EVALUATING THE EFFECTIVENESS OF COUNSELING CENTER-BASED PSYCHOTHERAPY OUTCOME MEASURES: A STATISTICAL COMPARISON OF THE COUNSELING CENTER ASSESSMENT OF PSYCHOLOGICAL SYMPTOMS AND THE OUTCOME QUESTIONNAIRE

by

Elizabeth Michelle Proemmel Duszak

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Educational Psychology

The University of Utah

August 2014

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Elizabeth Michelle Proemmel Duszak**

has been approved by the following supervisory committee members:

Lauren Weitzman	, Chair	**May 30, 2014** Date Approved
Zac Imel	, Co Chair	**May 30, 2014** Date Approved
Lois Huebner	, Member	**May 30, 2014** Date Approved
Amy Jo Metz	, Member	**May 30, 2014** Date Approved
David Robert Davies	, Member	**May 30, 2014** Date Approved

and by **Anne Cook** , Chair of

the Department of **Educational Psychology**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Evaluating the effectiveness of psychotherapy services, such as through client ratings of mental health symptoms, is a vital component of maintaining quality of care. However, the performance of psychotherapy outcome measures is not fully understood. Specifically, there are gaps in knowledge regarding the convergent validity of two widely disseminated measures, the Counseling Center Assessment of Psychological Symptoms (CCAPS) and the Outcome Questionnaire (OQ). The purpose of this study is to test the comparability of the OQ and the CCAPS as psychotherapy outcome measures. The first step to address this research question is to test the convergent validity of OQ Total scores and CCAPS Distress Index scores. Further analyses examine the relationship between these two general distress scores and the subscales of both instruments, which informs the question of whether the two instruments are providing similar or unique information. Clients at one college counseling center completed both the CCAPS and the OQ at every counseling session. The primary method of analysis was multivariate multilevel modeling, in which occasions were nested within clients. More specifically, the Bayesian mixed effects models fit provides point estimates and highest posterior density (HPD) intervals from the simulated parameters. In examining the correlation of the client-level random effects for the OQ Total score and CCAPS Distress Index, the mode of the posterior distribution of the correlated random effects was $r = 0.967$, *HPD*[.962, .971], suggesting that the two

measures are highly correlated. Unfortunately, when we included session number as part of the model, the multivariate multilevel model did not appear to converge appropriately. Analyses comparing various subscales on both instruments revealed high correlations frequently, though some smaller correlations did exist where they might be expected, thus demonstrating divergent validity. The CCAPS Distress Index and the OQ Total score provide *very* similar information. Further, the CCAPS subscales do provide some additive information beyond the general measure of distress. Thus, college counseling centers can consider other practical and psychometric factors in deciding which instrument to use, based on their center, clientele, and institution.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to thank my chair, advisor, and mentor, Lauren Weitzman, for all she has taught me so far and for being a resource in so many ways. I want to thank Zac Imel for his statistical knowledge and general approach to cochairing my dissertation, and to all of my committee members (Rob Davies, Lois Huebner, and A.J. Metz) for their feedback on this project and their role in my development throughout my time in graduate school. I am also thankful for the University of Utah Counseling Center staff, who all contributed to this project in various ways and have been a part of my life for so many years. I am thankful for Stacy Ackerlind, Kari Ellingson, Erika Hill, and so many others from my time in Suite 270, along with the staff at Weber State University's Counseling and Psychological Services, who have contributed to my professional development and personal growth. I appreciate the support, encouragement, and prayers from my parents, brother, and extended family (who continue to love me regardless of how busy I get or how far away I move), friends in the Counseling Psychology program (who understood what I was going through), and friends outside of the program (who reminded me that there was life beyond school). Finally, I am thankful for my husband, who put up with so much and always pushed me to continue. I am excited for what the future holds for us and I look forward to continued personal and professional relationships with so many of the people who have gotten me to this point.

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Evaluating the effectiveness of psychotherapy services is a vital component of maintaining quality of care. Agencies are increasingly utilizing client ratings of mental health symptoms to evaluate treatment effectiveness. Although this is an important advance, the performance of widely disseminated psychotherapy outcome measures is not clear. Specifically, agencies and clinicians may use different measures to evaluate treatment, and the extent to which these measures provide similar answers in regard to patient response is questionable. Moreover, it is not clear if newly developed measures designed to provide increased diagnostic specificity (e.g., measures of substance abuse and depression) actually provide more specific information than general measures of psychological distress.

To provide a context for evaluating current, widely used psychotherapy outcome instruments, I will provide a brief introduction into the history of monitoring treatment response in mental health, outline salient contextual influences that prompt the assessment of psychotherapy outcome, discuss well-established criteria for effective psychotherapy outcome measurement, and review commonly used psychotherapy outcome instruments. Finally, I will identify gaps in knowledge specifically regarding the

convergent validity of two widely disseminated measures, the Counseling Center Assessment of Psychological Symptoms (CCAPS; Center for Collegiate Mental Health [CCMH], 2012) and the Outcome Questionnaire (OQ; Lambert et al., 2004).

Once this groundwork has been laid, I will propose a rationale for examining the CCAPS and OQ using a large administrative database from a college counseling center that administered both measures to clients at every encounter. Multivariate multilevel modeling techniques, a contemporary method in psychotherapy outcome research, will be described as a methodology to examine the convergent validity of change on these two measures.

History of Psychotherapy Outcome Efforts

According to Lambert and Lambert (1999), "Outcome assessment is a branch of applied psychology that illuminates the strength of the effects of psychological interventions on patient functioning" (p. 115). Early efforts at measuring outcome were primarily theory-based, whether Freudian dynamic, client-centered, behavioral, or cognitive, and little to no research existed to support the use and interpretation of these measures (Lambert & Lambert, 1999). More recently, efforts to determine the best treatment for various diagnoses as well as changes in reimbursement requirements within managed care organizations have dramatically increased the use of outcome assessment and research (Beutler, 2001).

Psychotherapy Outcome in Clinical Practice

Three important influences for practitioners to measure psychotherapy outcome are external pressures, ethical obligations, and as a means to improve clinical service delivery. External pressures related to receiving funding and/or payment for clinical services have become increasingly salient in clinical practice over the last two decades (Bishop, 1995; Cormier & Nurius, 2003; Lambert & Lambert, 1999). One relatively early example of funding being tied to outcome assessment is the Community Mental Health Centers Amendment of 1975, which required program evaluation for federally funded agencies (Larsen, Attkisson, Hargreaves, & Nguyen, 1979). Tanner and Stacy (1985) mentioned that this government mandate was an impetus for increases in the use of, and research on, client satisfaction measures. More recently, many managed care organizations have imposed an expectation that practitioners will empirically demonstrate that their services are beneficial (Cormier & Nurius, 2003). The effect of third-party payment on measuring psychotherapy outcome is demonstrated in the frequency with which this reason is mentioned in articles (for a small sampling, see Callaghan, 2001; Deane, 1993; Holcomb, Beitman, Hemme, Josylin, & Prindiville, 1998). In a university and college counseling center (UCC) context, outcome assessment may not be a requirement, but it can provide data to support the importance of the counseling center in the institution of higher education and to advocate for funding (Bishop, 1995). Thus, whether demonstrable outcomes are required or preferred, many practitioners are assessing outcomes because of these external demands.

Another source of external pressure is the consumer rights movement (Deane, 1993; Ogles, Lunnen, & Bonesteel, 2001). For example, in a hospital setting with patients

with mental illness, the patient's voice was not valued previously, but the patient's voice is now being taken more seriously (Powell, Holloway, Lee, & Sitzia, 2004). From a consumer rights philosophy, the client expects to get better, and clinicians need to demonstrate that clients are getting what they pay for (Jacobson, Roberts, Berns, & McGlinchey, 1999). Further, outcome data are of value to consumers to determine which type of treatment may best meet their needs (Callaghan, 2001).

Outcome assessment is also important because of the ethical obligation to provide effective treatment. Cormier and Nurius (2003) argued that clinicians cannot meet this ethical obligation without assessing client outcome, because a formal assessment provides a more complete and less biased picture of client change than the therapists' opinion alone (Corrigan, 1990; Larson et al., 1979; Moore & Kenning, 1996). Others have expressed a similar idea, that assessments provide a way of confirming that clients are receiving adequate services (Bieschke, Bowman, Hopkins, Levine, & McFadden, 1995; Moore & Kenning, 1996).

Further, outcome assessment is important because it provides beneficial information for practitioners to improve their clinical service delivery. LaSala (1997) mentioned that assessing services is consistent with the values of individuals in the helping profession. Kendall, Holbeck, and Verduin (2004) highlighted how outcome assessment is a vital source of feedback for the practitioner, as it can be used to adjust treatment or suggest alternatives. Without this feedback, practitioners do not have needed information to improve treatment (Cormier & Nurius, 2003). Steenbarger and Smith (1996) described this process as a "continuous feedback loop in which services are delivered, evaluated, modified, and redelivered" (p. 148). Recent research has

documented that the act of monitoring client change improves outcome (Lambert, Harmon, Slade, Whipple, & Hawkins, 2005).

Properties of an Effective Psychotherapy Outcome Measure

To assess psychotherapy outcome (both for clinical use and for research), psychologists and other professionals cannot use *any* psychological measure and assume that change in the right direction is equivalent to a good outcome. There are practical guidelines for choosing a measure, as well as statistical considerations (i.e., psychometric properties). As Hill and Lambert (2004) noted, the results of outcome measures and research on outcomes can be greatly affected by the psychometric properties of the instrument(s) utilized. The American Psychological Association (APA) *Ethical Principles of Psychologists and Code of Conduct* (2002) instructs psychologists to use only measures with established validity and reliability.

Related to the practical considerations when selecting an instrument as an outcome assessment, Groth-Marnat (2003) identified numerous criteria: (a) brief to complete (less than 15 minutes); (b) specifically pertinent for outcomes (i.e., not a full battery for describing or diagnosing); (c) relevant to the group on which the instrument will be used (based on age and other client characteristics); (d) "usable and understandable" (p. 580) for professionals and nonprofessionals both; (e) supported by research that demonstrates that the measure changes in psychotherapy; and (f) backed by strong psychometric properties. Sound psychometric properties, as expanding upon and specified by Newman, Ciarlo, and Carpenter (1999) include, "a) reliability (test-retest, internal consistency, or interrater agreement where appropriate); b) validity (content,

concurrent, and construct validity); c) demonstrated sensitivity to treatment-related change; and d) freedom from response bias and non-reactivity (insensitivity) to extraneous situational factors that may exist" (p. 160). The establishment of each of these psychometric properties should be considered an ongoing process, with the more points of quality evidence that are accumulated, the stronger the support for the use of the measure (Groth-Marnat, 2003).

Validity

Validity refers to how much the evidence supports the specified interpretation of a given test—that the inferences made based on scores are justifiable (Crocker & Algina, 1986). As set forth in the *Standards for Educational and Psychological Testing* by the American Educational Research Association (AERA), APA, and National Council on Measurement in Education (NCME; 1999), validity is based on a *test score interpretation*; thus, validity support is needed for any type of outcome determination (e.g., "better"), even if other interpretations of the test have validity support (e.g., diagnostic validity). While different types of validity evidence are described below, it is important to consider them as a whole. This concept is particularly true in light of how these classifications have changed over time (Bold & Rounds, 2000) as they are now all largely considered to fall under the broad term "construct validity" (so much so that AERA, APA, and NCME [1999] considers the term "construct validity" to be redundant).

At a basic level, a measure to be used for psychotherapy outcome should relate to what the therapist and/or client want to change in psychotherapy. As discussed above, this hoped-for change may be general or specific; however, the measure should contain a

significant portion of the construct desired (*construct representation*; Messick, 1995). For example, if psychotherapy is for both depression and anxiety equally, a valid outcome assessment would not look solely at depression. The specific items should reflect the construct (*content validity*; Cronbach & Meehl, 1955): for example, items on a measure of depression should not be exclusively about negative thoughts if depression is also conceptualized as behavioral, emotional, and physical.

Validity can also be supported by convergent and divergent validity. In *convergent validity*, two measures or scores that are thought to be similar are tested to see if they are indeed similar. In *divergent validity*, two measures or scores that should not overlap are tested to see if they are indeed dissimilar. For example, if a researcher is trying to validate a new measure of extraversion (considered a stable trait), this measure should *not* relate to how hungry the participant is at the time of taking the measure: the measure of hunger should *not* correlate with scores on the extraversion scale.

Another aspect of validity is the sensitivity and specificity of the measure, which may often be overlooked but should be given careful consideration (Groth-Marnat, 2003). *Sensitivity* refers to the measure correctly identifying true positives (e.g., correctly identifying someone in treatment as distressed; Groth-Marnat, 2003). *Specificity* refers to the measure correctly identifying true negatives (e.g., correctly identifying someone not in treatment as not distressed; Groth-Marnat, 2003). In order to make this classification, a cutoff score is used, such as those described below related to clinical significance.

