2009-04-29

Relative Sensitivity to Change of Psychotherapy Outcome Measures for Children and Adolescents: A Comparison Using Parent- and Self-Report Versions of the CBCL/6-18, BASC-2, and Y-OQ-2.01

Debra Theobald McClendon
Brigham Young University - Provo

RELATIVE SENSITIVITY TO CHANGE OF PSYCHOTHERAPY OUTCOME MEASURES FOR CHILDREN AND ADOLESCENTS: A COMPARISON USING PARENT- AND SELF-REPORT VERSIONS OF THE CBCL/6-18, BASC-2, AND Y-OQ-2.01

by

Debra Theobald McClendon, MA

A dissertation submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Psychology

Brigham Young University

August 2009

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a dissertation submitted by

Debra Theobald McClendon

This dissertation has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	Jared S. Warren, Chair
Date	Gary M. Burlingame
Date	Dennis L. Eggett
Date	Harold L. Miller, Jr.
Date	George L. Bloch

BRIGHAM YOUNG UNIVERSITY

As Chair of the candidate's graduate committee, I have read the dissertation of Debra Theobald McClendon in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date	Jared S. Warren Chair, Graduate Committee

Accepted for the Department

Date	Harold L. Miller, Jr. Graduate Coordinator, Psychology

Accepted for the College

Date	Susan S. Rugh Associate Dean, College of Family, Home and Social Sciences

ABSTRACT

RELATIVE SENSITIVITY TO CHANGE OF PSYCHOTHERAPY OUTCOME MEASURES FOR CHILDREN AND ADOLESCENTS: A COMPARISON USING PARENT- AND SELF-REPORT VERSIONS OF THE CBCL/6-18, BASC-2, AND Y-OQ-2.01

Debra Theobald McClendon

Department of Psychology

Doctor of Philosophy

This repeated-measures study evaluated the relative sensitivity to change of the Child Behavior Checklist/6-18 (CBCL/6-18), the Behavior Assessment System for Children-2 (BASC-2), and the Youth Outcome Questionnaire-2.01 (Y-OQ-2.01). Participants were recruited from Valley Mental Health, a community outpatient clinic in Salt Lake City, UT. There were 178 participants for 136 cases, with 134 adults and 44 adolescents. Participants provided two through five data points for a total of 548 data points. Hierarchical Linear Modeling (HLM) was conducted for three major comparisons: adult informants, adult and adolescent dyads, and adolescents. Results indicated the Y-OQ-2.01 was the most change sensitive, while the BASC-2 and CBCL/6-18 were not statistically different from each other. Results also showed that the parent-report measures were more change-sensitive than the self-report measures completed by

adolescent informants. Sensitivity to change was also evaluated through the reliable change index (RCI) and the use of cut-off scores. In comparisons using the RCI, the Y-OQ-2.01 identified the most cases for reliable change. The Y-OQ-2.01 also had the greatest corroboration of its findings with the other two measures. In comparisons using cut-off scores, results are offered for three variations, as different standards were used to establish cut-off scores for the three measures. The third variation, for which cut-off scores for all three measures were adjusted to one standard deviation above the mean, is suggested to be the most appropriate when comparing measures. Those results indicated there was no statistical difference in how the measures performed relative to each other. Thus, based on the HLM and RCI results of this study, it is recommended that clinicians select the Y-OQ-2.01 for outcome use and tracking changes in child and adolescent symptoms and behaviors. **Keywords**: treatment outcome, child, adolescent, sensitivity to change.

ACKNOWLEDGMENTS

I acknowledge that this work is the culmination of many scholarly and personal partnerships. I express my most heart-felt appreciation to all who have been part of seeing this project come to light. I am particularly grateful for Jared Warren, for not only chairing my committee, but also for offering financial and human resource support through his research team. I was able to avoid many research pitfalls due to his generosity with information and research assistants. I express gratitude to my mentor, Gary Burlingame, in encouraging and financially supporting this project, thus showing his confidence in me despite the logistic complexities of the research design. I am exceedingly grateful to Dennis Eggett of the Statistics Department for his participation on this project. His patience with my anxieties regarding statistics was unwavering and his expertise priceless. Special thanks also go to the others on my graduate committee including Harold Miller and George Bloch. Their critical analyses and editorial suggestions have been invaluable in producing the project herein presented.

Data collection for this study continued over a period of approximately 18 months. I am especially grateful to Valley Mental Health for hosting this lengthy project and offer my thanks to Catherine Carter, Karen Platis, Tammy Patterson, Dave Justice, and Christina Arellano for their time, support, and patience. I am also appreciative of the receptionists who assisted us on a daily basis.

To the faculty members and the secretaries in Clinical Psychology, I express my thanks. I am also grateful for the many hours of work provided by Aimee Vincent, David Mason, Candilyn Newell, Carolyn Cornwall, Kat Tumblin and Sean Woodland who

assisted me with administrative duties. I am grateful to those research assistants who recruited community sponsors, with a special thanks to Alicia Sorenson. I am also appreciative of those who traveled to Valley Mental Health to recruit and meet with research participants at their own expense; this work could not have been completed without their many hours of travel and service. I am likewise grateful to those who assisted in scoring measures, data entry, packet preparation, phone calls, and mailings; these tasks were crucial to the success of this project.

Much appreciation goes to my family. I express particular gratitude to my husband, Richard. His generous statistical tutoring was essential to the completion of this project. Yet, more valuable to me than his statistical knowledge was his unceasing conviction and faith that this project would succeed. I thank my sister-in-law, Susan Theobald, who has been a never-failing source of emotional support and counsel throughout the many seasons of my life. I also thank my beautiful daughters, Katja and Chloe, for their love and smiles, as well as for the sacrifices they have made, though perhaps not within their awareness, in order for me to do this work.

Lastly, I thank my loving Heavenly Father who has supported and strengthened me in moments of weakness and provided peace and comfort in times of angst and doubt—throughout this project and the doctoral program at large. After 8 years it has all come together for completion. I am grateful for the many experiences and people that have contributed to my professional and personal development. Yet, above all, I am eternally grateful for the many tender mercies—and even miracles—that have been generously granted by my Father in Heaven to help me complete this journey. Thank you, Father.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

.

Relative Sensitivity to Change of Psychotherapy Outcome Measures for Children and Adolescents: A comparison using parent- and self-report versions of the CBCL/6-18, BASC-2, and Y-OQ-2.01

CHAPTER 1

Introduction

Outcome measures are increasingly used in all health care fields due to managed health care organizations. These organizations have been set up to provide quality care at minimal cost (Richardson & Austad, 1991). To accomplish this objective health care organizations demand accountability by health care professionals; they must demonstrate that the care they provide is beneficial to their patients. Clinical psychology, along with all mental health care professions, is subject to this demand as well. If practitioners in the mental health field desire to participate in treating clients that use these health care corporations to manage the access they have to services, then they must adapt clinical, administrative, and organizational procedures to meet the health care organizations' expectations (Richardson & Austad, 1991). Therefore, in recent years there has been a stronger focus on outcome assessment within clinical psychology; practitioners are increasingly required to demonstrate the efficacy of the treatment they provide (Koss & Shiang, 1994).

Many measures used for treatment outcome have been adopted as outcome instruments although they were initially designed for some other purpose, such as assigning an accurate diagnosis. Two such measures with widespread use in clinical child and adolescent psychology include the Child Behavior Checklist (CBCL; Achenbach,

1991) and the Behavioral Assessment System for Children (BASC; Reynolds & Kamphaus, 1992). In spite of their widespread use, these measures may not be valid for assessing outcome (Lambert & Hill, 1994). An outcome measure is intended to measure change due to a psychotherapeutic intervention. Although these adopted measures generally have excellent psychometric properties, they have unknown or restricted sensitivity to change because they tend to assess static constructs that take longer to show change (Berrett, 2000; Vermeersch et al., 2000). Yet, sensitivity to change is the most important characteristic of an outcome measure (Burlingame et al., 1995). The Outcome Questionnaire (Lambert et al., 1996) was specifically designed to track treatment progress and outcome in adults while the Youth Outcome Questionnaire (Y-OQ; Burlingame et al., 2001) is a similar measure specifically for child and adolescent populations. The Y-OQ was developed to address the shortcomings of other measures in child and adolescent research and clinical practice by allowing assessment of dynamic constructs that would be sensitive to behavioral and symptomatic changes when administered on a regular basis.

Using measures without established change sensitivity, such as the CBCL/6-18 and the BASC-2, is potentially problematic since their consumers attempt to assess the quality of treatments and calibrate the needs of individual clients based on the data these measures produce. Since it is not clear how these measures compare to those specifically designed to assess therapeutic outcome, such as the Y-OQ-2.01, a comparative study was necessary to evaluate their relative sensitivity to change. Significant differences found between these measures regarding sensitivity to change will allow the work of researchers, clinicians, and health care corporations to be maximized for the ultimate

benefit of clients by discontinuing the use of measures that are not as sensitive to change in favor of measures that would meet their purposes for assessment more appropriately. While no significant differences found between these measures will allow outcome measure consumers to select a measure based on other desired attributes, such as its ability to diagnose.

