98-106.

Safer, D. L., Agras, W. S., Lowe, M. R., & Bryson, S. (2004). Comparing two measures of eating restraint in bulimic women treated with cognitive-behavioral therapy. *International Journal of Eating Disorders, 36*, 83-88.

Satandar-Pinnock, K., Woodside, D. B., Carter, J. C., Olmsted, M. P., & Kaplan, A. S. (2003). Perfectionism in anorexia nervosa: A 6-24-month follow-up study. *International Journal of Eating Disorders, 33*, 225-229.

Schoemaker, C., Verbraak, M., Breteler, R., & van der Staak, C. (1997). The discriminant validity of the Eating Disorder Inventory-2. *British Journal of Clinical Psychology, 36*, 627-629.

Shah, N., Passi, V., Bryson, S., & Agras, W. S. (2005). Patterns of eating and abstinence in women treated for bulimia nervosa. *International Journal of Eating Disorders, 38*, 330-334.

Singer, J. D., & Goldstein, H. (2006). Linear regression and hierarchical linear models. In J. L. Green, Camilli, G., & Elmore, P. B. (Eds.), *Handbook of Complementary Methods in Education Research*. Washington, DC: Lawrence Erlbaum Associates, Inc.

Singer, J. D. and Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Press.

Spangler, D. L. (2010). The Change in Eating Disorder Symptom Scale: Scale development and psychometric properties. *Eating Behaviors, 11(3),* 131-137.

Spangler, D. L., & Stice, E. (2001). Validation of the beliefs about appearance scale. *Cognitive*

Therapy and Research, 25(6), 813-827.

Stice, E., Telch, C. F., & Rizvi, S. L. (2000). Development and validation of the Eating Disorder Diagnostic Scale: A brief self-report measure of anorexia, bulimia, & binge-eating disorder. *Psychological Assessment, 12,* 123–131.

Stice, E., Fisher, M., & Martinez, E. (2004). Eating disorder diagnostic scale: Additional evidence of reliability and validity. *Psychological Assessment, 16*(1), 60-71.

Stunkard, A. J., & Messick, S. (1985). The Three-Factor Eating Questionnaire to measure dietary restraint and hunger. *Journal of Psychosomatic Research, 29,* 71–83.

Stotland, S., & Zuroff, D. C. (1990). A new measure of weight locus of control: The Dieting Beliefs Scale. Journal of Personality Assessment, 54(1), 191-203.

Sysko, R., Walsh, B. T., & Fairburn, C. G. (2005). Eating Disorder Examination-Questionnaire as a measure of change in patients with bulimia nervosa. *International Journal of Eating Disorders, 37*, 100-106.

The Medical Management Institute (2008). (ICD-9-CM) *International Classification of Diseases 9th Revision-Clinical Modification.* Salt Lake City: Contexo Media.

Thompson-Brenner, H. J. (2003). *Implications for the treatment of bulimia nervosa: A meta-analysis of efficacy trials and a naturalistic study of treatment in the community.* ProQuest Information & Learning, US.

Tingey, R. C., Lambert, M. J., Burlingame, G. M., & Hansen, N. B. (1996). Clinical significant change: Practical indicators for evaluating psychotherapy outcome. *Psychotherapy Research, 6*(2), 144-153.

Troop, N. A., Serpell, L., & Treasure, J. L. (2001). Specificity in the relationship between depressive and eating disorder symptoms in remitted and nonremitted women.

International Journal of Eating Disorders, 30(3), 306-311.

Turner, H., & Bryant-Waugh, R. (2004). Eating Disorder Not Otherwise Specified (EDNOS): Profiles of clients presenting at a community eating disorder service. *European Eating Disorders Review, 12*, 18-26.

Tylka, T. L., & Subich, L. M. (1999). Exploring the construct validity of the eating disorder continuum. *Journal of Counseling Psychology, 46*(2), 268-276.

Vermeersch, D., Lambert, M., & Burlingame, G. (2000). Outcome questionnaire: Item sensitivity to change. *Journal of Personality Assessment, 74*(2), 242-261.

Vetrone, G., Cuzzolaro, M., Antonozzi, I., & Garfinkel, P. E. (2006). Screening for eating disorders: False negatives and eating disorders not otherwise specified. *European Journal of Psychiatry, 20*(1), 13-20.

Weinfurt, K. P. (2000). Repeated measures analyses: ANOVA, MANOVA, & HLM. In L. G. a. Y. Grimm, P.R. (Ed.), *Reading and Understanding More Multivariate Statistics*. Washington, D.C.: American Psychological Association.

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. *Psychotherapy, 33*(2), 275-283.

Whittal, M. L., Agras, W. S., & Gould, R. A. (1999). Bulimia nervosa: A meta-analysis of psychosocial and pharmacological treatments. *Behavior Therapy, 30*(1), 117-135.

Williamson, D. A., Barker, S. E., Bertman, L. J., & Gleaves, D. H. (1995). Body image, body dysphoria, and dietary restraint: Factor structure in nonclinical subjects. *Behavior Research and Therapy, 33*(1), 85-93.

Williamson, D. A., Womble, L. G., Smeets, M. A. M., Netemeyer, R. G., Thaw, J. M., Kutlesic, V., & Gleaves, D. H. (2002). Latent structure of eating disorder symptoms: A factor analytic and taxometric investigation. *American Journal of Psychiatry, 159*, 412-418.

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendation for future directions. *Journal of Personality Assessment, 82*(1), 50-59.

Wolk, S. L., Loeb, K. L., & Walsh, B. T. (2005). Assessment of patients with anorexia nervosa: Interview versus self-report. *International Journal of Eating Disorders, 37*, 92-99.

Wonderlich, S. A., Crosby, R. D., Joiner, T., Peterson, C. B., Bardone-Cone, A., et al. (2005). Personality subtyping and bulimia nervosa: psychopathological and genetic correlates. *Psychological Medicine, 35*, 649-657.

Zabinski, M. F., Pung, M. A., Wilfley, D. E., Eppstein, D. L., Winzelberg, A. J., Celio, A., & Taylor, C. B. (2001). Reducing risk factors for eating disorders: Targeting at-risk women with a computerized psychoeducational program. *International Journal of Eating Disorders, 29*, 401-408.