Reliability

Reliability is the consistency or replicability of an individual's scores when tested with the same or alternate test forms in similar circumstances (Crocker & Algina, 1986). Strong reliability is essential in the measurement of outcomes, because error is compounded in a change score (Hill & Lambert, 2004). There are several types of reliability: *internal consistency, alternate forms, test-retest,* and *interrater reliability*.

To understand reliability, it is important to understand the theory behind why scores might vary. Classical test theory (CTT) suggests that, in an ideal world, test scores would accurately represent the individual on the construct being measured—one's true score (Osterlind, 2006). However, some level of error in measurement will always exist (Groth-Marnat, 2003). Generalizability Theory augments CTT by dividing the types of error and providing ways to estimate the different types (Shavelson, Webb, & Rowley, 1989). Errors in measurement may be systematic or random. Systematic errors are variations in scores that do not relate to the construct being measured yet always affect an individual's score or group's scores in the same way (Crocker & Algina, 1986). For example, if a measure of life satisfaction utilizes language that requires a high reading level, then the measure is systematically affected by the respondent's reading level in addition to life satisfaction. Random error is caused by "chance happenings" (Croker & Algina, 1986, p. 106) and affects an individual's performance in an unpredictable way (Osterlind, 2006), such as motivation or environmental distractions. The standard error of measurement (SEM) is used to account for random error by providing a range from the observed score in which the true score is likely to fall. Thus, by calculating and reporting

the various types of reliability, scale developers and researchers are providing information about what type of error and what level of error that might be present.

Internal consistency relates to the relationship between items on a single administration (AERA et al., 1999). For example, on the measure of depression, one would expect that there is not a wide variety in the responses to specific items (unless, of course, the items are representing different aspects of depression, in which case there may be more variability). Internal consistency is frequently measured by Cronbach's α, which considers the variance of each item and the number of items on the instrument (Crocker & Algina, 1986).

Test-retest reliability is a measure of how much an individual's score may fluctuate over time and across repeated administration without intervention (Groth-Marnat, 2003). Test-retest reliability is particularly important for outcome assessment, as an individual completes the assessment more than once, and clinicians need to have a clear picture of how much of the change can be attributed to the intervention and how much is fluctuation that might occur even without the intervention.

Alternate form reliability would need to be established if there were different versions of the same assessment, especially if the client would complete more than one version in the course of psychotherapy. Without this measure of reliability, the professional could not know how the results compared across forms (Groth-Marnat, 2003). *Interrater reliability* would need to be established if the instrument was completed by observers, in order to confirm that the observers were rating observations in a similar manner. Most outcome measures do not utilize alternate forms or outside observers; thus, these two types of reliability are not discussed related to specific instruments.

Psychometric properties for psychotherapy outcome measures

There are several important psychometric considerations that are unique to psychotherapy outcome measures. For example, test-retest may be enough to establish reliability for an assessment that an individual would only take one or two times under normal circumstances, but outcome measures may be given multiple times in the course of psychotherapy, and thus reliability needs to be established across repeated administrations in the absence of psychotherapy. In addition, individuals' scores on the measure *should* change when receiving an intervention: this characteristic is known as *sensitivity to change* (Hill & Lambert, 2004). Based in part on the example of the OQ by Vermeersch and colleagues (2000, 2004), the criteria for a score to be sensitive to change are as follows:

1) Slope is in the correct direction (meaning that the person is getting better, not worse);
2) Slope is significantly different from zero (meaning that the person is getting significantly better)
3) Slope is significantly greater for treated than for untreated individuals (meaning that the person is getting better faster than an untreated person).

Clinical significance

Psychologists recognized decades ago that a statistically significant difference between treated individuals and untreated individuals (even with large effect sizes) is not enough to determine that a treatment is effective (Jacobson, Follette, & Revenstorf, 1984; Jacobson & Truax, 1991). In addition to these statistical comparisons, *clinical*

significance is a concept that refers to treated individuals returning to "normal" (Jacobson et al., 1999; Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999) and also relates to how convincing the change is (Kendall et al., 1999). Clinical significance provides one means of interpreting score changes.

"Normal" or "functional" versus "clinical" or "dysfunctional" are terms generally used in discussions of clinical significance, though these terms have significant conceptual implications beyond what may be explicitly stated. What is considered normal? Is it *not* meeting diagnostic criteria for a particular disorder, or is "normal" based on the general population (Kendall et al., 1999)? What level of symptoms is still normal and how does this vary based on the specific symptom(s) under consideration? For example, much of the general population experience symptoms associated with depression or anxiety at low levels, but any experience of hallucinations would be considered dysfunctional (Kendall et al., 1999). Is normal different for more chronic conditions (e.g., schizophrenia), such that a positive treatment outcome does not mean being symptom-free but a predefined reduction of symptoms (Jacobson & Truax, 1991)? With these questions and the unique purpose of each study, some variations in the definition of normal will occur across studies (Jacobson et al., 1999).

In spite of these variations, more accurate and complete meaning on the benefit of treatment can be made across studies by utilizing one of the specific mathematical ways to determine clinical significance. One commonly accepted method is to determine a cutoff score in which scores are statistically more likely to be part of the dysfunctional or functional populations (Jacobson & Truax, 1991). When the variances for the two groups are equal, the cutoff score places the individual either closer to the mean of the normal

population or closer to the mean of the clinical population. A cutoff score can be determined with unequal variances as well (Jacobson & Truax, 1991). Other options include scores at least two standard deviations from the mean of the clinical population or scores within at least two standard deviations from the mean of the normal population (Jacobson & Truax, 1991). Published averages are often based on scores falling within *one* standard deviation of the mean (Kendall et al., 1999). Kendall and colleagues (1999) proposed steps for equivalency testing, which provides a statistical test to determine that the scores for the treated individual are equivalent to scores for the normal population. Tingey, Lambert, Burlingame, and Hansen (1996) suggest using multiple normative samples for social validation, such that a positive outcome is movement from one sample to another. Methods utilizing a normative comparison group provide the advantage of determining clinical significance based on information independent of the sample of treated individuals (Kendall et al., 1999)—an *external* standard (Jacobson & Truax, 1991).

In addition to returning to normal functioning, Jacobson and colleagues proposed that, for a change to be clinically significant, the change must also be reliable. The *Reliable Change Index* (RCI) is a measure of how much change has occurred (Jacobson & Truax, 1991), and thus whether the change is "real" or possibly due to measurement error. The formula is the pre-post test difference divided by the standard error of differences, with a RCI greater than 1.96 meaning that the change is sufficiently large to exceed the margin of measurement error (Jacobson & Truax, 1991).

The criteria of returning to normal functioning and being reliable are combined by Jacobson and Truax (1991) to create four possible classifications: *recovered* (when the

change is greater than the RCI and the post score is within the normal range), *improved but not recovered* (when the change is greater than the RCI but still in the abnormal range), *no change* (if the change is not greater than the RCI), and *deteriorated* (if the change is greater than the RCI and the post score is further from normal). An article by Vonk and Thyer (1999) provides one example of the use of clinical significance specifically in a UCC. They administered the SCL to clients at intake and termination, and then presented support for the effectiveness of short-term treatment in a UCC by utilizing both statistically significant change and clinically significant improvement.

Some validity support for the construct of clinical significance exists. In a study by Ankuta and Abeles (1993), clients who had clinically significant change were more satisfied with psychotherapy and self-reported greater benefit from psychotherapy than those with nonsignificant change or no change. In another by Lunnen and Ogles (1998), perceived change, satisfaction, and the strength of the therapeutic alliance were all higher for clients with clinically significant change than for those who had no change or deteriorated.

While the construct of clinical significance has some validity support and is popular in both clinical and research settings, it does have limitations. For example, reliable change will be more easily achieved for those who have a higher level of pathology and thus greater opportunity for change (Mintz & Keisler, 1982). On the other end of the spectrum, clients who seek psychotherapy but have levels of distress on a given measure already below the cutoff value cannot possibly meet a definition of clinically significant change that includes moving from the dysfunctional to the functional range. Floor and ceiling effects may restrict an individual determination of

clinical significance or even the likelihood of clients as a group reaching clinical significance on a given measure (Lunnen & Ogles, 1998).

The value of the Reliable Change Index has several additional criticisms. Some (e.g., Hsu, 1999) have suggested that, instead of utilizing raw scores, residualized change scores should be used to increase reliability. However, research on different methods of calculating RCI generally yield consistent results (McGlinchey, Atkins, & Jacobson, 2002). RCI does take into account random error by using test-retest reliability as noted above; however, it does not account for systematic error that would be present at equal levels in a test-retest reliability study. Further, RCI assumes that the random error is consistent for everyone, though it is conceivable that clients at one end of the distress spectrum would have a different value for random error than clients at the other end.

In summary, clinical significance is useful in *clinical practice* because of its ease to understand and interpret at the individual level. It has the advantage in *outcome research* of being a standard index across different outcome measures. However, there are concerns about the validity of the index. One important disadvantage of clinical significance in outcome research is that it is a categorical variable and thus reduces power to detect differences. In Jacobson and colleague's definition of clinical significance, the categorical variable is actually a combination of two variables: the amount of change (and whether that change is statistically reliable), and whether the threshold of dysfunctional-to-functional has been crossed. Further, clinical significance in itself does not provide a test of statistical significance across values of a variable of interest (e.g., treatment A versus Treatment B) but instead provides a computational analysis of each value of the variable separately (e.g., the percent of clients in each category of clinical

significance in Treatment A). Thus, clinical significance should not be used as the only analysis in psychotherapy outcome research.

Additional methods of interpreting change

Research on psychotherapy outcomes may use standard statistical tests and effect sizes to make comparisons between treatments. However, these methods do not provide information at the individual level during treatment. Clinical significance can be used to evaluate an individual's outcome at any given time but takes into account only the initial and current scores and not the trajectory of change over time. Additional methods of tracking and interpreting change in psychotherapy are emerging in the field. For example, Lueger and colleagues (2001) developed expected treatment response values based on theoretical underpinnings and client characteristics. As another example, Finch, Lambert, and Schaalje (2001) developed expected recovery curves that they generated based on the outcomes observed in a large data set. Beutler (2001) emphasized that a quality assurance system must be able to identify clients for whom psychotherapy is not working. Clinicians and researchers continue to evaluate these various methods, both for their clinical usefulness and their statistical support.

Specific Psychotherapy Outcome Measures

Psychotherapy outcome can be and is conceptualized in a variety of ways. Froyd, Lambert, and Froyd (1996) used the term *content* to describe the topic of the assessment, whether intrapersonal, interpersonal, or social. A clinician could choose to look at general levels of distress (e.g., OQ [Lambert et al., 2004]) or consider a specific domain (e.g.,

depression with the Beck Depression Inventory [BDI; Beck, Steer, & Brown, 1996]). Outcome could also be determined by the presence (or absence) of a diagnosis (Kendall et al., 2004). Cormier and Nurius (2003) provided a framework for measuring change in *goal behaviors* that is highly individualized, by using the dimensions of frequency, duration, magnitude (intensity), and occurrence of these behaviors. The following provides an overview of specific instruments commonly used to assess psychotherapy outcome. Each of these instruments is currently used in UCCs. Details about their content, development, and psychometric properties are provided so that readers can begin to evaluate and compare these instruments. This review also highlights gaps in the literature particularly related to the comparability of change in scores across instruments.

Symptom Checklist-90-R (SCL)

The SCL is a measure of experienced symptoms, designed to be appropriate as a one-time assessment of symptoms with clinical or nonclinical populations, as well as with repeated administrations to assess change in psychotherapy (Derogatis, 1994). It consists of 90 items that contribute to nine primary symptom dimensions (Somatization, Obsessive-Compulsive, Interpersonal Sensitivity, Depression, Anxiety, Hostility, Phobic Anxiety, Paranoid Ideation, and Psychoticism) and three global indices (Global Severity Index, Positive Symptom Distress Index, and Positive Symptom Total).

The original instrument developed was the Hopkins Symptom Checklist (HSCL). Derogatis (1994) described how the SCL was further developed to be clinically useful as a self-report measure. He also emphasized that the constructs in the instrument be both consistent with their use in the literature and supported by empirical findings. Support for

the internal structure included a factor analysis using 1,002 psychiatric outpatients: almost all items loading correctly on the theorized dimensions (Derogatis, 1994). Further, this structure remained consistent in multiple studies across various populations (e.g., gender and social class).