Therefore, the purpose of this study was to investigate the relative sensitivity to change of the Youth Outcome Questionnaire (Y-OQ-2.01) and two measures often used for outcome assessment that were not designed for that purpose, the Child Behavior Checklist (CBCL/6-18) and the Behavior Assessment for Children (BASC-2). Discussion of this research and its findings proceeds as follows:

Chapter 2 contains an exploration of the literature in regards to the purpose of outcome measures and their importance to researchers, clinicians, and health care corporations. It also examines measuring outcome, including the psychometric considerations of reliability, validity, and sensitivity to change; the discussion focuses on sensitivity to change. The specific characteristics of the CBCL/6-18, BASC-2, and Y-OQ-2.0 are then discussed, followed by a comparison of all three measures with regard to recommendations made by researchers and clinicians, as reported in the literature, on key features of outcome measures. Finally, chapter 2 presents the statement of the problem, a review of the analysis of therapeutic change, and the hypotheses of the present study.

Chapter 3 discusses the methods of this study including a description of the clinic used for data collection, response rate, demographic features of the sample, and logistical procedures used. This chapter also discusses how reliable change indices were calculated to determine those participants whose data would be retained for analysis and presents a

summary of their demographics. Attention then turns to a literature review of Hierarchical Linear Modeling, the chosen analytic procedure.

Chapter 4 presents the results of this study. It reviews descriptive results from the reliable change index (RCI) and cut-off score comparisons. It also reviews the HLM analyses that include comparisons for adult informants, adult and adolescent dyads, and adolescent informants.

Chapter 5 discusses the major findings of this study along with their implications for researchers and clinicians working with children and adolescents. It also addresses limitations of this study while making recommendations for future studies.

CHAPTER 2

Literature Review

Purpose of Outcome Measures

Within the mental health community, using outcome measures has become increasingly important in recent years to researchers, clinicians, and health care corporations (Burlingame, Lambert, Reisinger, Neff, & Mosier, 1995; Burlingame et al., 2001; Burlingame, Wells, Lambert, & Cox, 2004).

Researchers

Standardized outcome assessment is a fundamental component of psychotherapy research (Ogles, Lambert, & Fields, 2002). Although research has well documented the effectiveness of psychotherapy with adult populations (Lambert & Ogles, 2004), the parallel research with children and adolescents is less well-developed (Kazdin, 1993). Nevertheless, research in this area suggests that psychotherapy for children and adolescents is effective (Kazdin, 1993) and that individual therapy with children is approximately as effective as with adults (Brown, 1987). Advances continue in this area, with studies examining the effectiveness of therapy for children and adolescents exceeding 1,500, including over 500 modes of treatment (Kazdin, 2004). However, according to Kazdin (2004), most of these treatments have never been subjected to empirical research. As empirical inquiry proceeds, Kazdin (1995) recommends that standardized outcome measures be used to profile children and adolescents in a consistent way to further researchers' efforts by allowing them to integrate studies about specific problems. Durlak et al. (1995) also indicate that the use of outcome measures would strengthen the child psychotherapy research.

Clinicians

Clinicians are theoretically the fundamental consumers of psychotherapy research so they can implement effective research developments into their treatments (Burlingame et al. 2004). However, clinicians are often suspicious of the research process. For many, the research process seems dissimilar from clinical practice and even appears irrelevant (Burlingame et al., 2004). Often, research topics and individuals used as subjects within the research are so different from what clinicians work with on a daily basis that generalizability to clinical practice is limited (Kazdin, 1991). Another reason for the discrepancy between clinicians and researchers is that third-party payers do not view the research findings as irrelevant; instead, they rely on these findings. Many clinicians fear losing their jobs or positions on provider panels if they do not make substantial gains in brief time frames with more severely disturbed clients than the researchers ever attempted to treat in their studies (Brown, Burlingame, Lambert, Jones, & Vaccaro, 2001). However, Durlak et al. (1995) reviewed the literature on clinicians' concerns and found that, in spite of clinicians' suspicions, research findings are applicable to clinical practice.

One of the most helpful contributions researchers have made to clinicians is the development of various outcome measures; however, many clinicians have not yet adopted the use of these instruments in their clinical practice. Recently, Hatfield and Ogles (2004) discovered that, out of 874 randomly selected practicing psychologists, only 37% used some form of outcome assessment in their practice. When those who did not use outcome measures (n=550) were asked their reasons for not doing so, they indicated, in order of importance: "adds too much paperwork", "takes too much time", "extra burden on clients", "feel it is not helpful", and "do not have enough resources."

Yet despite these negative perceptions, there are indications of a trend toward increased outcome measure use among clinicians in general (Phelps, Eisman, & Kohout, 1998). Although Hatfield and Ogles (2004) reported 37% of practicing psychologists in their study use outcome measures, it is a substantial increase over the 29% reported by Phelps et al. (1998) and the 23% reported by Bickman et al. (2000). Furthermore, it is important to note that of those clinicians in the Hatfield and Ogles (2004) study who reported using outcome measures, the proportion of child and adolescent clinicians using outcome measures actually was higher (54%).

This trend of increased outcome measure use may be due in part to the advantages they provide to the clinician as they provide psychotherapy treatment to their clients. Both researchers (Lambert et al., 2001; Wells, Burlingame, Lambert, Hoag, & Hope, 1996) and clinicians in the Hatfield and Ogles (2004) study come together in agreement regarding these advantages:

1. Outcome measures serve as an intake measure of current functioning, initial severity, and an index of risk factors that might moderate expectations for rapid improvement (Lambert et al., 2001; Wells et al., 1996).
2. Outcome measures track change—the progress that has been obtained since therapy started (Lambert et al., 2001; Wells et al., 1996). "Tracking client progress" was the most important reason stated among clinicians who used outcome measures. In addition, it was also rated as the most useful type of information that outcome measures produce among those clinicians who did *not* use outcome measures (Hatfield & Ogles, 2004).

3. Using standardized outcome measures can provide additional outside validation of clinical judgment, which can aid practitioners in providing better services for their clients. “Determine if there is a need to alter treatment” was the second most important reason stated among clinicians using outcome measures (Hatfield & Ogles, 2004).
4. Outcome measures provide a potential summary source for demonstrating the effectiveness of therapeutic interventions (Wells et al., 1996) by the use of aggregated data. Through analysis of data from their own clientele, clinicians can make decisions about the effectiveness of their own delivery of psychotherapeutic services (Hatfield & Ogles, 2004; Lambert, Hansen, & Finch, 2001).
5. The use of outcome measures represents a more ethical practice (Clement, 1994). “Ethical practice” was the third reason given by psychologists for using outcome measures (Hatfield & Ogles, 2004).

The trend may also be due in part to private insurance companies and managed care companies increasingly requiring practitioners to administer outcome-assessment instruments in order to make decisions for insurers and care managers about the effectiveness and efficiency of services. These external pressures may influence the degree to which clinicians use outcome measures, regardless of whether clinicians would choose to use them of their own accord (Hatfield & Ogles, 2004). However, in the Hatfield and Ogles (2004) study, “Required by MCO/insurance” (MCO=managed care organization) and “Required by work setting” were not central for clinicians who used outcome measures (i.e. they were indicated as the 5th and 6th most important reasons).

Health Care Corporations

The cost of providing mental health treatment to children and adolescents is difficult to estimate. However, evidence from several sources indicates a concern for cost-containment and accountability in the health care market (Burlingame et al., 1995). Between the years of 1986 and 1996, expenditures for mental health services grew at an average annual rate of more than 7 percent, which was equivalent to the growth seen in *total* health care costs during the 1990s (U.S. Dept of Health and Human Services, 1999). According to Burlingame et al. (2004), managed health care systems and the so-called *era of accountability* are the health care industry's response to these increasing costs. Therefore, in spite of the concerns of clinicians described previously, third-party payers require providers to be able to document therapeutic progress. "How much therapeutic effect can be 'bought' for how much therapy?" is the central question for which these companies are searching for answers (Linden & Wen, 1990). This zeitgeist has put psychotherapy outcome research into a central role as researchers are pressured to provide health care corporations with valid, reliable, and sensitive measures to track the quality of mental health treatment.

Health care corporations enter data from standardized outcome measures into a data bank for analysis. According to Wells et al. (1996) this allows health care corporations to report therapeutic effectiveness to subscriber companies and profile individual providers; establish decision algorithms to empirically determine appropriate session limits (e.g., expectancy tables); and answer further research questions, such as evaluating the efficacy of new treatment modalities (Wells et al., 1996).