The SCL and its scales have been compared to numerous instruments and their scales to provide convergent and discriminant validity data. These instruments include the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1940), the Center for Epidemiological Studies Depression Scale (Radloff, 1977), the Hamilton Rating Scale for Depression (Hamilton, 1967), and the General Health Questionnaire (Goldberg, 1972). Derogatis (1994) summarized the studies: there were generally high correlations for like constructs and low correlations for less similar constructs. Construct validity for the SCL has preliminary support for its sensitivity and specificity. Specifically, the clinical cutoff score for the Global Severity Index is 62/63 (nonclinical/clinical), based on large clinical and nonclinical samples. This cutoff has demonstrated acceptable levels of sensitivity and specificity among medical patients but is still being explored in other settings (Derogatis, 1994).

Reliability for the nine symptom dimensions of the SCL are based on three different studies (Derogatis, 1994). Internal consistency reliability was determined by utilizing 209 “symptomatic volunteers” (Derogatis, Rickels, & Rock, 1976) and 103 psychiatric outpatients (Horowitz, Rosenberg, Baer, Ureno, & Villasenor, 1988). Values for α ranged from .77 (on psychoticism for the volunteers) to .90 (on depression for both sets of results). Test-retest reliability was determined utilizing 94 psychiatric outpatients, with reliability ranging from .78 (on hostility) to .90 (on phobic anxiety) over a 1-week

period. Test-retest reliability with elapsed time of 10 weeks for the 103 psychiatric outpatients ranged from .68 (somatization) to .83 (paranoid ideation).

The SCL has been utilized extensively as an outcome measure. Results across numerous studies demonstrate that clients' scores on the SCL *do* change in psychotherapy, and the change is greater than that of controls. This change can be found across the distress continuum (mild to severe) and for a variety of treatment interventions (Derogatis, 1994). Two studies (Schmitz, Hartkamp, & Franke, 2000; Schauenberg & Strack, 1999) have calculated clinical significance values based on samples in Germany. Todd, Deane, and McKenna (1997) presented research comparing SCL scores among adolescents, undergraduate college students, and adults, and the authors discussed the implications that these differences have on the interpretation of clinical significance. No information about clinical significance is included in the administration manual (Derogatis, 1994).

Beck Depression Inventory (2nd ed.; BDI) and the Beck Anxiety Inventory (BAI)

The BDI (Beck et al., 1996) and the BAI (Beck & Steer, 1996) are both 21-item self-report measures of symptoms. The BDI was developed specifically to assess for symptoms of depression as delineated by the DSM-IV (APA, 1994). As such, cutoff scores are based not on statistical differences between clinical and nonclinical populations but instead on the presence or absence of a diagnosis, and the manual includes suggestions for adjusting these cutoff scores based on desired levels of specificity and sensitivity (Beck et al., 1996). The development of the BAI involved two

cycles of administering pilot instruments, analyzing the results (including factor and diagnostic validity analyses), and subsequently reducing the number of items to result in the current version (Beck & Steer, 1996).

Validity and reliability support for the newest version of the BDI is based on two different groups: 500 psychiatric outpatients from four settings, and 120 college students from an introductory psychology class (Beck et al., 1996). Validity research on the BDI includes desired convergent and discriminant validity with several other measures. Factor analysis revealed two factors, though these factors differed somewhat between the clinical and nonclinical groups (Beck et al., 1996). Specifically, the factors for the clinical group were Somatic-Affective and Cognitive, but the factors for the nonclinical group were Cognitive-Affective and Somatic. Reliability evidence demonstrates the BDI's internal consistency: Cronbach's α was .92 for the clinical group and .93 for the nonclinical group. In addition, item-total correlations were significant on every item in both groups, even after adjusting for the multiple statistical tests (Beck et al., 1996).

Validity and reliability support for the BAI is based on a sample of 393 outpatients diagnosed with mood and anxiety disorders and a nonclinical sample of 243 people from three different settings (Beck & Steer, 1993). Validity research on the BAI includes desired convergent and discriminant validity with several other measures. Reliability evidence demonstrates the BAI's internal consistency: Cronbach's α ranges from .85 to .93 based on type of anxiety disorder diagnosis (Beck & Steer, 1993).

Outcome Questionnaire – 45.2 (OQ)

The OQ measures change in psychotherapy as its primary purpose (Lambert et al., 2004). As such, it is designed to be administered multiple times during the course of psychotherapy. To maximize its utility as an outcome measure, the developers aimed to create an instrument that was brief, sensitive to change in psychotherapy, available for a relatively low cost, and characterized by strong reliability and validity support (Lambert et al., 2004). It has three subscales: Symptom Distress (22 items), Interpersonal Relations (11 items), and Social Role Functioning (nine items), which together cover a wide range of symptoms of mental disorder in the adult population and measure a person's overall level of distress (Lambert et al., 2004). The OQ has been developed utilizing several different normative groups, including an undergraduate student population in a classroom setting, UCC clients, and inpatient samples (Lambert et al., 2004).

In scale development for the OQ, items were selected that could occur across a variety of specific disorders and complaints that would likely affect an individual's quality of life (Lambert et al., 2004). Confirmatory factor analysis (CFA) was conducted for three different models: the three subscales as three separate factors; the symptom distress subscale as one internal factor, and the interpersonal and social role subscales collapsed as one external factor; and all items as one factor. Based on a sample of 1085 people (from multiple settings), all three of these models were sufficient models (Lambert et al., 2004). Because the three scales correlate highly, the OQ may be best interpreted as one overall measure of distress (Lambert et al., 2004). However, some research (e.g., Kim, Beretvas, & Sherry, 2010) does not support either a one-factor or a three-factor structure, thus calling into question what, exactly, the OQ is measuring.

Convergent validity for the OQ demonstrates high correlations (all significant beyond the .01 level of confidence) between the OQ Total score, the subscale scores, and 11 different measures considered a counterpart for one or more of the OQ subscales, based on a sample of 157 nonclinical college students (Lambert et al., 2004). Clinical samples from three different settings (n = 183), were given the OQ and three measures comparable to one of the OQ subscales (the General Symptom Index of the SCL-90 [Derogatis, 1994], The Inventory of Interpersonal Problems [Horowitz et al., 1988], and The Social Adjustment Scale [Weissman & Bothwell, 1976]). Correlation coefficients were all significant at the .05 level (Lambert et al., 2004). Based on these findings, Lambert and colleagues (2004) determined that the OQ Total score could be viewed as an overall level of distress, but the interpretation of the subscale scores—with somewhat lower correlations to instruments hypothesized to cover similar content—is less certain. In addition to the convergent validity, research has provided evidence for divergent validity. Specifically, Durham and colleagues (2002) found that only 0.7% of OQ score variance across time could be attributed to the Marlowe-Crown Social Desirability Scale (Crowne & Marlowe, 1960) and the Test Taking Survey (developed for the study to assess mechanical responding).

Construct validity for the OQ is supported by its sensitivity and specificity. Specifically, the expected differences between clinical and nonclinical samples do exist, and unexpected differences between same-type samples (both clinical or both nonclinical) do not exist (Lambert et al., 2004). In addition, one study demonstrated that the OQ correctly identifies people as either clinical or nonclinical about 83% of the time (Lambert et al., 2004). The overall cutoff score utilized for the OQ is 63/64, where 63 is

nonclinical and 64 is clinical; the cutoff scores for the subscales are as follows: Symptom Distress = 36/37, Interpersonal Relations = 15/16, and Social Role Functioning = 12/13 (Lambert et al., 2004). These cutoff scores are based on community nonpatient sample data and multiple-site outpatient sample data as described elsewhere.

Reliability properties for the OQ include internal consistency, test-retest, and repeated-administration. Internal consistency is based on a nonclinical college student sample of 157 and a clinical employee assistance program sample of 298, with Cronbach's α ranging from .70 (on the social role scale) to .93 (for the overall score). The student sample was also used for test-retest reliability, with Pearson product-moment correlation coefficients ranging from .78 (on symptom distress) to .84 (for the overall score). An additional sample of 56 nonclinical college students were given the OQ 10 times over 10 weeks, with the correlation decreasing over each administration, to .66 for the correlation between Week One and Week Ten (Lambert et al., 2004). In another study (Durham et al., 2002), college students completed the OQ weekly, biweekly, monthly, or two times only, in a 9-week period. Across frequency of administration, the largest drop occurred between the first and second administrations, and it was not a clinically significant drop, and little change occurred in scores after the second administration.

Research generally supports the OQ as an outcome measure, in that it is sensitive to change in psychotherapy. In one study utilizing 5,553 treated individuals from counseling centers and 248 untreated college students (Vermeersch et al., 2004), 43 of the items met the first criterion—the slope was in the correct direction. Of these, 35 items met the second criterion—the slope was significantly different from zero. Most of these (34 items) also had slopes significantly greater than the untreated individuals. All three

subscales and the Total score met all three criteria. In addition, there was a large effect size for group differences (between slopes of clinical and nonclinical samples) of the Total score and the symptom distress subscale, and a medium effect size for group differences of the interpersonal relations and symptom distress subscales and 15 of the items. The reliable change index (RCI, as defined by Jacobson and Truax [1991]) was determined based on clinical samples used above. The RCIs are as follows: Total = 14, Symptom Distress = 10, Interpersonal Relations = 8, and Social Role Functioning = 7 (Lambert et al., 2004).

Counseling Center Assessment of Psychological Symptoms (CCAPS)

The CCAPS is an instrument designed and normed specifically for counseling centers to use with the college student population (Locke et al., 2011). Its developers aimed for the CCAPS to be both statistically sound and clinically useful, with the intent of using it locally and nationally for research and evaluation in addition to its clinical use (Locke et al., 2011). It was *not* designed to provide diagnostic information. There are two versions currently in use: the CCAPS-62 and the CCAPS-34 (CCMH, 2012). The CCAPS-62 has 62 items across eight subscales (ranging from 5 to 12 items per scale): Depression, Generalized Anxiety, Eating Concerns, Social Anxiety, Hostility, Family Distress, Substance Use, and Academic Distress. The CCAPS-34 retains 34 of the 62 items across seven of the original eight subscales (ranging from four to six items per scale). In addition to the subscale scores, the CCAPS includes a Distress Index score, which pulls 19 items from multiple scales to provide a value for a client's general psychological functioning (CCMH, 2012). The Distress Index is comprised of the same

19 items on both the CCAPS-62 and the CCAPS-34. The developers of the CCAPS emphasize that the instrument is multidimensional and the scales provide unique information about the *ways* in which a client is distressed, rather than just *how* distressed the client is; thus, they encourage continued use of the scales, in addition to the Distress Index (CCMH, 2012). Current clinical norming data consist of 59,606 students seeking counseling at 97 colleges and universities in 2010-2011 (CCMH, 2012). Given that the CCAPS is still in an earlier stage of instrument development, data supporting its initial reliability and validity will be described here in more detail.

The CCAPS was originally developed by a team of professionals at a UCC (Locke et al., 2011). This team identified 11 common concerns for college students, generated 167 items, and gathered data on these items from 113 students in an undergraduate subject pool. The instrument was modified based on factor analysis and item loading, to shorten to 101 items. Data were then gathered from 2,155 students seeking services, and another factor analysis reduced the instrument to 70 items and nine factors (CCAPS-70). Then 52 counseling centers utilized the instrument and pooled their data from 22,060 students seeking services. Factor analysis and item loading resulted in the CCAPS-62 with its eight factors. Generally, the factor loadings remained consistent across the studies to refine the measure, thus providing support for the robust nature of the factors (Locke et al., 2011). The 34 items on the CCAPS-34 were determined utilizing "advanced statistical techniques combined with input from counselors to create a maximally-sensitive short version of the CCAPS" (CCMH, 2012). These statistical techniques included both classical test theory and Item Response Theory (Locke et al., 2012). Two changes to the subscales from the CCAPS-62 to the CCAPS-34 are the

following: 1) the Family Distress subscale is not included and 2) the Substance Use subscale becomes the Alcohol Use subscale (Locke et al., 2012). CCMH developed the Distress Index by examining a second-order factor model, a bifactor model, and a total score for statistical fit and clinical merit, resulting in the selection of the bifactor model (CCMH, 2012).

As shown in Table 1.1, correlations between subscales on the CCAPS-62 range from 0.05 (Social Anxiety and Substance Use) to 0.66 (Anxiety and Depression), based on the clinical data utilizing administrations of the CCAPS-70 (Locke et al., 2011). In this sample, all of the correlations were statistically significant, in part due to the large sample size. The highest correlations were between the Depression subscale and four other subscales, along with the Anxiety subscale and two other subscales. Confirmatory factor analysis on the CCAPS-34 resulted in a similar pattern of intercorrelations between subscales in a sample of 482 undergraduate students (Locke et al., 2012).

Convergent validity was assessed using data from 499 students from a subject pool who were given the CCAPS-62 and nine other instruments: one referent measure for each of the eight scales, plus the Marlowe-Crown Social Desirability Scale-Short Version (MCSD; Reynolds, 1982)—to determine if any of the scales were too highly correlated with social desirability (Locke et al., 2011). Results indicated that the Pearson product-moment correlations were highest between each subscale and its referent additional measure. In addition, while all of the correlations between the subscales and the MCSD were statistically significant, they were relatively weak (Locke et al., 2011). Locke and colleagues (2012) conducted a similar study using the CCAPS-34 and the same additional instruments, with similar results: subscales correlated highest with the appropriate

Table 1.1

Correlation Between Subscales on the CCAPS-62

Scale	1	2	3	4	5	6	7	8
Depression	--							
Eating Concerns	0.36	--						
Substance Use	0.18	0.19	--					
Generalized Anxiety	0.66	0.30	0.19	--				
Hostility	0.56	0.25	0.24	0.5	--			
Social Anxiety	0.54	0.27	0.05	0.44	0.31	--		
Family Distress	0.38	0.21	0.08	0.33	0.39	0.25	--	
Academic Distress	0.59	0.22	0.17	0.45	0.35	0.31	0.23	--

referent measure. McAleavey and colleagues (2012) conducted a third study using the CCAPS-62 and the same additional instruments, in a clinical population, and again found that CCAPS subscales correlated highest with the appropriate referent measure.

The reliability of the instrument has some positive support. Internal consistency of the CCAPS-62 is based on data from 499 students in a subject pool (Locke et al., 2011). Cronbach's α for each of the subscales is as follows: Depression = .91; Eating Concerns = .88; Substance Use = .85; Generalized Anxiety = .85; Hostility = .86; Social Anxiety = .82; Family Distress = .81; and Academic Distress = .78. Reported internal consistency for a clinical population, along with the means and standard deviations, for each of the subscales on the CCAPS-62 and the CCAPS-34 can be found in Table 1.2 (CCMH, 2012).

Table 1.2

Descriptive Statistics for Each CCAPS Subscale

	CCAPS-62 (*N* = 59,606)			CCAPS-34 (*N* = 9,560)		
CCAPS Scales	Mean	*SD*	Alpha	Mean	*SD*	Alpha
Depression	1.58	0.93	0.91	1.53	1.03	0.88
Generalized Anxiety	1.60	0.92	0.85	1.81	1.00	0.83
Social Anxiety	1.81	0.95	0.84	1.77	1.00	0.82
Academic Distress	1.85	1.02	0.82	1.88	1.12	0.82
Eating Concerns	1.00	0.88	0.90	0.99	1.16	0.89
Family Distress	1.28	0.96	0.83	N/A	N/A	N/A
Hostility	1.04	0.87	0.86	0.92	0.86	0.84
Substance Use/ Alcohol Use	0.76	0.87	0.84	0.67	0.91	0.83
Distress Index	1.64	0.84	0.92	1.64	0.84	0.92

Test-retest reliability has been assessed in a general student sample for both the CCAPS-62 and the CCAPS-34 (Locke et al., 2011; Locke et al., 2012). Students from a subject pool completed one of the assessments and then completed that same version either 1 or 2 weeks later. Pearson product-moment correlation coefficients displayed in Table 1.3 ranged from r = .78 to r =.93 on the CCAPS-62 at 1-week, r = .76 to r = .92 on the CCAPS-62 at 2-weeks, r = .79 to r =.87 on the CCAPS-34 at 1-week, and r = .74 to r = .86 on the CCAPS-34 at 2-weeks.

Table 1.3

Test-Retest Reliability for Subscales in CCAPS-62 and CCAPS-34

CCAPS Scales	CCAPS-62		CCAPS-34	
	1-week (n = 46)	2-week (n = 52)	1-week (n = 86)	2-week (n = 47)
Depression	0.927	0.917	0.866	0.864
Generalized Anxiety	0.782	0.842	0.857	0.850
Eating Concerns	0.893	0.896	0.815	0.771
Social Anxiety	0.826	0.888	0.851	0.805
Hostility	0.907	0.834	0.813	0.751
Substance Use/ Alcohol Use	0.866	0.900	0.792	0.781
Academic Distress	0.923	0.759	0.794	0.742
Family Distress	0.920	0.914	N/A	N/A

According to the *CCAPS 2013 Clinician's Guide* (CCMH, 2013), the CCAPS instrument may be used as a therapeutic outcome measure (pre-post change), utilizing the more informative CCAPS-62 at initial appointment and at termination. It can also be used for treatment monitoring (session-to-session change) with the shorter CCAPS-32 (CCMH, 2013). The *CCAPS 2012 Technical Manual* (CCMH, 2012) includes reliable change indices for each subscale on both the CCAPS-62 and the CCAPS-34. It also includes clinical cutoff scores to distinguish between those who are more similar to a clinical population and those who are more similar to a nonclinical population.

McAleavey and colleagues (2012) generated these scores utilizing the formula recommended by Jacobson and Traux (1991) for determining cutoff scores in overlapping populations. The sample consisted of 15,873 college students who completed the CCAPS-62 and indicated either that they were not receiving any treatment or who were in counseling at their college. While these two groups were statistically significantly different on all of the subscales except substance use, the authors noted that there was a high level of overlap in distributions between the two groups and thus that the clinical cutoff scores should be interpreted with caution. With the data and analyses currently available, more information is needed particularly about the CCAPS as a psychotherapy outcome measure, including sensitivity to change.

Research Questions and Rationale

Each of the instruments described above has been developed with a focus on best practice guidelines for scale construction and adherence to sound psychometric properties. However, all instruments have limitations, and vary in their psychometric strengths and weaknesses as well as general assessment characteristics (e.g., clinical utility for a given population). The OQ has been extensively researched as an outcome measure and has been used in multiple settings with vast data now available about how clients change over time. On the other hand, the CCAPS has unique advantages both psychometrically and practically. It was developed and normed specifically for the college student population; it is available free of charge for counseling centers; and is already integrated into Titanium Schedule, the electronic management system utilized by many counseling centers. Further, the CCAPS was selected as part of the standardized

data set for the Center for Collegiate Mental Health (CCMH), a national collaborative research center whose goal is "brining science and practice together" (CCMH, 2010, p. d) by gathering this standardized data at counseling centers across the country and utilizing it for multiple purposes. With the CCMH network, researchers will continue to gather large amounts of data on the CCAPS, strengthen the norming data, and provide these data back to participating centers for use in clinical practice, as well as informing public policy, higher education administrators, and other constituencies. However, the CCAPS, as a relatively new instrument, is less familiar to many counseling center practitioners than the OQ and the clinical meaning of change in scores on the CCAPS is less clear.

Given these factors, the purpose of this study is to test the comparability of the OQ and the CCAPS as psychotherapy outcome measures. Are they providing similar or unique information about clients? If the two measures correlate highly within clients, then the measures are providing similar information, whereas if the two measures do not correlate highly within clients, then the measures are providing unique information. In order to make comparisons between the OQ and the CCAPS, the CCAPS Distress Index will be used. Thus, one aspect of the present research will answer the question "Does the CCAPS Distress Index work as a general measure of distress and of psychotherapy outcome vis a vis the OQ Total score?" The first step to address this research question is to test the convergent validity of OQ Total scores and CCAPS Distress Index scores. I hypothesized that these two scores will be highly correlated (Hypothesis 1), such that a client who scores high on one measure will also score high on the other measure, whereas a client who scores low on one will also score low on the other. Second, it is important to test the correlation of change in OQ Total score with change in CCAPS Distress Index

score across clients. It is hypothesized that change in OQ Total score will be highly correlated with change in CCAPS Distress Index score (Hypothesis 2), such that a client who changes a great deal on one measure will also change a great deal on the other measure, whereas a client who changes a little on one will only change a little on the other.

CHAPTER 2

METHODS

Participants

The sample consists of counseling clients at a university counseling center (UCC) in a large public institution located in the mountain west region. The client population consists of 55.9% females, 42.8% males, and .5% transgender. Individuals reported their race/ethnicity as 78.3% Caucasian/White, 6.1% Hispanic/Latino(a), 4.4% Asian American/Asian, 2.6% multiracial, and less than 2% African American/Black, American Indian/Alaskan Native, Native Hawaiian/Pacific Islander, or other. Undergraduate students make up 65.9% of the client population (20.1% senior, 20.0% junior, 14.5% sophomores, and 11.3% freshman), while 27.3% are graduate students and 2.1% are faculty or staff. For this research project, we utilized data provided by 2,320 clients for 16,779 sessions between January 2011 and May 2013. The mean number of sessions was 7.23, median was 4, with a range of 1 to 109 sessions.

Measures

Counseling Center Assessment of Psychological Symptoms (CCAPS)

The Counseling Center Assessment of Psychological Symptoms (CCAPS; Center for Collegiate Mental Health [CCMH], 2012) is a relatively new instrument that was selected by CCMH as part of the standardized data set to be utilized by participating counseling centers across the country. The CCAPS-62 consists of eight scales and 62 items, while the CCAPS-34 is reduced to seven scales and 34 items (Appendix A lists all items by subscale). The Distress Index consists of 19 items (consistent across both versions of the CCAPS) and provides a measure of general psychological distress (CCMH, 2012). The instruments were developed using a rational-empirical approach, with factor analysis supporting the items on each scale (Locke et al., 2011). Test-retest reliability for the CCAPS-34 ranges from .707 (Academic Distress) to .843 (Eating Concerns) at a 1-week interval, and from .768 (Academic Distress) to .825 (Social Anxiety) at a 2-week interval, in two samples of students in a nonclinical setting (CCMH, 2012).

Outcome Questionnaire – 45.2 (OQ)

The Outcome Questionnaire -45.2 (OQ; Lambert et al., 2004) was specifically designed to measure change in psychotherapy (Lambert et al., 2004). It consists of 45 items and three subscales and can be found in Appendix B. Based on factor analysis and convergent validity results, the three subscales have some support, but the Total score has the strongest psychometric support (Lambert et al., 2004). It is reliable—with both

internal consistency and repeated-measures reliability, and sensitive to change in psychotherapy (Lambert et al., 2004). A more in-depth description of its development, validity, and reliability is included in Chapter 1.

Procedure

As standard practice, clients at the UCC where the data were collected are asked to complete both the CCAPS-62 and the OQ, along with the rest of the CCMH standardized data set and other questions, prior to intake. Clients who return for individual or group counseling are asked to complete both the CCAPS-34 and the OQ before every session. Respondents complete both instruments either on paper or electronically. All information is stored in Titanium Schedule and the OQ Analyst for both clinical and research purposes, and these programs calculate scores for each of the scales of the OQ and CCAPS. Before beginning multilevel modeling, I divided the scores for the OQ Total score by 45 to create a mean item score. This calculation does not alter the distribution of the scores but does place the OQ Total score on the same scale as each of the CCAPS scale scores: 0 to 4. The consistent scoring across measures allows for a clearer interpretation of the results that would otherwise be challenging if the scores were on different scales (Baldwin et al., 2014). Based on the guideline by Speer and Newman (1996) that 90% of the items on a measure of psychotherapy outcome should be completed to be considered valid, I coded a measure as missing for a given administration if it was missing more than 10% of items (more than 5 on the OQ, 6 on the CCAPS-62, or 3 on the CCAPS-34). Further, if the CCAPS Distress Index or the OQ Total was a score of zero, then that measure was considered missing for that occasion.

This step was included because of the questionable meaning of a score of zero, particularly in a clinical sample. Specifically, of the 16,779 measurement occasions, 10 OQ Total scores were zero, of which 7 also had a score of zero on the CCAPS Distress Index (the other scores were .05, .05, and .15). There were 97 times the CCAPS Distress Index score was zero: 7 with OQ Total scores of zero, only 1 with an OQ score in the clinical range, and 15 without valid OQ scores. With these two exclusion criteria and occasions where one or both instruments were not complete, the final dataset consisted of a total of 13,450 valid OQs and 14,818 valid CCAPS. One of the advantages of using multilevel modeling (the analysis method for this research, as described below) is that missing data do not necessitate that the client be excluded altogether from the analysis (Hox, 2010) and the model can accommodate having only one outcome measure at a given time point as well as having different numbers of total measurement occasions.

Data Analysis

Analyses utilized multivariate multilevel modeling. Multilevel modeling takes into account the hierarchical nature of the variables wherein lower level scores are considered "nested" within higher level scores (Hox, 2010). In the multivariate multilevel model used in this research, the focus is on occasions (sessions) nested within clients, and thus, the model allows for the two outcome variables to be correlated for each person (Baldwin et al., 2014). Regarding sample size, there are not clear standards for a minimum number of individuals, only that the sample size be large enough to provide an accurate estimate of the parameters and for any asymptotic characteristics to be revealed (MacCallum et al., 1997).