As asserted herein, the needs of researchers, clinicians, and health care corporations have intersected, theoretically and practically. A critical method for addressing these needs is to measure treatment outcome.

Measuring Outcome

Many outcome measures used by clinicians have little or no empirical foundation (Froyd, Lambert, & Froyd, 1996). Employing outcome measures that have demonstrated acceptable reliability and validity estimates and are sensitive to change is crucial in order to accurately and reliably assess change as it occurs during and after therapy (Lambert & Hill, 1994; Wells et al., 1996). This practice will allow clinicians to provide the best possible services to their clients (Hatfield & Ogles, 2004).

Reliability

The reliability of a measure is important since it is an estimate of the amount of error contained therein (Allen & Yen, 1979). If estimated reliability is low, a clinician's confidence regarding the measure's ability to accurately reflect changes in an individual's symptoms and behavior is also low due to the presence of greater amounts of error. If estimated reliability is high, then a clinician's confidence in the measure's ability to reflect changes in an individual's symptoms and behavior is also high due to the presence of limited amounts of error. High estimated reliability also indicates that, if an individual is given a measure a second time (or repeatedly), similar results would be found if that individual remained stable on the symptoms and behaviors the measure was designed to assess.

In outcome measurement estimated reliability serves an additional function. The measure is often given to individuals before and after treatment and/or several times

during treatment. A change score is then calculated from the results of those measurements. The change score reflects both the true difference in the symptoms and behaviors being measured and the error inherent within the measure (Allen & Yen, 1979; Lambert & Hill, 1994). Thus, reliability of the change scores will be directly affected by the estimated reliability of the outcome measure. An outcome measure with low estimated reliability will produce measurements that will vary over time and across studies due to the higher level of error within the measure; these measurements may lead to mixed findings regarding the effectiveness of a given treatment. As the purpose of an outcome measure is to assess the effectiveness of a given treatment by way of measuring changes in an individual's symptoms and behavior, an outcome measure with low reliability cannot be used to accurately measure change and will prove uninformative to clinicians.

Researchers (Burlingame et al., 1995; Durlak et al., 1995) recommend internal-consistency reliability-coefficients, which estimate the homogeneity of items on a measure, at or above 0.80. Test-retest reliability assesses the temporal stability of an outcome measure, usually by administering an instrument to subjects twice, without a significant variable introduced in the intervening time period. Researchers (Durlak et al., 1995; Reisinger & Burlingame, 1997) recommend that this statistic be above 0.70.

Validity

The validity of a measure is also important to the study of outcome sensitivity since it is concerned with the measure's ability to measure what it purports to measure (Reisinger & Burlingame, 1997). This general definition of validity is actually a description of construct validity: how well a given instrument measures a theoretical

concept. There are several other types of validity to consider when selecting a particular outcome measure for use in clinical practice. Face validity refers to the measure's appearance of validity, content validity the adequacy with which the measure's items assess the construct domain, and criterion-related validity the measure's ability to correlate highly with other measures designed for a similar purpose or purporting to measure similar symptoms and behavior. Researchers (Burlingame et al., 1995; Reisinger & Burlingame, 1997) suggest that validity coefficients be no lower than 0.50 for an outcome measure and consider a validity coefficient above 0.75 to be excellent.

Validity for change. An outcome measure is intended to measure change in symptoms and behavior following psychotherapeutic treatment. Some outcome measures were specifically designed to track treatment progress, including outcome; others have been adopted as such although they were previously designed for some other purpose, such as assigning an accurate diagnosis. Regardless of developmental origin, it is still vital to the validity of the measure for use in outcome measurement to be able to detect change. "Validity alone is not sufficient to make a measure responsive to treatment effects. What is required is validity for change. A measure can be a valid indicator of a characteristic but still not be a valid indicator of change on that characteristic" (Lipsey, 1990, p. 100). If a measure demonstrates acceptable construct, face, content, and criterion-related validity, and yet is limited in its ability to detect change when change has occurred, then the validity of the instrument for use as an outcome measure is compromised. For example, a measure designed to provide diagnostic profiles is not clearly useful in assessing outcome (Lambert & Hill, 1994), although it may be clearly useful in aiding clinicians to assign appropriate diagnoses.

Lipsey (1990) gives guidance on evaluating a measure's validity for change. When identifying experimental studies in which the treatments, samples, and measures are similar to those one might have planned to conduct, if those studies show large effects, this indicates that the measures must have validity for measuring change. Effect size is represented by **d** and is a measure of the degree to which the population means of two samples differ (μ_1-μ_0) in terms of the standard deviation of the parent population (Howell, 2002). More specifically, **d** for any outcome measure is the difference between post-treatment means for treatment and comparison groups divided by the outcome measure's standard deviation. According to Cohen (1988), an effect size is considered small when **d = 0**.20, moderate when **d** = 0.50, and large when **d** = 0.80 or greater.

Validity for change has not been clearly discussed in the child and adolescent outcome literature; perhaps it has been considered synonymous to sensitivity to change. For example, Burlingame et al. (2001) noted that an early study showing large effect sizes for the CBCL (Webster-Stratton, 1984) "provide[d] evidence for the measure's overall sensitivity to change" but then noted that the "CBCL's sensitivity to change resulting from psychotherapy has been questioned" (Drotar, Stein, & Perrin, 1995, p. 363). It appears that the Webster-Stratton (1984) article provided evidence not for sensitivity to change in a general or "overall" sense, but evidence of validity for change. Sensitivity to change is much more complex and more difficult to achieve than validity for change.

Sensitivity to change

To understand sensitivity to change, it is first important to understand the concept of sensitivity as used in the psychometric literature. Sensitivity deals with the issue of inclusion—how well an instrument selects subjects for the trait it is measuring. Sensitivity is calculated by dividing the number of true positives identified by the measure by the total number of actual positives (which includes those missed by the measure). Values range from 0 to 1.0 and the higher the value, the more sensitive the measure. Sensitivity has a reciprocal relationship to specificity. Specificity deals with the issue of exclusion—how well an instrument deselects a subject for the trait it is measuring. Specificity is calculated by dividing the number of true negatives identified by the measure by the total number of actual negatives (which includes those missed by the measure). Values range from 0 to 1.0 and the higher the value, the more specific the measure. Sensitivity and specificity are largely determined by the set-point for the cut-off score of the measure (Allen & Yen, 1979).

When sensitivity is expanded to repeated-measures or multi-wave data, where change scores are of interest, it becomes the new but related issue of sensitivity to change. It is identified by examining subjects of varying severity levels via sensitivity and specificity analyses (Burlingame et al., 2001). Sensitivity to change is also examined by identifying the clinical significance associated with cut-off scores and a reliable change index (RCI; Jacobson et al., 1984; Jacobson & Truax, 1991). The assumption of using a cut-off score is that a client's outcome score will drop from a clinical range to a normal range following successful psychotherapeutic treatment. This score is a set-point that serves to define a client's score relative to either the clinical or normal populations

(Burlingame et al. 2001). The RCI is unique to each outcome measure and establishes a confidence-interval based on error variance that must be exceeded in order to label a client's change as "reliable" (Burlingame et al., 2005). The goal of the RCI is to allow an outcome measure to identify when a client has made changes as a result of the therapeutic intervention.

Sensitivity to change is important when assessing the value of an outcome instrument, since it reflects a measure's ability to detect changes that occur as a result of participation in psychotherapeutic treatment (Lambert & Hill, 1994). This can also be considered equivalent to the measure's responsiveness (Vermeersch, Lambert, & Burlingame, 2000). Lipsey (1990) defined sensitivity more precisely: "Measurement sensitivity…means that measured values fully reflect any change of interest on the characteristic measured and do not reflect an appreciable amount of…variance from any other source" (Lipsey, 1990 p. 120).

These definitions suggest that an outcome measure's sensitivity to change is directly related to the ability of the instrument to do what it purports to do, which is to measure an individual's change in symptoms and behavior over time due to a psychotherapeutic intervention. Therefore, the concept of sensitivity to change is best conceptualized as an issue of construct validity (Vermeersch et al., 2000).

Criteria for establishing sensitivity to change. According to Kazdin (2005), evidenced-based assessment requires "delineating the different purposes of assessment, and then, for each purpose, identifying the special requirements and then the criteria for stating when these requirements are met" (p. 548). Reliability, validity, and sensitivity to change have each been discussed in terms of their importance to outcome measures,

however, according to Burlingame et al. (2005): "Sensitivity to change is the most important characteristic of a treatment outcome instrument" (underlines included in original source). Outcome measures are qualitatively different from other measures, such as those used to make diagnostic decisions that do not need to be sensitive to change. Given the importance of change sensitivity for psychotherapy outcome measures, researchers have begun to propose criteria for establishing change sensitivity that are disparate from criteria used to establish psychometrics for other measures (Guyatt, 1988; Meier, 1997).