In a standard univariate multilevel model for linear change, the model is the following:

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_i + v_i x_{it} + e_{it} \qquad (1)$$

where y_{it} is the response variable y for individual i at occasion t; β_{0i} is the mean intercept; β_1 is the mean slope; x_{it} is the measure of time for individual i at occasion t; u_i is the random variation of the intercept for individual i; v_i is the random variation of the slope for individual i; and e_{it} is the residual error for individual i at occasion t. The results of this model include estimates for each of the following: the fixed effect intercept (β_0) and slope (β_1); random effects variances for intercept ($\sigma^2_{u0} = \text{var}(u_{0i})$), slope ($\sigma^2_{u1} = \text{var}(u_{1i})$), and residual ($\sigma^2_e = \text{var}(e_{it})$); and random effects covariance of the intercept and slope ($\sigma_{u01} = cov(u_{0i}, u_{1i})$). Random effects are the additional terms in multilevel modeling that allow for the dependence of observations (Raudenbush & Bryk, 2002); in this study, random effects are the variability between individuals. When creating two separate univariate multilevel models for two outcome measures, the equations would be the following:

$$y_{1it} = \beta_{10} + \beta_{11} x_{it} + u_{1i} + v_{1i} x_{it} + e_{1it} \qquad (2)$$

$$y_{2it} = \beta_{20} + \beta_{21} x_{it} + u_{2i} + v_{2i} x_{it} + e_{2it} \qquad (3)$$

To combine these formulas to include both outcome variables in one model, two dummy variables are created (one for each outcome variable) and the data are organized in a *long* format, rather than more commonly recognized wide format (Baldwin et al., 2014). In this format, data are treated as if there is only one outcome value (MacCallum et al., 1997), and then the dummy variables are coded as 1 in the column for the measure

from which the score came and as 0 in the column for the other measure. This multivariate structure results in the following equation:

$$y_{kit} = \beta_{10}h_i + \beta_{20}j_i + \beta_{11}x_{it}h_i + \beta_{21}x_{it}j_i + u_{1i}h_i + u_{2i}j_i + v_{1i}x_{it}h_i + v_{2i}x_{it}j_i + e_{1it}h_i + e_{2it}j_i \quad (4)$$

where *k* indices the outcome variable (either CCAPS or OQ); *h* and *j* are the two dummy variables (where *h* =1 for CCAPS and 0 for OQ, and *j* = 1 for OQ and 0 for CCAPS); and the remaining formula consists of the combined univariate models (Baldwin et al., 2014). The results of this model include separate estimates for each outcome measure of two fixed effects and four random effects as in the univariate model; it also provides estimates of the random effects of the covariances across each pair of the outcome measures' slopes and intercepts (MacCallum et al., 1997).

The index of the correlation of scores within a client is provided by the intraclass correlation (ρ). The intraclass correlation is defined as the proportion of variance explained by the client (Raudenbush & Bryk, 2002) and was computed as the ratio of the variance of the client random effects to the total variance (the sum of the residual variance and the variance of the client random effects). In this study, a higher intraclass correlation indicates that more of the overall variance is unique to client, and less variance is unaccounted-for differences between the CCAPS and the OQ.

In multilevel modeling with longitudinal data, time can be measured using real time (e.g., number of days since first session) or ordinal positions (MacCallum et al., 1997). In this study, the session number was used as an ordinal representation of time. Session number is the common way to measure time in treatment outcome, particularly as treatments occur at each session number. Real time would provide information about days since first session but would provide less direct information regarding how much

treatment the person had received, and this could vary widely from person to person and even within persons.

To test Hypothesis 1—that client differences on the OQ and the CCAPS were highly correlated, I utilized a two-level, empty (e.g., no predictors) multivariate multilevel model with a random effect for clients (i.e., a random intercept). Multivariate multilevel models provide information about correlation between outcomes across multiple levels of analyses. The outcomes were the OQ Total score and the CCAPS Distress Index score. In the initial model, repeated administrations of the OQ and CCAPS were "nested" within clients. Thus, the model provided information about 1) the variability of OQ Total and CCAPS Distress Index scores within clients (e.g., how much the scores change for a particular client), and 2) the correlation of variability in OQ Total and CCAPS Distress Index scores within clients.

To test this hypothesis, I used Bayesian mixed effects models fit via the MCMCglmm package in R (Hadfield, 2010; R Core Development Team, 2012). This procedure simulates parameters for each model using Markov chain Monte Carlo procedures (MCMC). These simulated parameters are called the posterior distribution and provide point estimates and highest posterior density (HPD) intervals. HPD intervals are the Bayesian equivalent of confidence intervals. In this case, the primary parameter of interest is the correlation of random effects. The model utilized a noninformative prior distribution, meaning that the parameter values were weakly constrained, which is standard in Bayesian analyses and does not influence substantive results (Gelman & Hill, 2007). The MCMC chain consisted of 50,000 iterations, including a burn-in of 5,000 iterations, and thinning interval of 20. I used the mode and the 95% HPD interval of the

simulated posterior distribution to determine the correlation between the OQ and CCAPS scores.

To test Hypothesis 2—that client changes on the OQ and the CCAPS were highly correlated, a second multivariate multilevel model included session number as a parameter at level one. This model also included a random effect for session number at the client level (level two; e.g., a random slope). This random effect allowed for an estimate of variability across clients in how much change occurs across sessions (e.g., some clients could have scores that decrease more than other clients). Thus, the models provided an estimate of the correlation between clients' differences in OQ Total score change and clients' differences in CCAPS Distress Index score change. I again employed Bayesian analyses to test the hypothesis.

As secondary analyses, I explored the relationship between each CCAPS scale score and the OQ Total score, along with select pairs of subscales between instruments, using multilevel modeling. These exploratory models had the potential to indicate that certain scales provide unique information beyond the relationship between the CCAPS Distress Index score and the OQ Total score.

CHAPTER 3

RESULTS

The mean of the 14,818 valid CCAPS Distress Index scores was 1.59 (*SD* = .74); the mean of the 13,450 valid OQ Total scores was 1.58 (*SD* = .55). Figure 3.1 is a scatterplot of the two scores with box and whisker plots included along the axes for both scores. I also calculated the Pearson product-moment correlation between the two measures. The OQ Total and CCAPS Distress Index scores were strongly correlated, $r(12{,}811) = 0.900$, $p < .01$. This result does not take into account the dependency of observations within persons, but provides a rough initial exploration of the similarity of the CCAPS and OQ.

Hypothesis 1

To determine if the OQ Total score and CCAPS Distress Index score are highly correlated, we examined the correlation of the client-level random effects for the OQ and CCAPS using a multivariate multilevel model. The mode of the posterior distribution of the correlated random effects was $r = 0.967$, *HPD*[.962, .971]. The correlation was very large, and the HPD interval did not include zero, providing strong evidence that client-level differences in the distress indices for these measures are highly correlated.

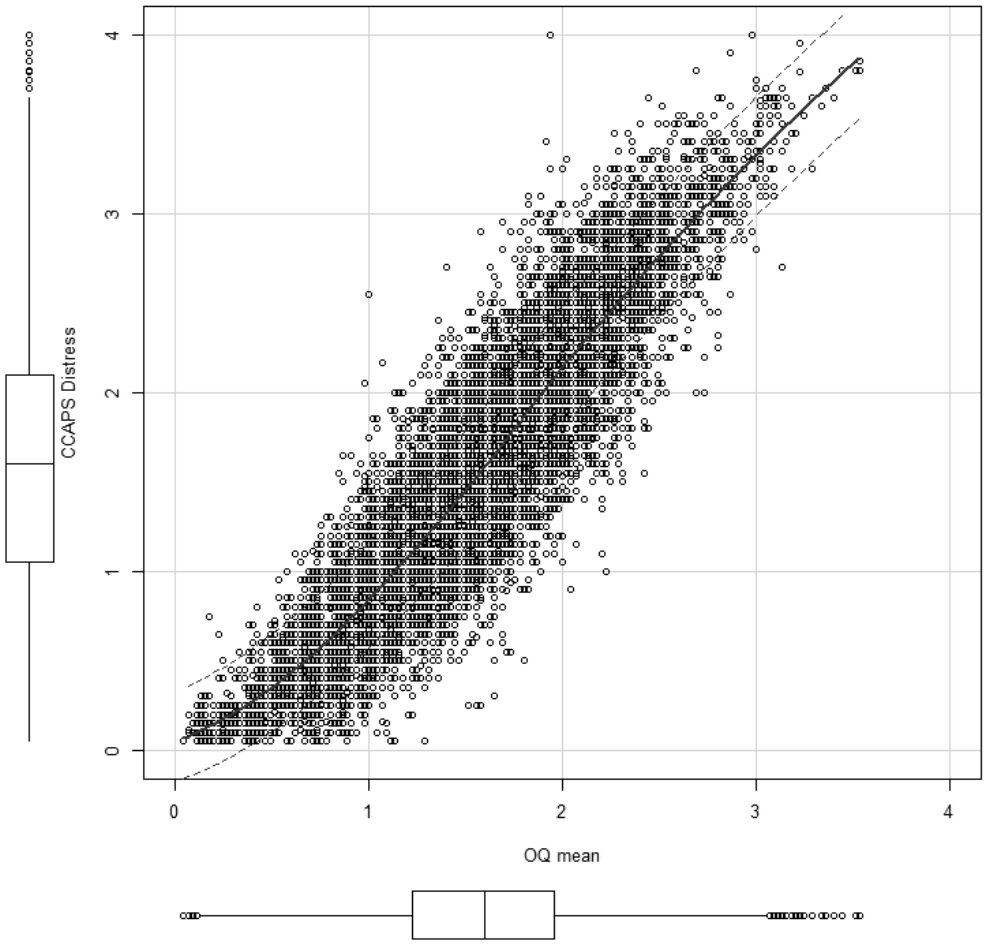


Figure 3.1 – Scatterplot of CCAPS Distress and OQ Total Scores

To further understand the convergent validity of these two measures, I explored one way in which they may be used in a clinical setting: determining if a given score is in the clinical or nonclinical range (using the cutoff score for each measure in which a person is statistically more likely to be in one group versus the other). The cutoff score for the CCAPS Distress Index is 1.21, which resulted in 26.92% of the clients being nonclinical and 73.05% being clinical. The cutoff score for the OQ Total score (mean) is 1.40, which resulted in 29.45% of the clients being nonclinical and 70.55% being clinical.

Table 3.1 contains the percentage of clients based on their classification on both the OQ and the CCAPS. Most clients (88.46%) were categorized as either clinical on both measures or nonclinical on both measures, though the McNemar's test (a type of chi-square analysis with paired samples) still revealed significant differences, χ^2 (1, N = 1837) = 9.552, p = .002.

Hypothesis 2

For the second hypothesis, the goal was to examine the correlation of client level differences in OQ and CCAPS distress index change over time. We selected a subset of the full dataset as the complete dataset includes clients at various stages of treatment (i.e., some clients began counseling before January 2011), whereas the subset of data included only clients who started treatment after January 2011. The reason for this step is that it allows the session number (as defined within the dataset) to reflect their treatment session number accurately and consistently, which is important when considering how the two measures change over the course of psychotherapy. The original dataset consisted of 16,779 sessions from 2,320 clients; after excluding the clients whose first session (within

Table 3.1

Percentage of Clients Based on Nonclinical and Clinical Categorizations on the OQ and the CCAPS