Three criteria have been suggested when selecting items for measures that assess change:

1. Items should show change resulting from an intervention (Tryon, 1991) and overall change reflected in client scores on a given item must occur in the theoretically proposed direction (Vermeersch et al., 2000). Although a client may worsen on an item during the initial stages of therapy, resulting in changes in item scores opposite to those proposed, the overall change reflected in an item must occur in the theoretically proposed direction
2. Items should not change when there is no intervention, such as when the client is placed in control conditions (Tryon, 1991). More specifically, since several studies have identified the presence of re-test effects that result in decreased endorsement of symptomatology over time by clients who go untreated (Aneshensel, Estrada, Hansell, & Clark, 1987; Durham, Burlingame, & Lambert, 1998; Jorm, Duncan-Jones, & Scott, 1989), clients receiving an

intervention should change significantly more than clients under control conditions (Vermeersch et al., 2000).

3. Changes in scores should not be attributable to measurement error or to confounding factors such as social desirability, practice effects, mechanical responding, or mere regression (Vermeersch et al., 2000).

Limits to sensitivity. There are reliable differences in the sensitivity to change of outcome measures (Casey & Berman, 1985). Information regarding an outcome measure's sensitivity to change is needed before the instrument can be confidently used to evaluate the effects of treatment on individuals (Deyo & Inui, 1984; Lipsey, 1983). Although this information may not always be readily available, Vermeersch et al. (2000) identify five limits to a measure's ability to be change-sensitive that may help researchers and clinicians evaluate a measure of interest:

1. Scales within the measure may include items that are not relevant to the construct of interest. This is largely related to the use of multi-trait scales (Fitzpatrick et al., 1992).
2. Measures eliciting responses that are categorically arranged (e.g., *yes–no* or *true–false*) or restricted to a small range (e.g., a Likert-scale range from 0 to 2) may be scaled in units that are too gross to detect or be sensitive to change (Lipsey, 1990).
3. Scales may contain instructions to the respondents that are not conducive to the detection of change. For example, outcome measures that ask clients to answer items according to how they have felt over an extended period of time (e.g., over the past 6 months) are not likely to be sensitive in detecting

changes resulting from treatments that have been delivered weekly over a brief period of time (Berrett, 2000).

4. Instruments may include items that tap into constructs that are more static and therefore less susceptible to change; while others may include items that tap into constructs that are more dynamic and therefore more susceptible to change. Measures that tap into static constructs may have greater difficulty detecting client changes within brief periods.
5. Measures may contain items that are subject to floor or ceiling effects. This is problematic because it can limit the ability of the item to detect growth or depreciation at the upper or lower end of the construct of interest (Lipsey, 1990).

Sensitivity to change relative to measures' varying forms. Many measures have different forms that vary based on the informant from whom the data are obtained. The most commonly used forms for children and adolescents are parent-, teacher-, and self-report. Although it would be expected that these different forms would have varying levels of sensitivity to change, this feature has yet to be discussed in the literature. Drotar et al. (1995) discussed the relative sensitivity of the CBCL's parent form to the teacher and self-report forms for screening purposes (identifying disorders) and found the CBCL parent-form to be the most sensitive. However, this type of sensitivity cannot be considered synonymous with change sensitivity. In an epidemiological study that examined parent–child/adolescent dyad data on a generic measure of health and well-being, the Child Health Questionnaire, Waters, Stewart-Brown, and Fitzpatrick (2003) reported that adolescents are more sensitive to their general health, body pain, mental

health than are their parents. Although these content areas are included in most outcome measures, this finding should not be inferred to mean that children and adolescents would be more sensitive to documenting change in these areas due to a psychotherapeutic intervention. Finally, Casey and Berman (1985) reviewed 75 studies and found that different measures did report the effects of treatment differently; observers, therapists, parents, and subject performance reports produced significantly higher effects than teacher- and child self-report. This finding may indicate that parent-report forms would be more sensitive to change than a child or adolescent's self-report form, but since it is reported in terms of treatment effects it is unclear if similar conclusions would be made regarding sensitivity to change.

Recommendations for Properties of Outcome Measures

To aid clinicians, researchers, and health care managers in choosing an ideal outcome instrument, the following recommendations from researchers and clinicians have been identified from the literature:

1. An outcome measure should have excellent reliability (Lambert & Hill, 1994; Vermillion & Pfeiffer, 1993; Weber, 1997).
2. An outcome measure should have excellent validity (Lambert & Hill, 1994; Vermillion & Pfeiffer, 1993; Weber, 1997); including validity for change (Lipsey, 1990).
3. An outcome measure should contains items that are sensitive to symptomatic and behavioral changes that occur with treatment (i.e., items are sensitive to change; Achenbach & Rescorla, 2004; Lambert & Hill, 1994; Vermillion & Pfeiffer, 1993).

4. An outcome measure should be normed so clinicians can make an assessment of the clinical significance of treatment effects, as opposed to only statistical significance (Durlak et al., 1995; Vermillion & Pfeiffer, 1993); this would best be accomplished by the use of cut-off scores and an RCI (Jacobson et al., 1984; Jacobson & Truax, 1991).
5. An outcome measure should be completed by clients in a matter of minutes; (Burlingame et al., 1995; Kazdin, 2005; Lambert & Hill, 1994; Lipsey, 1990) so it does not take too much time or put extra burden on clients (Hatfield & Ogles, 2004).
6. An outcome measure should be scorable and interpretable in a matter of minutes (Burlingame et al., 1995; Kazdin, 2005; Lambert & Hill, 1994) so it does not add too much paperwork or otherwise tax the available human resources (Hatfield & Ogles, 2004).
7. An outcome measure should provide relevant information regarding the client and changes in her or his symptoms and behavior (Lambert & Hill, 1994) so it is practical (Kazdin, 2005; Vermillion & Pfeiffer, 1993) and helpful to the clinician (Hatfield & Ogles, 2004). Lambert et al. (2001) found that giving feedback to clinicians concerning client change (as assessed by an outcome measure) resulted in better therapeutic outcome and more therapy sessions for clients who were at a high risk for treatment failure.
8. An outcome measure should allow frequent use to track progress, or monitor treatment (Burlingame et al., 1995; Hatfield & Ogles, 2004; Kazdin, 2005) so that treatment can be altered accordingly or problem areas can be identified

dynamically rather than waiting an extended period of time or until the end of treatment before making an assessment. Much of the research on therapeutic change has been based on data on a client's status at two time points, for example, scores on a pretest and a posttest (Bryk & Raudenbush, 1987). In general, two time points provide an inadequate basis for studying change (Bryk & Weisberg, 1977; Rogosa, Brandt, & Zimowski, 1982).

9. An outcome measure should be cost efficient (i.e., inexpensive) so it does not adversely impact the consumer's financial resources (Burlingame et al., 1995; Hatfield & Ogles, 2004; Weber 1997).

The Child Behavior Checklist/6-18 (CBCL/6-18)

The CBCL/6-18 is the newest version of the CBCL, the most commonly used child and adolescent measure used in the assessment of psychosocial dysfunction (Kazdin, 1994); it is probably the most commonly used measure for outcome purposes as well (Hatfield & Ogles, 2004). The CBCL/6-18 is only one of a family of measures called the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach & Rescorla, 2004) that was developed for assessing problems, competencies, and adaptive functioning for people of all ages. This system also includes the Child Behavior Checklist for Ages 1 ½-5 (CBCL/1 ½-5), Caregiver-Teacher Report Form for Ages 1 ½-5 (C-TRF), Teacher's Report Form (TRF), the Youth Self-Report (YSR), the Direct Observation Form (DOF), the Semi-structured Clinical Interview for Children and Adolescents (SCICA), and the Test Observation Form (TOF).

Development

According to Achenbach, the developer of the CBCL, "It was the lack of satisfactory constructs and operational definitions for childhood disorders that prompted us to develop the CBCL in order to assess parents' perceptions of their children's competencies and problems" (1991, p. 83). Achenbach (1991) reported that competencies can be as important as problems for understanding children and adolescent's adaptive development. Therefore, researchers developed the CBCL following a two-stage process that addressed both the competencies and the problems of children and adolescents.

The first stage of the CBCL's development involved a literature review of the assessment of competence in order to obtain candidate items and then pilot tested descriptions of positive characteristics in various formats. Prior to this development, there was little research to confirm which competencies reportable by parents discriminated between those children who were adapting and those children who were identified as needing help for behavioral or emotional problems (Achenbach, 1991).