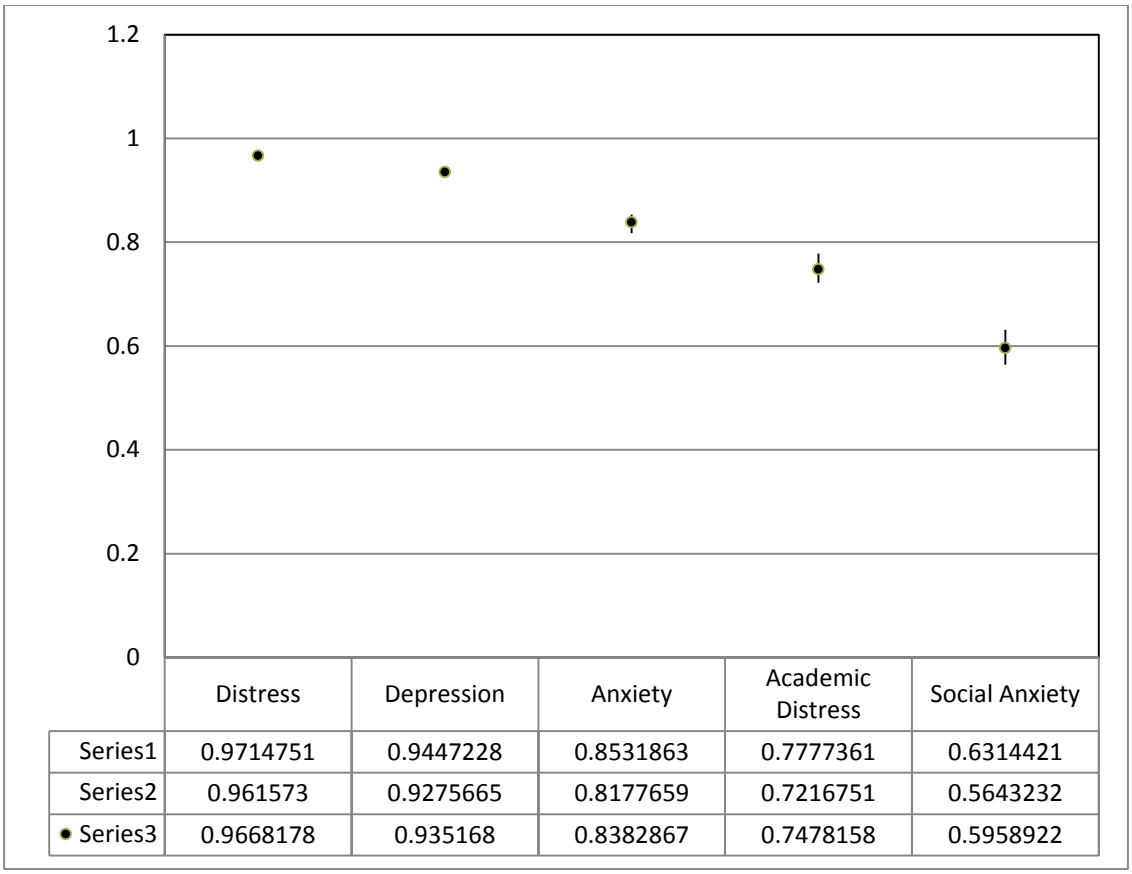
		OQ	
		Nonclinical	Clinical
CCAPS	Nonclinical	22.43%	4.52%
	Clinical	7.02%	66.03%

the dataset) was not intake, the resulting dataset consisted of 11,481 sessions from 1,745 clients.

Unfortunately, when we included session number and a random slope for session number allowing person-level variability over time, the multivariate multilevel model did not appear to converge appropriately. We explored several different modeling strategies, including the standard maximum likelihood multilevel modeling package lme4 and Bayesian models. Results were not consistent. When using lme4, it appeared that the variance between persons in slopes was very low after accounting for between-person differences. Because of the difficulty of fitting a correlation between two variables when the variability is small, the model did not provide meaningful results. When using Bayesian models, the model also had trouble converging, as evident in the diagnostic plots from the posterior distributions (i.e., "poor mixing" in the time series trace plot), which suggests that the model had difficulty converging on a value. Thus, I have decided not to report the results for these analyses.

Additional Exploratory Analyses

Additional analyses compared each of the CCAPS scale scores to the OQ Total score. Figure 3.2 presents the Bayesian Model results, including mode and HPD for the correlations of the OQ Total score with the following CCAPS scale scores: Distress Index (for comparison), Depression, Anxiety, Academic, and Social Anxiety. In addition to higher correlations on the CCAPS Depression and Anxiety scales, the HPD intervals for each of these correlations are tighter. While these two scales are particularly correlated with the OQ Total score, the correlations between the OQ Total score and both

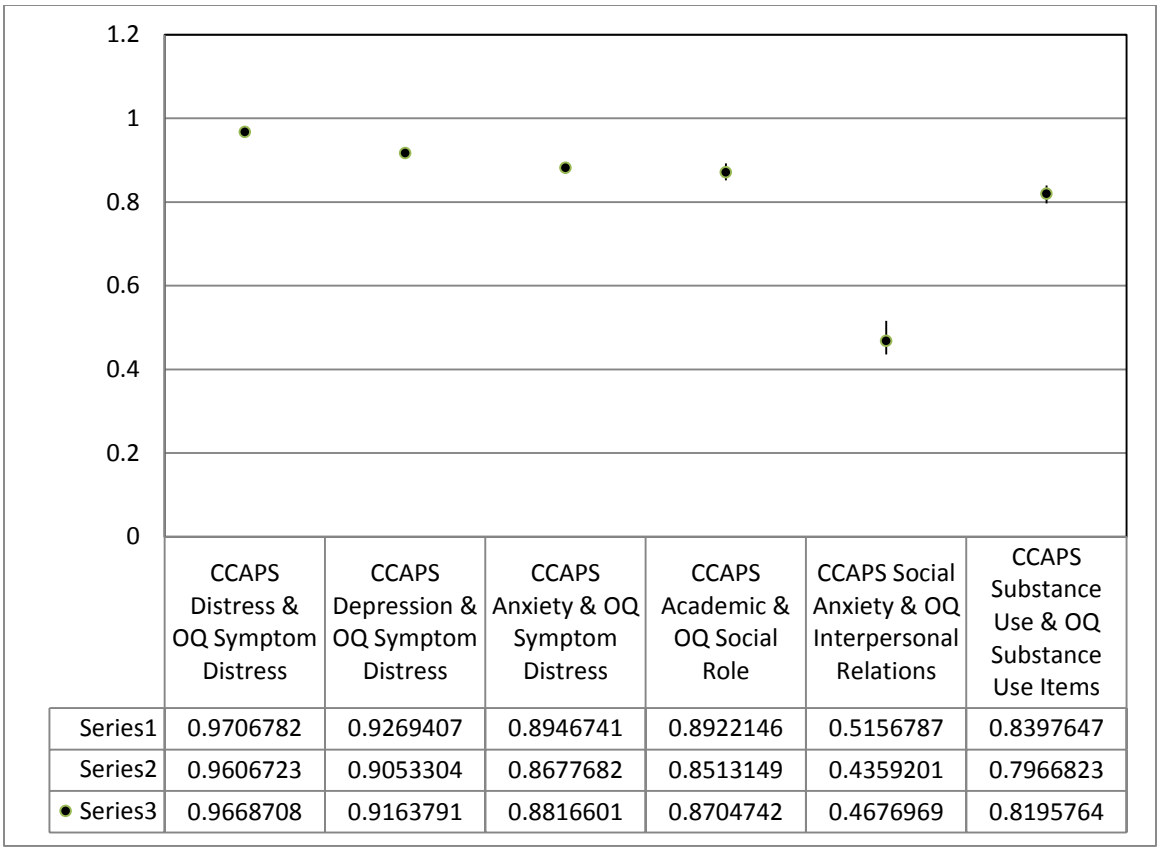


	Distress	Depression	Anxiety	Academic Distress	Social Anxiety
Series1	0.9714751	0.9447228	0.8531863	0.7777361	0.6314421
Series2	0.961573	0.9275665	0.8177659	0.7216751	0.5643232
Series3	0.9668178	0.935168	0.8382867	0.7478158	0.5958922

Figure 3.2 – Bayesian Model Results for OQ Total Score and Select CCAPS Subscales
Note: Series 1 is the upper limit of the highest posterior density (HPD) interval, Series 2 is the lower limit of the HPD interval, and Series 3 is the mode of the distribution.

the Academic and Social Anxiety scales of the CCAPS are also high and do not include zero, suggesting that both are correlated with the OQ Total score. The scales Hostility, Eating Concerns, and Substance Use are not included because the distribution is so skewed for each that a model correlating these with a normal distribution is not warranted. The distributions remained skewed even after taking the log of the score, a typical strategy when dealing with nonnormal distributions (MacCallum et al., 1997).

To further explore the relationship between the CCAPS and the OQ instruments, I ran Bayesian models with select pairs of OQ subscales and CCAPS subscales that are conceptually similar to each other. These results are presented in Figure 3.3. Specifically,



	CCAPS Distress & OQ Symptom Distress	CCAPS Depression & OQ Symptom Distress	CCAPS Anxiety & OQ Symptom Distress	CCAPS Academic & OQ Social Role	CCAPS Social Anxiety & OQ Interpersonal Relations	CCAPS Substance Use & OQ Substance Use Items
Series1	0.9706782	0.9269407	0.8946741	0.8922146	0.5156787	0.8397647
Series2	0.9606723	0.9053304	0.8677682	0.8513149	0.4359201	0.7966823
Series3	0.9668708	0.9163791	0.8816601	0.8704742	0.4676969	0.8195764

Figure 3.3 – Bayesian Model Result for Select OQ Subscale and CCAPS Subscales Pairs
Note: Series 1 is the upper limit of the highest posterior density (HPD) interval, Series 2 is the lower limit of the HPD interval, and Series 3 is the mode of the distribution.

the OQ Symptom Distress subscale score is highly correlated with the CCAPS Distress Index, Depression, and Anxiety subscale scores, along with tighter HPD intervals around the mode. The pair OQ Social Role – CCAPS Academic Distress also resulted in high correlation and a fairly tight HPD interval around the mode. On the other hand, the pair OQ Interpersonal - CCAPS Social Anxiety was not as high, suggesting that these two subscales are measuring conceptually different constructs. After I calculated a mean for the OQ items related to substance use, the Bayesian model revealed a fairly high correlation between this mean and the CCAPS Substance Use subscale.

CHAPTER 4

DISCUSSION

The purpose of this study was to provide information about the comparability of the OQ and the CCAPS as a psychotherapy outcome measure. An element of this inquiry is about how the CCAPS Distress Index performed as a general measure of distress in psychotherapy. I address this question in Hypothesis 1, using multilevel modeling to examine the correlation between the OQ Total and CCAPS Distress Index scores as repeated measures within clients. The model revealed very high correlation between the CCAPS Distress Index and the OQ Total scores within clients, which gives strong support that the CCAPS Distress Index does provide *very* similar information to the OQ Total score. In an effort to better understand this correlation, I examined the items for a face validity comparison of content. Of the 20 items that comprise the CCAPS Distress Index, 11 have an item on the OQ that asks about a similar symptom or concept. These items from both instruments can be found in Table 4.1. The other nine items do not have a similar item on the OQ. Despite the unique items, the high correlation suggests that the two measures are so similar that they are basically redundant. It may be that, while both measures do not ask about the exact same symptoms (e.g., panic attack), the underlying construct(s) are consistent across instruments (e.g., anxiety).

Table 4.1

CCAPS Distress Index Items with Similar OQ Items

CCAPS Item	**OQ Item**
I am unable to keep up with my school work	I am not working/studying as well as I used to
I am not able to concentrate as well as usual	I have difficulty concentrating
I feel isolated and alone	I feel lonely
I feel sad all the time	I feel blue
My heart races for no good reason	My heart pounds too much
I have sleep difficulties	I have trouble falling asleep or staying asleep
I feel tense	I have sore muscles
I get angry easily	I feel irritated
I am afraid I may lose control and act violently	I feel angry enough at work/school to do something I might regret
I feel worthless	I feel worthless
I have thoughts of ending my life	I have thoughts of ending my life

A statistically significant difference was found in the chi square analysis that examined the classification of clients into clinical and nonclinical groups on the two measures. Clients were more likely to be in the "clinical on the CCAPS but nonclinical on the OQ" group than in the "clinical on the OQ but nonclinical on the CCAPS" group. This statistical significance may be attributed, at least in part, to the large sample size. Another contributing factor may be the different groups used as norming samples. Specifically, the CCAPS used college students in counseling and college students not in counseling to make the determination of clinical versus nonclinical, while the OQ used samples of adults in a range of psychotherapy and psychiatric treatment services (including in-patient care) and adults in the general population. This greater range in the clinical norming sample (compared to the CCAPS) could mean that clients need to be

"more distressed" before being classified as more similar to the clinical population than the nonclinical population. Despite this small difference in clinical classification, most clients (over 90%) were classified congruently: either as nonclinical on both measures or as clinical on both measures. Thus, the CCAPS Distress Index does have convergent validity with an established outcome measure of general distress.

To further assess the comparability of the OQ and the CCAPS as a psychotherapy outcome measure, I attempted to confirm that the relationship between the OQ Total and the CCAPS Distress Index scores stayed consistent throughout the course of psychotherapy (Hypothesis 2). As noted above, the multilevel model including session number did not work, providing inconsistent results across methods. One possible explanation is that the high correlation between the two scores at the client level limits the ability for the models exploring correlation in change over time to fit.

The exploratory analyses compared pairs of the CCAPS subscales scores with OQ Total and OQ subscales scores. The CCAPS Depression, Anxiety, and Academic Distress subscales correlate more highly with the OQ Total score, suggesting that these subscales are more similar to general distress. This is not surprising particularly for the Depression and Anxiety subscales, given that 12 of the 20 items of the CCAPS Distress Index are pulled from these two subscales. The CCAPS subscale Social Anxiety was correlated, while the Hostility, Eating Concerns, and Substance Use subscales were so skewed that a valid comparison with the OQ Total score was not possible. In further investigating relationships between OQ and CCAPS subscales, there seems to be much overlap between overall distress (as measured by the OQ Total score and the CCAPS Distress Index) and symptoms (as measured by the OQ Symptom Distress and CCAPS

Depression and Anxiety subscales). Academic distress and functioning are related to overall distress but less so than depression and anxiety symptoms. These results suggest that the CCAPS subscales do provide some additive information beyond the general measure of distress. These analyses were one step in the process of understanding the CCAPS subscales, and the extent to which the subscale scores reflect "conceptually and psychometrically distinct domains" as the Center for Collegiate Mental Health (CCMH) purports (CCMH, 2012, p. 9).