The second stage of the CBCL's development was to develop the procedures for assessing child and adolescent behavioral and emotional problems. Researchers began by obtaining descriptions of problems that concerned parents and mental health professionals. These descriptions were obtained from earlier research, reviews of the clinical and research literature, and consultation with clinical and developmental psychologists, child psychiatrists, and psychiatric social workers. The CBCL problem section was pilot tested and revised through a series of nine editions completed by parents of children seen in a variety of settings from 1970 to 1976. Parents' ratings were then

used to derive syndromes of co-occurring problems through factor analysis, and these empirically-based syndromes were used to construct scales by which to establish statistical norms.

Using the CBCL/6-18

The CBCL/6-18 is filled out by a child or adolescent's parent or other significant adult figure. In filling out the CBCL/6-18 competencies portion, an adult is asked to rate the child or adolescent on how well or how often, as compared to other children or adolescents of the same age, he or she engages in: sports, hobbies, organizations, chores/jobs, friends, family relationships, and academics.

For the problems portion of the CBCL/6-18, a parent is asked to rate the child or adolescent on 118 problems that may be observed in their child or adolescent presently or within the last six months. The ratings are on a 3-point Likert scale (0 = Not true, as far as you know; 1 = Somewhat or sometimes true; 2 = Very true or Often true). The three-step response scale was chosen because, according to Achenbach (1991), it is usually easier to fill out than a *present/absent* scale for untrained raters such as parents and "more finely differentiated response scales were rejected because fine gradations in problems are unlikely to be captured by a questionnaire" (p. 14). Achenbach further stated that scoring problem items on more differentiated scales became vulnerable to respondent characteristics, and which reduced the discriminative power of items below that obtained with the three-step scales (Achenbach, Howell, Quay, & Conners, 1991). Furthermore, those scales did not increase differentiation of syndromes derived from ratings of behavioral or emotional problems (Achenbach & Edelbrock, 1978). Some items, in

addition to being rated a 0, 1, or 2, request the parent to provide a brief description of the problem to provide potentially valuable information to the clinician.

The CBCL/6-18 usually takes 15-17 minutes to complete. The Youth Self Report (YSR) is utilized by 11- to 18-year olds to report their own competencies and problems. It has many items in common with the CBCL/6-18 but is tailored specifically to the child or adolescent acting as her or his own informant. The CBCL/6-18 is self-administered, so it does not take any clinician time for administration, and scoring is most frequently accomplished quickly by computer. Forms cost 60 cents each; there is no per-use charge for scoring or administration by computer software (http://www3.parinc.com/products/product.aspx?Productid=CBCL-S). The purchase price for the computer scoring software is $345. Also, a Web-Link makes the need for supplies of forms obsolete (Achenbach & Rescorla, 2004).

Content Domains

The CBCL/6-18 competencies section is scored on the following scales: Activities, Social, School, and Total Competencies. The Total Competencies is the sum of the other three scales. The CBCL/6-18 problems section contains two major headings: Internalizing and Externalizing. Internalizing problems are identified as: Anxious/Depressed, Withdrawn/Depressed, and Somatic Complaints syndromes. According to Achenbach and Rescorla (2004), these primarily reflect problems within the self. Externalizing problems are identified as: Aggressive Behavior and Rule-Breaking syndromes, which the authors indicate as primarily reflecting conflicts with other people and with social mores. Those scales not included under the internalizing and externalizing rubrics are Social problems, Thought problems, Attention problems and

Other problems. Critical Items are listed on reports by the software scoring program and they identify items of particular clinical concern. The CBCL/6-18 also has diagnostically-oriented scales that are designed to discriminate between people referred for mental health services and those who have not been referred for such services: Affective Problems, Anxiety Problems, Somatic Problems, Attention Deficit/Hyperactivity Problems, Oppositional Defiant Problems and Conduct Problems. The Total Problems score is the sum of all the problem scales and is the most global index of psychopathology on the CBCL/6-18 (Achenbach & Rescorla, 2004).

Psychometric Properties

Reliability. Internal consistency estimates for the CBCL using a test/re-test interval of seven days produced a mean correlation of 0.87 for all of the competence scales, while the mean for the total competence scale is 0.87. The mean of all of the problem scales is 0.89, while the total problems scale mean is 0.93 (Achenbach, 1991). When using test/re-test assessments obtained over an 8- to 16-day interval, internal reliability estimates of all the CBCL scale scores are 0.90 (Achenbach & Rescorla, 2004). These high estimates of internal consistency indicate that the CBCL/6-18 is a reliable measure.

Validity. Content validity of the CBCL is supported by the numerous competencies and problems it assesses that are of clinical concern to parents and mental health workers. Construct validity of the CBCL is supported by the correlations of its scales with analogous scales on the Quay-Peterson (1983) Revised Behavior Problem Checklist and Conners' (1973) Parent Questionnaire. Criterion-related validity is supported by the ability of the CBCL to predict membership in a clinical (inpatient and

outpatient) or normal population with an average classification accuracy of 82.5%. If a borderline group is allowed, the accuracy of classification increases to 89.1% (Achenbach, 1991). More recent research (Burlingame et al., 2001) has found varying levels, indicating the CBCL will correctly identify members of clinical population 75% of the time and members of a normal population 87% of the time. Validity estimates range from 0.59-0.86 (Achenbach, 1991).

The CBCL appears valid for change. It has been used successfully as an outcome measure in hundreds of studies. See for examples: (Crawford, Field, Fisher, Kaplan, & Kolb, 2004; Larson, 1998; Packman, 2002; Seligman, Ollendick, Langley, & Baldacci, 2004; Webster-Stratton, 1984). Effect sizes in these studies are within the *large* and moderate ranges.

Cut-off scores. The CBCL makes good use of T-scores. A separate cutoff T-score exists for each scale: the competence scales, the syndrome scales, the internalizing and externalizing scales, and the total problem scale. In each case there is a borderline clinical range below the cut-off which separates the normal and clinical ranges. For the competence scales, T-scores below the borderline range are within the clinical range, while all scores above it are within the normal range. For all other scales, T-scores below the borderline range are within the normal range, while T-scores above it are in the clinical range (Achenbach, 1991). This difference in scoring is due to normal subjects having greater numbers of competencies while having fewer problems. The CBCL does not have a raw cut-off score in order to clearly ascertain a client's status by simply summing item scores, yet on the Total Problems scale T-scores of 60-63 are classified as *borderline* and scores of 64 and above are classified as *clinically significant.*

Reliable change index. In the appendix of each ASEBA manual the authors have provided information by which clinicians can assess the quality of change they observe on the scale scores of the ASEBA instruments. This information includes the standard error of measurement (*SEM*) for each scale separately for samples of referred and non-referred children and adolescents of each age and gender, as well as by each type of informant from which the data were obtained. If the change in scale score exceeds one SEM, then the change observed exceeds that which would be expected by chance 68% of the time. If clinicians desire a 95% confidence interval, then they can simply multiply the SEM by 1.96.

Although the authors indicate that Jacobson and Truax (1991) have suggested the use of the RCI as a statistical basis for documenting changes from pre- to post-treatment assessments, it does not appear that they have calculated this statistic for the CBCL. A study performed by Behrens and Satterfield (2006) used the Jacobson and Truax method for evaluating RCI with the CBCL/6-18, and they indicated that it defined a range of two standard deviations using raw score points (they did not use T-scores), 21 points on the CBCL/6-18, and 29 points on the YSR.

Use as an Outcome Measure

The CBCL has been used to provide broad measures of change after treatment in clinical research designs. By 1999, there were already over 300 publications reporting treatment research using the ASEBA instruments (Newman, Ciarlo, & Carpenter, 1999).

The CBCL can also be used in evaluating outcomes for individual children. It can be used to track treatment progress, as well as providing pre- and post-assessment outcomes. To monitor treatment for individual children and adolescents the authors

advocate using the instrument over uniform intervals appropriate for the treatment, "such as every 3 months" (Achenbach & Rescorla, 2004 p. 203). If clinicians choose to administer the CBCL more frequently than the 6 months recommended in the standard instructions, the authors indicate the rating interval used in considering the first administration should also be shortened in order to maintain even intervals. However, in 1991, Achenbach recommended allowing at least 2 months between assessments, both to minimize possible "practice effects" and to allow time for "behavioral changes to occur and become apparent to raters" (p. 74). Achenbach (1991) also indicated: "Because of the time required for behavioral change to stabilize and become clearly recognized by parents, rating periods of less than 2 months are probably not worth using" (p. 228). More recently, Achenbach and Rescorla (2004) altered this time-frame, indicating that the CBCL should "probably not be readministered at intervals of less than about 1 month" (p. 203). The aspects of functioning it measures take time to change (i.e., it measures static, rather than dynamic, variables). Moreover, as previously noted for the recommended 2-month interval, time is needed for changes to stabilize and for informants to become aware of the changes.

The results of these repeated administrations can be used to track functioning in relation to scale score norms for the child or adolescent's age, gender, and type of informant (if using other forms). The data provided by the CBCL allows clinicians to see actual change in scale scores and whether scores have moved from the clinical range to the borderline or normal range (Achenbach, 1991; Achenbach & Rescorla, 2004).

Sensitivity to change. The CBCL suggested cut-off of 60 has a sensitivity index of 0.754 and a specificity of 0.870 (Burlingame et al., 2001); this means it will correctly

identify those within the clinical range 75% of the time and those within the sub-clinical or normal range 87% of the time.

The child and adolescent clinical research literature on outcome measures includes statements regarding the CBCL's sensitivity to change resulting from psychotherapy. The CBCL is a traditional measure that was originally designed to measure relatively stable dimensions, not to measure therapeutic change (Berrett, 2000). Without evidence of sensitivity to change, the use of traditional measures, such as the CBCL, to assess psychotherapeutic change may be inappropriate, as their lack of sensitivity to change often prohibits the measure from demonstrating therapeutic change, even when significant change has occurred (Berrett, 2000). Drotar et al. (1995) have discussed the limited sensitivity to change of the CBCL, yet it is still commonly employed as an outcome measure in child and adolescent therapy research and its authors advocate its use as such (Achenbach & Rescorla, 2004).

For Drotar et al. (1995), the main concern with the CBCL's sensitivity to change is its restricted ability to accurately measure ratings below the clinical level. Clinicians anticipate their treatments will facilitate a child or adolescent's change from the clinical level to the sub-clinical, or normal, level. If an assessment is done once the adolescent has dropped below the clinical level (i.e., has only mild symptoms), the CBCL may no longer provide "sensitivity for detecting variation" (Drotar et al., 1995, p. 185). Therefore, Drotar et al. (1995) give the following recommendation:

> Researchers should be cautious about making inferences based on scores from the CBCL and related instruments that are within the normal range. Investigators who are interested in detection of more subtle adjustment problems or assessment

of psychological competence may wish to consider additional or alternative instruments that have been designed specifically for this purpose. (p. 190)

Several aspects of the CBCL's design may contribute to concern about its sensitivity to change. The CBCL contains a Likert scale ranging from 0 to 2, as discussed previously. A relatively narrow range for responding, such as this, may not be sensitive to changes that take place in a child or adolescent's functioning (Lipsey, 1990). Additionally, the instructions in the CBCL instruct the parent to answer each item based on the child's behavior now or in the past six months. Since psychotherapeutic treatments in today's managed care era tend to be brief, with many clients only receiving 4-20 sessions, changes that occur as a result of therapy may not be reflected in the results since parents are completing the CBCL in relation to the past six months, encompassing all of the child or adolescent's time in therapy as well as pre-therapy functioning (Berrett, 2000). This problem can produce possible confusion that can confound the results.

There are also concerns about how scoring may influence the CBCL's sensitivity. Subjective scoring for items requiring open-ended descriptions and suppression of T-scores for clients in the non-clinical range (i.e., all raw scores below the 69th percentile are assigned the same T-score) further inhibit the measure's sensitivity to distinctions between those with mild symptoms who are in the sub-clinical range (Drotar et al., 1995). Achenbach (1991) suggested using raw scale scores in statistical analyses rather than T-scores to address the problem of suppression for those in the non-clinical range, because the raw scores reflect all differences among individuals. However, this method reduces the ability make comparisons across age and sex groupings (Achenbach, 1991) and does not solve the problem of how to interpret the differences that are observed (Drotar et al.,

1995). Furthermore, regarding the content of the CBCL relative to outcome, Mosier (2001) indicates that social competence (i.e., school failure, special education, relationship with parent, physical or mental disabilities) is the only domain covered by the CBCL that is relevant to outcome (Mosier, 2001).

The Behavior Assessment System for Children-2 (BASC-2)

The BASC-2 (Reynolds & Kamphaus, 2004) is a recent revision of the Behavior Assessment System for Children (Reynolds & Kamphaus, 1992). The BASC-2 was designed to "facilitate the differential diagnosis and educational classification of a variety of emotional and behavioral disorders of children and to aid in the design of treatment plans" (Reynolds & Kamphaus, 2004, p. 1). The BASC-2 is a multi-method system for ages 2-21; its components may be used individually or in any combination. It contains a Self-report of Personality (SRP); Parent Rating Scale (PRS); Teacher Rating Scale (TRS); Structured Developmental History (SDH); and Student Observation System (SOS).

Development

The development of the original BASC involved several stages. The conceptualization of the measurement system was based on a comprehensive review of behavior-rating and self-report instruments, a goal to assess both adaptive and maladaptive behaviors, and consultations with child and adolescent clinicians (Reynolds & Kamphaus, 2004). Item content came through consultations with teachers, parents, and children; psychologists; and reference sources, such as the Diagnostic Statistical Manual. Once item content was established, there were two item tryouts in 1986 and 1987 involving teachers, parents, and children from Kentucky, Nevada, Texas, Georgia,

California, Florida, Minnesota, and Ohio that preceded the final selection of items (Reynolds & Kamphaus, 1992).

Standardization of the BASC was accomplished by selecting a much larger general sample that was representative of the United States through the regions of Southwest, South, North Central, and Northeast. The sample sizes included thousands of teachers, parents, and children (Merenda, 1996). Statistical analysis consisted of item-to-scale correlations as well as confirmatory factor analysis (Reynolds & Kamphaus, 2004).

The first step in creating the BASC-2 was to make a comprehensive review of the original BASC with the goal of increasing the consistency of item content between the TRS and the PRS across age levels for each form (Reynolds & Kamphaus, 2004). Some items that had previously been omitted for a particular age group were re-included. Also, items that appeared only on the TRS of the original BASC but were deemed reasonable for both school and home settings were added to the PRS and vice versa. New items were also written for all BASC scales, with particular attention to those scales with reliabilities that were not as high as the authors desired. Finally, several new scales were created to broaden the content domains and allow closer comparisons between forms.

Using the BASC-2

The BASC-2-PRS is completed by a child or adolescent's parent or other significant adult figure. The other forms (SRP and TRS) are filled out by the child and their teacher, respectively. An adult is asked to read phrases that describe how children may act and then rate their child's behavior in the last several months relative to the phrase. The questions are based on a four-point Likert scale. The four-choice response format uses letters instead of the standard numbers. Parents are asked to circle N for

Never, S for Sometimes, O for Often, and A for Almost Always in response to the behaviors they have observed. The BASC-2-PRS has three versions with varying numbers of items for different age groups (Preschool, 134 items; Child, 160 items; Adolescent, 150 items) and takes approximately 10-20 minutes to complete. The TRS is similar to the PRS, although it usually requires 10-15 minutes to complete. The SRP form asks children and adolescents to describe their emotional-responses and self-perceptions. The question format consists of true/false questions and the four-point Likert scale as described for the PRS and TRS. The SRP is slightly longer but has three forms that vary by age: ages 6-11, 139 items; ages 12-21, 176 items; and ages 18-25, 185 items. The SRP takes approximately 30 minutes to complete.

In hand scoring the BASC, the letter responses of N, S, O, and A correspond to the scores 0, 1, 2, and 3 points, respectively, while scoring via computer the letters correspond to 1, 2, 3, and 4, respectively. The BASC system is self-administered so it does not take any clinician time for administration, and scoring is most frequently accomplished quickly by computer (Reynolds & Kamphaus, 2004); this takes approximately 5 minutes (Gladman & Lancaster, 2003). The BASC system also has hand-scored forms that have a built-in scoring system that is revealed when the clinician separates the two parts of the carbonized form to reveal an inner page with the items already scored; calculating the scale and composite scores by hand follows easily but can be time intensive (Gladman & Lancaster, 2003). The hand-scored forms cost $1.34 each, while the computer scored non-scannable forms are $1.12 each; the computer scored scannable forms are $1.76 each (http://ags.pearsonassessments.com/Group.asp?nGroupInfoID=a30000). A variety of

software programs with a variety of features allow clinicians to select the one that is most appropriate for their needs. For example, the BASC ASSIST software to score non-scannable forms costs $259.00, while the version used to score scannable forms costs $605.00.

Content Domains

The BASC-2-PRS includes the following scales: Activities of Daily Living, Adaptability, Aggression, Anxiety, Attention Problems, Atypicality, Conduct Problems, Depression, Functional Communication, Hyperactivity, Leadership, Learning Problems, Social Skills, Somatization, Study Skills, and Withdrawal. The TRS and PRS overlap, but each also possesses unique aspects geared toward the specific informant. These scales are used to form the following composite scales: Externalizing problems, Internalizing problems, Adaptive Skills, and Behavioral Symptoms Index (Reynolds & Kamphaus, 1992, 2005).