Limitations

One limitation of this study is that the data were from counseling center clients at one university. There are many ways in which the sample is consistent with other UCCs, which is demonstrated by similar demographic characteristics in the Center for Collegiate Mental Health Annual Report (CCMH, 2013), including more women than men, and predominately White and heterosexual. Differences do exist between this sample and both what may be present at other UCCs and other research being done to validate the CCAPS. Specifically, this sample consisted of a larger proportion of graduate students and faculty and staff: 28% of this sample, compared to 14% of the CCMH Annual Report (CCMH, 2013). Further, the university in the present study has a larger nontraditional population even among undergraduate students. Combined, these two differences are reflected in the age of the sample, with an average age of 25.6 years old. It is unknown how these differences might have affected the results in a way that would not be generalizable to other centers.

Another limitation, as is true in any real-world research, is the missing data, particularly the missing administrations of the OQs. It was standard practice and policy at the counseling center during data collection for all clients to complete both the OQ and the CCAPS before every individual counseling session; however, a percentage of clients (19.20%) did not have an OQ associated with their counseling appointment. It is possible that clients were more likely to take the CCAPS rather than the OQ if they only completed one instrument, or that counselors were more likely to notice that the CCAPS was missing and ask their clients to complete it. Another possibility is that there was an error in the system in connecting OQs to appointments; because the CCAPS is administered and stored in the same software program that is used for scheduling, this potential problem is unique to the OQ. A review of the data did not reveal any specific trends for how or why clients did not have an OQ on so many occasions, which supports the conclusion that there is not a systematic reason for the missing data. While multilevel modeling is better able to accommodate missing data points than other standard analyses, the large percentage of missing OQ administrations is a limitation for the current study.

Future Research

The next step in this research will be to further investigate the current data set and what is occurring when session number is added to the multilevel model (Hypothesis 2). I hope to understand why the different approaches to fitting the model provided contradictory results and to see if there is a way to fit the model. Is it that the high correlation between the two measures when treated as repeated measures is limiting the ability of the model to fit with an additional variable? If so, is there a way to adjust the

model to take this into account and still provide results about change over time? Another possibility is that there is something about the data set that limits the model fit, such as the variability of scores at first session or the variability of client change over time. It may be that the model needs to include an interaction between the number of sessions and rate of change (slope), as previous research has demonstrated that rate of change is not constant across total dose of psychotherapy (Baldwin, Berkeljon, Atkins, Olsen, & Nielsen, 2009). This line of research will provide information about the relationship between the two measures over the course of psychotherapy.

Another important direction for continuing to examine and improve the utility of the CCAPS as a psychotherapy outcome measure is to generate recovery curves with existing data, and use these to create predictive recovery curves. One method of doing so is described by Finch and colleagues (2001). These predictive recovery curves have important utility in the practice of psychotherapy: they can provide information to clinicians and clients about expected recovery and deviations from expected recovery. Research has shown that this feedback can actually improve outcomes, when using the OQ (Lambert et al., 2005), so research could investigate whether this finding is consistent when using the CCAPS.

Additional research can further enhance our understanding of the relationship between client general distress and domain-specific concerns, particularly in looking at change during the course of psychotherapy. This study included only beginning exploratory analyses using the CCAPS subscales (beyond the CCAPS Distress Index) and the OQ subscales (Symptom Distress, Interpersonal Relations, and Social Role Performance). Further research comparing the CCAPS subscales and the OQ subscales

may provide valuable information about the overlap of the two measures, as well as about the general versus specific nature of client concerns. Beutler (2001) briefly mentioned that the number of dimensions of a psychotherapy outcome measure has an impact on the amount of information available not only about complex client concerns but also about how clients might differentially change on different dimensions. Research along these lines could increase our understanding of the nature of change during psychotherapy when clients present with general distress versus when they present with domain-specific concerns. For example, if a client presents with disordered eating, what does change look like on the CCAPS Eating Concerns subscale, and what does it look like on the CCAPS Distress Index or the OQ Total score? What are the rates of change for these separately, and how does the relationship between domain-specific concerns and general distress change or stay consistent throughout psychotherapy? The answers would have treatment implications as well; if general distress is highly related to eating concerns, then a more broad approach to treatment may be useful, but if the relationship is not strong, a more focused approach may be warranted. This example considers eating concerns, but similar research questions and clinical implications are applicable for each of the CCAPS subscales.

Implications for Practice in University Counseling Centers

In recent years, many UCCs have been reevaluating which outcome measure to use to track client change in counseling. The rapid development of the CCAPS and its implementation in many UCCs has prompted much debate. The use of the CCAPS has been advanced by its availability in the Titanium software and its selection as part of the

standardized data set by CCMH. However, many people question the wide-spread adaptation of the instrument without more research on its validity and reliability, particularly as a psychotherapy outcome measure.

The high correlation between the CCAPS Distress Index and the OQ Total scores provide support for the use of the CCAPS as a psychotherapy outcome measure. The extremely high correlations between the CCAPS Distress Index and the OQ Total score suggest that it is superfluous to administer both measures as was done during the data collection for the current study. This finding also suggests that much of the research that utilizes the OQ would be consistent with CCAPS data. For example, research has found that when the OQ is utilized to provide session-to-session feedback to counselors, clients experience improved outcomes (Lambert et al., 2005). It is very likely that developing a similar software package for the CCAPS that provides counselors with feedback based on the CCAPS would similarly improve client outcomes.

Conclusion

This study answers the important question of how the OQ and CCAPS compare statistically as psychotherapy outcome measures. As mentioned above, there are many factors to take into consideration when selecting a psychotherapy outcome measure (e.g., cost, length). The high correlation found in this study between the CCAPS Distress Index and OQ Total scores suggests that the measures are essentially interchangeable. The choice of a counseling center-based psychotherapy outcome measure will thus need to be guided by different factors. To illuminate this dilemma, it may be useful to conceptualize the decision of counseling centers to administer the OQ or the CCAPS as analogous to a

gymnastics meet. Here, the result of one event—the statistical properties of the two measures as psychotherapy outcome measures—may be viewed as a tie, necessitating turning to other events (e.g., practical considerations and research base) to determine a "winner." Advantages of the CCAPS include the ease to administer and score through Titanium software, the focus in development and norming on the college student population specifically, and the inclusion of both a measure of general distress and domain-specific information. Advantages of the OQ-45 include a solid research history that supports its sensitivity to change in psychotherapy, along with strong reliability and validity properties. Different UCCs may "judge" these "events" differently based on their center, clientele, and institution, but this research provides a final score for one aspect of the decision.

APPENDIX A

CCAPS-62 AND CCAPS-34 ITEMS BY SUBSCALE

CCAPS Version		Scale	CCAPS Item	Reverse Scored	Distress Index
34	62				
	8	Depression	I feel disconnected from myself		
4	9		I don’t enjoy being around people as much as I used to		Yes
5	10		I feel isolated and alone		Yes
	12		I lose touch with reality		
11	20		I feel worthless		Yes
12	23		I feel helpless		Yes
	28		I am enthusiastic about life	Yes	
	37		I have unwanted thoughts I can't control		
21	40		I feel sad all the time		Yes
25	46		I have thoughts of ending my life		Yes
	55		I like myself	Yes	
	58		I find that I cry frequently		
	62		I feel that I have no one who understands me		
	3	Generalized Anxiety	There are many things I am afraid of		
2	4		My heart races for no good reason		Yes
7	14		I am anxious that I might have a panic attack in public		Yes
9	17		I have sleep difficulties		Yes
10	18		My thoughts are racing		Yes
15	27		I have spells of terror or panic		Yes
17	30		I feel tense		Yes
	33		I am easily frightened of startled		

	39		I experience nightmares or flashbacks		
1	2	Social Anxiety	I am shy around others		
	16		I become anxious when I have to speak in front of audiences		
19	35		I make friends easily	Yes	
22	41		I am concerned that other people do not like me		Yes
24	44		I feel uncomfortable around people I don't know		
26	47		I feel self conscious around others		Yes
	54		I feel comfortable around other people	Yes	
	6	Academic Distress	I enjoy my classes	Yes	
8	15		I feel confident I can succeed academically	Yes	
28	51		I am not able to concentrate as well as usual		Yes
30	53		It's hard to stay motivated for my classes		Yes
33	59		I am unable to keep up with my school work		Yes
3	5	Eating Concerns	I feel out of control when I eat		
6	13		I think about food more than I would like to		
	19		I am satisfied with my body shape	Yes	
	22		I am dissatisfied with my weight		
13	25		I eat too much		
	31		When I start eating I can't stop		
	34		I diet frequently		
	48		I purge to control my weight		
	61		The less I eat, the better I feel about myself		
	1	Family Distress	I get sad or angry when I think of my family		
	7		I feel that my family loves me	Yes	
	11		My family gets on my nerves		
	21		My family is basically a happy one	Yes	
	38		There is a history of abuse in my family		

	42		I wish my family got along better		
18	32	Hostility	I have difficulty controlling my temper		
20	36		I sometimes feel like breaking or smashing things		Yes
23	43		I get angry easily		Yes
	45		I feel irritable		
29	52		I am afraid I may lose control and act violently		Yes
32	57		I frequently get into arguments		
34	60		I have thoughts of hurting others		
	24	Substance / Alcohol Use	I use drugs more than I should		
14	26		I drink alcohol frequently		
16	29		When I drink alcohol I can't remember what happened		
27	49		I drink more than I should		
	50		I enjoy getting drunk		
31	56		I have done something I have regretted because of drinking		

APPENDIX B

OQ-45.2 INSTRUMENT

Outcome Questionnaire (OQ®-45.2)

Instructions: Looking back over the last week, including today, help us understand how you have been feeling. Read each item carefully and mark the box under the category which best describes your current situation. For this questionnaire, work is defined as employment, school, housework, volunteer work, and so forth.

Name:_____ Age:_____Yrs

ID#_____ Sex M ☐ F ☐

Session #_____ Date_____/_____/_____

		Never	Rarely	Sometimes	Frequently	Almost Always
1.	I get along well with others.	4	3	2	1	0
2.	I tire quickly.	0	1	2	3	4
3.	I feel no interest in things.	0	1	2	3	4
4.	I feel stressed at work / school	0	1	2	3	4
5.	I blame myself for things.	0	1	2	3	4
6.	I feel irritated.	0	1	2	3	4
7.	I feel unhappy in my marriage / significant relationship.	0	1	2	3	4
8.	I have thoughts of ending my life.	0	1	2	3	4
9.	I feel weak.	0	1	2	3	4
10.	I feel fearful.	0	1	2	3	4
11.	After heavy drinking, I need a drink the next morning to get going. (If you do not drink, mark "never")	0	1	2	3	4
12.	I find my work / school satisfying.	4	3	2	1	0
13.	I am a happy person.	4	3	2	1	0
14.	I work / study too much.	0	1	2	3	4
15.	I feel worthless.	0	1	2	3	4
16.	I am concerned about family troubles.	0	1	2	3	4
17.	I have an unfulfilling sex life.	0	1	2	3	4
18.	I feel lonely.	0	1	2	3	4
19.	I have frequent arguments.	0	1	2	3	4
20.	I feel loved and wanted.	4	3	2	1	0
21.	I enjoy my spare time.	4	3	2	1	0
22.	I have difficulty concentrating.	0	1	2	3	4
23.	I feel hopeless about the future.	0	1	2	3	4
24.	I like myself.	4	3	2	1	0
25.	Disturbing thoughts come into my mind that I cannot get rid of.	0	1	2	3	4
26.	I feel annoyed by people who criticize my drinking (or drug use) (If not applicable, mark "never")	0	1	2	3	4
27.	I have an upset stomach.	0	1	2	3	4
28.	I am not working /studying as well as I used to.	0	1	2	3	4
29.	My heart pounds too much.	0	1	2	3	4
30.	I have trouble getting along with friends and close acquaintances.	0	1	2	3	4
31.	I am satisfied with my life.	4	3	2	1	0
32.	I have trouble at work / school because of drinking or drug use. (If not applicable, mark "never")	0	1	2	3	4
33.	I feel that something bad is going to happen.	0	1	2	3	4
34.	I have sore muscles.	0	1	2	3	4
35.	I feel afraid of open spaces, of driving, or being on buses, subways, and so forth.	0	1	2	3	4
36.	I feel nervous.	0	1	2	3	4
37.	I feel my love relationships are full and complete.	4	3	2	1	0
38.	I feel that I am not doing well at work / school	0	1	2	3	4
39.	I have too many disagreements at work / school.	0	1	2	3	4
40.	I feel something is wrong with my mind.	0	1	2	3	4
41.	I have trouble falling asleep or staying asleep.	0	1	2	3	4
42.	I feel blue.	0	1	2	3	4
43.	I am satisfied with my relationships with others.	4	3	2	1	0
44.	I feel angry enough at work / school to do something I might regret.	0	1	2	3	4
45.	I have headaches.	0	1	2	3	4

SD	IR	SR

Total=

For More Information:
American Professional Credentialing Services LLC
E-mail: APCS@ERDLS.COM
Web: WWW.OQFAMILY.COM
Toll-Free: 1-888-MH SCORE (1-888-647-2673)
Fax/Voice: 1-873-388-8888

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060-1073. doi: 10.1037//0003-066X.57.12.1060

Ankuta, G., & Abeles, N. (1993). Client satisfaction, clinical significance, and meaningful change in psychotherapy. *Professional Psychology: Research and Practice, 24,* 70-74. doi:10.1037/0735-7028.24.1.70

Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose–effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology, 77,* 203-211. doi:10.1037/a0015235

Baldwin, S. A., Imel, Z. E., Braithwaite, S. R., & Atkins, D. C. (2014, February 3). Analyzing multiple outcomes in clinical research using multivariate multilevel models. *Journal of Consulting and Clinical Psychology.* Advance online publication. http://dx.doi.org/10.1037/a0035628

Beck, A. T., & Steer, R. A. (1996). *Beck Anxiety Inventory: Manual.* San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory—Second Edition: Manual.* San Antonio, TX: Pearson.

Bieschke, K. J., Bowman, G. D., Hopkins, M., Levine, H., & McFadden, K. (1995). Improvement and satisfaction with short-term therapy at a university counseling center. *Journal of College Student Development, 36,* 553-559.

Bishop, J. B. (1995). Emerging administrative strategies for college and university counseling centers. *Journal of Counseling and Development, 74,* 33-38.

Bold, D. M., & Rounds, J. (2000). Advances in psychometric theory and methods. In S. D. Brown & R. W. Lent (Eds.), *Handbook of counseling psychology* (3rd ed., pp. 140-176). New York, NY: John Wiley & Sons.

Callaghan, G. (2001). Demonstrating clinical effectiveness for individual practitioners and clinics. *Professional Psychology: Research and Practice, 32,* 289-297. doi:10.1037/0735-7028.32.3.289

Center for Collegiate Mental Health. (2012). *CCAPS 2012 Technical Manual.* University Park, PA: Author.

Center for Collegiate Mental Health. (2013). *Clinician's Guide to the Counseling Center Assessment of Psychological Symptoms.* University Park, PA: Author.

Center for the Study of Collegiate Mental Health. (2010, March). *2010 Annual Report* (Publication No. STA 11-000).

Cormier, S., & Nurius, P. S. (2003). *Interviewing and change strategies for helpers: Fundamental skills and cognitive behavioral interventions* (5th ed.). Pacific Grove, CA: Brooks/Cole.

Corrigan, P. W. (1990). Consumer satisfaction with institutional and community care. *Community Mental Health Journal, 26,* 151-165.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Harcourt Brace Jovanovich.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302. doi:10.1037/h0040957

Crowe, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349-354.

Deane, F. P. (1993). Client satisfaction with psychotherapy in two outpatient clinics in New Zealand. *Evaluation and Program Planning, 16,* 87-94. doi:10.1016/0149-7189(93)90020-9

Derogatis, L. R. (1994). *Symptom Checklist-90-R: Administration, scoring, and procedures manual.* Minneapolis, MN: National Computer Systems, Inc.

Derogatis, L. R., Rickels, K., & Rock, A. (1976). The SCL-90 and the MMPI: A step in the validation of a new self-report scale. *British Journal of Psychiatry, 128,* 280-289.

Durham, C. J., McGrath, L., Burlingame, G. M., Schaalje, G., Lambert, M. J., & Davies, D. (2002). The effects of repeated administrations on self-report and parent-report scales. *Journal of Psychoeducational Assessment, 20,* 240-257. doi:10.1177/073428290202000302

Finch, A. E., Lambert, M. J., & Schaalje, B. G. (2001). Psychotherapy quality control: The statistical generation of expected recovery curves for integration into an early warning system. *Clinical Psychology & Psychotherapy, 8*, 231-242. doi:10.1002/cpp.286

Froyd, J., Lambert, M., & Froyd, J. (1996). A review of practices of psychotherapy outcome measurement. *Journal of Mental Health, 5*, 11-16. doi:10.1080/09638239650037144

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press. doi:10.1017/CBO9780511790942

Goldberg, D. (1972). *The detection of psychiatric illness by questionnaire.* Oxford, England: Oxford University Press.

Groth-Marnat, G. (2003). *Handbook of psychological assessment* (4th ed.). New York, NY: Wiley.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software, 33*, 1–22.

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

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology, 10*, 249-254.

Hill, C. E., & Lambert, M. J. (2004). Methodological issues in studying psychotherapy processes and outcomes. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 84-135). New York, NY: Wiley.

Holcomb, W. R., Beitman, B. D., Hemme, C. A., Josylin, A., & Prindiville, S. (1998). Use of a new outcome scale to determine best practices. *Psychiatric Services, 49*, 583-595.

Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureno, G., & Villasenor, V. S. (1988). Inventory of interpersonal problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology, 56,* 885-892.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Hsu, L. M. (1999). A comparison of three methods of identifying reliable and clinically significant client changes: Commentary on Hageman and Arrindell. *Behaviour Research & Therapy, 37*, 1195-1202.

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

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67,* 300-307.

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

Kendall, P. C., Holmbeck, G., & Verduin, T. (2004). Methodology, design, and evaluation in psychotherapy research. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 16-43). New York, NY: Wiley.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67,* 285-299.

Kenny, D. A., & Hoyt, W. T.(2009) Multiple levels of analysis in psychotherapy research. *Psychotherapy Research, 19,* 462-468. doi: 10.1080/10503300902806681

Kim, S.-H., Beretvas, S. N., & Sherry, A. R. (2010). A validation of the factor structure of the OQ-45 scores using factor mixture modeling. *Measurement and Evaluation in Counseling and Development, 42,* 275-295. doi: 10.1177/0748175609354616

Lambert, M.J., Burlingame, G.M., Umphress, V., Hansen, N.B., Vermeersch, D.A., Clouse, G.C., & Yanchar, S. C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology and Psychotherapy, 3,* 249-258.

Lambert, M. J., Harmon, C., Slade, K., Whipple, J. L., & Hawkins, E. J. (2005). Providing feedback to psychotherapists on their patients' progress: Clinical results and practice suggestions. *Journal of Clinical Psychology: In Session, 61,*165-174.

Lambert, M. J., & Lambert, J. M. (1999). Use of psychological tests for assessing treatment outcome. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed., pp. 115-151). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Lambert, M. J., Morton, J. J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R. C., . . . Burlingame, G. B. (2004). *Administration and scoring manual for the Outcome Questionnaire-45.* Orem, UT: American Professional Credentialing Services.

Larsen, D. L., Attkisson, C. C., Hargreaves, W. A., & Nguyen, T. D. (1979). Assessment of client/patient satisfaction: Development of a general scale. *Evaluation and Program Planning, 2*, 197-207.

LaSala, M. C. (1997). Client satisfaction: Consideration of correlates and response bias. *Families in Society: The Journal of Contemporary Human Services, 78*, 54-64.

Locke, B. D., Buzolitz, J., Lei, P., Boswell, J. F., McAleavey, A. A., Sevig, T. D., . . . Hayes, J. A. (2011). Development of the Counseling Center Assessment of Psychological Symptoms-62 (CCAPS-62). *Journal of Counseling Psychology, 58*, 97-109. doi:10.1037/a0021282

Locke, B. D., McAleavey, A. A., Zhao, Y., Lei, P., Hayes, J. A., Castonguay, L. G., ... Lin, Y. (2012). Development and initial validation of the Counseling Center Assessment of Psychological Symptoms–34. *Measurement and Evaluation in Counseling and Development, 45*, 151-169. doi:10.1177/0748175611432642

Lueger, R. J., Howard, K. I., Martinovich, Z., Lutz, W., Anderson, E. E., & Grissom, G. (2001). Assessing treatment progress of individual patients using expected treatment response models. *Journal of Consulting and Clinical Psychology, 69*, 150-158. doi:10.1037/0022-006X.69.2.150

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

MacCallum, R. C., Cheongtag, K., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research, 32*, 215-253. doi:10.1207/s15327906mbr3203_1

McAleavey, A. A., Nordberg, S. S., Hayes, J. A., Castonguay, L. G., Locke, B. D., & Lockard, A. J. (2012). Clinical validity of the Counseling Center Assessment of Psychological Symptoms-62 (CCAPS-62): Further evaluation and clinical applications. *Journal of Counseling Psychology, 59*, 575-590. doi:10.1037/a0029855

McGlinchey, Atkins, & Jacobson (2002). Clinical significant methods: Which one to use and how useful are they? *Behavior Therapy, 33*, 529-550.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi:10.1037/0003-066X.50.9.741

Mintz, J., & Keisler, D. J. (1982). Individualized measures of psychotherapy outcome. In P. C. Kendall & J. N. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 491-534). New York, NY: Wiley.

Moore, K. E., & Kenning, M. (1996). Assessing client satisfaction in a psychology training clinic. *Journal of Mental Health Administration, 23*, 180-189.

Newman, F. L., Ciarlo, J. A., & Carpenter, D. (1999). Guidelines for selecting psychological instruments for treatment planning and outcome assessment. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed., pp. 153-170). Mahwah, NJ: Lawrence Erlbaum.

Ogles, B., Lunnen, K., & Bonesteel, K. (2001). Clinical significance: History, application, and current practice. *Clinical Psychology Review, 21*, 421-446. doi:10.1016/S0272-7358(99)00058-6

Osterlind, S. J. (2006). *Modern measurement: Theory, principles, and applications of mental appraisal.* Upper Saddle River, NJ: Pearson.

Powell, R. A., Holloway, F., Lee, J., & Sitzia, J. (2004). Satisfaction research and the uncrowned king: Challenges and future directions. *Journal of Mental Health, 13*, 11-20. doi: 10.1080/09638230410001654495

R Core Development Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385-401.

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology, 38,* 119 –125. doi:10.1002/1097-4679(198201)38:1_119::AID-JCLP2270380118_3.0.CO;2-I

Schauenberg, H., & Strack, M. (1999). Measuring psychotherapeutic change with the Symptom Checklist SCL 90 R. *Psychotherapy and Psychosomatics, 68,* 199-206. doi:10.1159/000012333

Schmitz, N., Hartkamp, N., & Franke, G. H. (2000). Assessing clinically significant change: Application to the SCL-90-R. *Psychological Reports, 86,* 263-274.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*, 922-932. doi:10.1037/0003-066X.44.6.922

Speer, D. C., & Newman, F. L. (1996). Mental heath services outcome evaluation. *Clinical Psychology: Science and Practice, 3,* 105—129.

Steenbarger, B. N., & Smith, H. B. (1996). Assessing the quality of counseling services: Developing accountable helping systems. *Journal of Counseling and Development, 75*, 145-150.

Tanner, B. A., & Stacey, W., Jr. (1985). A validity scale for the Sharp Consumer Satisfaction Scales. *Evaluation and Program Planning, 8*, 147-153.

Tingey, R., Lambert, M., Burlingame, G., & Hansen, N. (1996). Assessing clinical significance: Proposed extensions to method. *Psychotherapy Research, 6*, 109-123. doi: 10.1080/10503309612331331638

Todd, D. M., Deane, F. P., & McKenna, P. A. (1997). Appropriateness of SCL-90-R adolescent and adult norms for outpatient and nonpatient college students. *Journal of Counseling Psychology, 44*, 294-301. doi:10.1037/0022-0167.44.3.294

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

Vermeersch, D A., Whipple, J. L., Lambert, M. J., Hawkins, E. J., Burchfield, C. M., & Okiishi, J. C. (2004). Outcome Questionnaire: Item sensitivity to changes in counseling center clients. *Journal of Counseling Psychology*, *51*, 38-49.

Vonk, M. E., & Thyer, B. A. (1999). Evaluating the effectiveness of short-term treatment at a university counseling center. *Journal of Clinical Psychology, 55,* 1095-1106.

Weissman, M. M., & Bothwell, S. (1976). Assessment of social adjustment by patient self-report. *Archives of General Psychiatry*, *33*, 1111-1115.