Psychometric Properties

Reliability. Internal consistency estimates for the BASC-2-PRS using a test/re-test interval of 9-70 days produced mean correlations from 0.78-0.92 for the composite scales across all three age groups. Reliability of the BASC-2-PRS composite scales is estimated to be very high, ranging from the low to middle 0.90s using coefficient alpha. These reliability estimates are quite consistent between females and males, between clinical and non-clinical groups, and at different age levels (Reynolds & Kamphaus, 2004). These high estimates of internal consistency indicate that the BASC-2 is a reliable measure. Median inter-rater reliabilities are slightly lower at 0.74, 0.69, and 0.77 for preschool, child, and adolescent levels respectively, although this is not unexpected

according to general research on inter-rater correlations (De Los Reyes & Kazdin, 2004). The BASC-SRP reliability also was high.

Validity. Content validity of the BASC is supported by the numerous competencies and problems it assesses that are of clinical concern to parents, mental health workers, teachers, and children. Construct validity of the BASC-PRS is supported by the correlations of its scales with analogous scales on Child Behavior Checklist (Achenbach, 1991) and with externalizing scales of the Conners' Parent Rating Scales (Conners, 1989). The Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1943 [renewed 1970]), Achenbach's Youth Self-Report (Achenbach, 1985), and the Behavior Rating Profile (Brown & Hammill, 1983) showed a number of high correlations with the BASC-SRP scales. Criterion-related validity is indicated by the authors (Reynolds & Kamphaus, 2004), although an average classification accuracy is not presented. Validity is also supported by scale inter-correlations and factor analysis for the grouping of scales into composites. Validity estimates for the BASC-2 do not appear to be presented by the authors in the manual (Reynolds & Kamphaus, 2004).

The original BASC appears valid for change. It has been used successfully as an outcome measure in hundreds of studies (see for examples, Evans, Axelrod, & Langberg, 2004; Lehner-Dua, 2002; Packman, 2002). Effect sizes in these studies are within the *large* range, with some subscale effect sizes falling within the *moderate* range, and a few subscales falling in the *small* range. Although the BASC-2 is new, it is expected that it, too, is valid for change as its psychometric properties are improved from those of the original BASC.

Cut-off scores. General norms for the BASC are based on a large national sample representative of the general population with regard to age, gender, ethnicity, and clinical or special education classification (Reynolds & Kamphaus, 1992, 2004). Normative scores are provided for each scale of the BASC including a T-score and a percentile. Notably, the BASC makes use of Linear T-scores, so it is not appropriate to interpret these T-scores in terms of the normal distribution. The T-scores must be interpreted in light of their corresponding percentiles because the relationship with linear T-scores and percentiles varies with the shape of the score distribution (Reynolds & Kamphaus, 2004). The BASC does not have a raw cut-off score in order to clearly ascertain a client's status by simply summing item scores, yet T-scores of 60-69 are classified as *at-risk* and scores of 70 and above are classified as *clinically significant.*

Reliable change index. The BASC-2 does not make use of the RCI in order to allow clinicians to easily evaluate the reliability of change in their clients' symptoms and behavior.

Use as an Outcome Measure

The results of repeated administrations of the BASC can be used to track functioning in relation to scale-score norms for the child or adolescent's age, gender, and type of informant (if using other forms). The data provided by the BASC allows clinicians to see actual change in scale scores and whether scores have moved from the clinical range to normal range.

Although the original BASC had some limitations for use in outcome assessment because it did not contain enough items to assess changes in the patterns of illicit substance abuse or other severe behavior problems (Kamphaus, Reynolds, Hatcher, &

Kim, 2004), it appears that the authors corrected for those limitations in the BASC-2. However, outcome studies with this newer version are as yet unavailable.

Sensitivity to change. There are no published data on sensitivity to change for the BASC or BASC-2. There have been several explorations of the BASC's sensitivity in assigning diagnoses (Doyle, Ostrander, Skare, Crosby, & August, 1997; Ostrander, Weinfurt, Yarnold, & August, 1998), but this type of sensitivity cannot be considered synonymous with the measure's sensitivity in assessing changes due to a psychotherapeutic intervention.

The Youth Outcome Questionnaire-2.01 (Y-OQ-2.01)

The Youth Outcome Questionnaire (Y-OQ-2.01) is one of a family of measures that has been developed by researchers at Brigham Young University in collaboration with several managed care organizations that oversee health care throughout the Western United States. The Y-OQ-2.01 is a measure that assesses child and adolescent client improvement after a therapeutic intervention. The OQ family of instruments also includes measures to track adult outcome [OQ-45.2 (Lambert & Burlingame, 1993), OQ-30.1, and OQ 10.2]; several screening tools for primary care settings (OQ-PCM and Y-OQ-PCM), and a prognostic tool (Y-OQ-PA). There is a self-report version (Y-OQ SR-2.0) as well as several shorter versions of the Y-OQ (Y-OQ-30.1 and Y-OQ-12) that are currently under development.

The Y-OQ-2.01 was originally designed as the child and adolescent equivalent to the OQ-45.2 (Lambert et al., 1996). It was constructed specifically to track treatment progress (track actual change in client functioning), as opposed to assigning diagnoses, guiding treatment planning, or other such purposes (Burlingame et al., 2004).

Development

Lipsey (1990) has noted: "The importance of having valid, reliable and sensitive dependent measures in treatment effectiveness research is so great that it will generally warrant considerable advance preparation" (Lipsey, 1990, p. 103). Reisinger and Burlingame (1997) further indicated that, since sensitivity to change is a relatively new concept to psychometrics, many older instruments were not designed or evaluated with that feature in mind. Whereas the CBCL and BASC were carefully designed to assess the areas of problems and competencies and then later adopted for use in outcome measurement, the Y-OQ was envisioned to be an outcome measure from its inception, and this vision guided its development.

The Y-OQ was specifically developed as a response to the health care industry's demands for measuring outcome in mental health treatment; to serve as a quality indicator for managed care providers, third-party payers, and accrediting agencies; and as an instrument individual clinicians could use to track patient progress. The authors of the Y-OQ (Burlingame et al., 1996; Wells et al., 1996) have stated that it was specifically constructed to assess the occurrence of observed behavior change, to be brief, to be sensitive to change over short periods of time, and to be used on a session-to-session basis while being available at a nominal cost and maintaining high psychometric standards of reliability and validity.

The development of the Y-OQ followed a multi-stage process in order to meet the above-stated criteria. Y-OQ researchers (Burlingame et al., 1996; Wells et al., 1996) first performed literature searches of narrative and meta-analytic reviews regarding child clinical treatments in order to identify content domains that had been shown to be

empirically sensitive to change. Among those identified, only those content domains in which the average treated client demonstrated improvement of one-half a standard deviation were included for scale development. Second, two types of focus groups were conducted: the first was with children and adolescents, who had been treated in outpatient and inpatient settings of a managed health care organization, and their parents; the second was with psychiatrists, clinical psychologists, social workers, and other support staff from inpatient and outpatient settings. The inpatient provider focus group resulted in a Critical Items subscale; items these providers suggested were sensitive to change occurring during the average inpatient hospitalization. In the third stage of the Y-OQ development, researchers examined 100 treatment charts from inpatient and outpatient sites to identify recorded treatment-related change related to stated therapeutic goals.

Using the Y-OQ-2.01

The Y-OQ-2.01 is a 64-item instrument designed to measure the level of current distress a child or adolescent (age 4-17) is experiencing. Parents rate items on a five-point Likert scale (0=Never or Almost Never to 4=Almost Always or Always). Scores range from -16 to 240 (negative Y-OQ scores are possible because items that assess adaptive behavior are reverse-scored and can yield negative numbers). The self-report version (Y-OQ SR-2.0) is for adolescents 12-18 and is formatted in like manner. The Y-OQ-2.01 takes approximately 5-7 minutes to complete and is scored with use of computer scoring software; it can also be scored quickly by hand. The Y-OQ-2.01 can be purchased with a one-time nominal licensing cost as opposed to a fee-per-administration basis. This fee varies, based on the size of the group obtaining the license. A private-practice clinician can obtain the license for a price of $75, while a small group of clinicians pays $250

(http://www.oqfamily.com/LicenceAggrement2005.pdf). Larger groups pay a higher licensing fee. Once this license is obtained users are not required to purchase forms but are allowed to photocopy the measure for their use as agreed upon within the license.

Content Domains

Six content subscales were included in the Y-OQ-2.01 based on the results of the multi-stage developmental process. These subscales tap diverse areas of behavioral difficulties as well as elements of healthy behavior: Intrapersonal Distress, Somatic, Interpersonal Relations, Social Problems, Behavioral Dysfunction, and Critical Items. The total score on the Y-OQ-2.01 is the summation of items from each subscale and is designed to reflect the total amount of distress a child or adolescent is experiencing. The total score is the best index to track a client's global change and has the highest estimated reliability and validity compared to the reliability and validity of the individual subscales (Wells et al., 1996).

Psychometric Properties

Reliability. Internal consistency estimates of the Y-OQ-2.01 subscales based on normative samples range from 0.74-0.93, while the total score estimate is 0.97. This total score estimate suggests the measure assesses a strong single factor, which is useful for clinicians since this is the score typically used in order to track client change (Wells et al., 1996). Also, internal consistency reliability estimates were calculated for each of the normative samples (Burlingame et al, 2004). In these samples, the Y-OQ-2.01 total score demonstrated high internal consistency estimates of 0.94 across all normative sample settings (Burlingame et al., 2001). Additionally, Gironda (2000) calculated an internal consistency estimate for the total Y-OQ-2.01 score of 0.95.

Validity. Construct validity of the Y-OQ is evident in the differences between average total scores from community normal, outpatient, and inpatient samples. Community normal participants had the lowest scores and inpatient clients had the highest scores, while the outpatient sample's mean score fell between the other two. Also, the Y-OQ is able to predict membership in a clinical (inpatient and outpatient) or normal population with an average classification accuracy of 85 percent (Burlingame et al., 1996; Wells et al., 1996) based on total means using traditional cut-off scores. Criterion-related validity is supported by high correlations between the Y-OQ total and subscale scores and other measures used frequently for outcome, such as the Child Behavior Checklist (Achenbach, 1991) and the Conners' Rating Scale (Conners, 1990).

The Y-OQ appears valid for change (see for examples, Clark, 2002; Crawford et al., 2004). Effect sizes in these studies are within the *large* range. The Clark (2002) study, which examined the effects of wilderness therapy for adolescents, revealed that, while each of the scales from the Million Adolescent Clinical Inventory (MACI; Millon, Millon & Davis, 1993) produced no effect size to moderate effects sizes (0.02-0.75), the total score of the Y-OQ-2.01 produced a large effect size of 1.87.

*Cut-off sc*ores. The Y-OQ-2.01 makes use of a cut-off score. To identify whether a child or adolescent is within the clinical or normal range, a score of 46 distinguishes the cut-off. All scores above the cut-off are in the clinical range, while all scores below the cut-off are in the sub-clinical or normal range (Wells et al., 1996). The cut-off score for the Y-OQ SR-2.0 is 47 (Wells, Burlingame, & Rose, 2003).

Reliable change index. To evaluate change in scores between administrations of the Y-OQ-2.01, the developers have calculated an RCI value of 13 points. Therefore, a

subject's score must change by at least 13 points in order for the change to be considered reliable (Burlingame et al., 1996). If the RCI is greater than 13 points, the probability is less than .05 that the mean difference between the subject's scores occurred by chance (Mosier, 2001). RCIs have also been calculated for each of the subscales. The Y-OQ SR-2.0 has an RCI value of 18 points (Wells et al., 2003).

Use as an Outcome Measure

The Y-OQ was developed for use as an outcome measure. The results of repeated administrations can be used to track functioning in relation to scale-score norms for the child or adolescent's age, gender, and type of informant (if using other forms). The data provided by the Y-OQ allow clinicians to see actual change in scale scores and whether scores have moved from the clinical range to the normal range.

*Sensitivity to chang*e. Sensitivity to change over brief periods of time is evidenced by reliable decreases in Y-OQ total scores over the course of therapeutic treatment for children and adolescents (Burlingame et al., 2001; Mosier, 1998). In a study by Burlingame et al (2001) in which the RCI value of 13 and cut-off score of 46 were utilized to assess sensitivity to change in a combined clinical sample, 147 (17%) subjects were designated as recovered, 308 (37%) as improved, 260 (31%) as unchanged, and 125 (15%) as deteriorated. The average amount of change between pre- and posttest scores was 17.7 points; this change is greater than the 13-point RCI value and reflects a change of greater than 2 points per week.

The Y-OQ's sensitivity to change is also evidenced by its ability to better discriminate between children and adolescents with varying severity levels via sensitivity and specificity analyses than the CBCL (Burlingame et al., 2001). According to

Burlingame et al. (2001) the Y-OQ cut-off score of 46 has a sensitivity index of 0.820 and a specificity of 0.894; this means it will correctly identify those within the clinical range 82% of the time and those within the sub-clinical or normal range 89% of the time. Whereas, as indicated previously, the CBCL suggested cut-off of 60 (Achenbach & Edelbrock, 1991) has a sensitivity index of 0.754 and a specificity of 0.870; this means it will correctly identify those within the clinical range 75% of the time and those within the sub-clinical or normal range 87% of the time.

Comparison of the CBCL/6-18, BASC-2 and Y-OQ-2.01

The CBCL/6-18 and the BASC-2 were designed to assess competencies and problems in order to assist diagnosis and treatment planning. The Y-OQ-2.01 was designed for use as an outcome measure and to track client progress. Although these measures were designed for varying purposes they are all commonly used for outcome purposes and are herein examined from that perspective.

Based on the above the reviews of the CBCL/6-18, BASC-2, and Y-OQ-2.01 Table 2.1 provides a relative comparison of all three measures in terms of the previously identified recommendations made by researchers and clinicians to aid consumers in choosing an ideal outcome measure. Some of the numbers presented are from previous versions of the measures, when present statistics were not available. The rating system employed uses the ratings of Poor, Moderate, and Good. These labels have been selected to summarize the previous descriptions. Each rating reflects how the measure compares to the other two measures on a particular dimension. The ratings of Restricted or Unknown have been used when a rating within the general continuum from Poor to Good was not clearly appropriate.

Table 2.1

Comparison of the CBCL/6-18, BASC-2 and Y-OQ-2.01 Based on Researchers' and Clinicians' Recommendations

	CBCL/6-18	*BASC-2*	*Y-OQ-2.01*
Recommended Reliability --.80	0.87-0.93	0.78-0.95	0.94-0.97
Recommended Validity *Moderate -- .50* *Excellent - -.75*	0.59-0.86	Not presented	0.85
Valid for Change *Poor -- .20* *Moderate -- .50* *Good - -.80*	Moderate to Good	Moderate to Good (for most scales)	Good
Sensitive to Change	Restricted	Unknown	Good
Normed *-Cut-off scores* *- RCI*	-T-score cut-off -No RCI	-T-score cut-off -No RCI	-Raw score cut-off -RCI
Can be Completed Quickly	15-17 minutes	10-20 minutes, (30 minutes for SRP)	5-7 minutes
Can be Scored and Interpreted Easily and Quickly	Good-(Computer) Moderate-(Hand)	Good-(Computer) Poor-(Hand)	Good-(Computer) Good-(Hand)
Provides Relevant Information (Content Domains) Regarding Client Change	Activities Social Competence School Competence Anxiety/Depression Withdrawl/Depression Somatic Complaints Social Problems Thought Problems Attention Problems Other Problems Critical Items	Activities of Daily Living Adaptability Aggression Anxiety Attention Problems Atypicality Conduct Problems Depression Functional Communication Hyperactivity	Intrapersonal Distress Somatic Complaints Interpersonal Relations Social Problems Behavior Dysfunction Critical Items

		Leadership Learning Problems Social Skills Somatization Study Skills Withdrawl	
Can be Used Frequently to Track Progress	Closest recommended interval is 1 month, most studies use 3 month intervals	Closest recorded interval for outcome use is 2 ½ months, most studies use 3 month intervals	Weekly
Cost Effective (i.e., Inexpensive)	Forms: $.60 Scoring program: $345.00	Forms: $1.12-1.76 Scoring programs: start at $259.00 -depends on desired features	Forms: Free Scoring program: starts at $75.00 -depends on size of practice

Statement of the Problem

Sensitivity to change is the most important feature of an outcome measure (Burlingame et al., 2005). In order to meet the needs of researchers, clinicians, and health care managers, the measure must be sensitive to intra-individual changes that occur as a result of a therapeutic intervention. Therefore, an outcome measure's sensitivity to change is vital in evaluating the effectiveness of psychotherapy for children and adolescents.

This review has examined three measures commonly used in child and adolescent outcome research and in clinical practice: the CBCL/6-18, the BASC-2, and the Y-OQ-2.01. These three measures all have excellent reliability and validity estimates. It also appears these measures are valid for evaluating change as evident by the many research studies that have employed them for outcome use and produced moderate to high effect sizes. However, there are concerns for the CBCL's sensitivity to change within the

normal ranges, and the degree of sensitivity to change for the BASC-2 is unknown. Furthermore, sensitivity to change has not been directly compared across the CBCL/6-18, BASC-2, and Y-OQ-2.01. A direct comparison of the sensitivity to change of these measures will allow researchers, clinicians, and health care managers to identify the measure that is most appropriately employed as an outcome measure. This would enable outcome measure consumers to maximize the benefits that have been discussed previously. Since sensitivity to change is the most important aspect of an outcome measure (Burlingame, et al., 2005), if these measures are not significantly different in regards to sensitivity to change, then outcome measure consumers can select a measure based on other attributes (such as number of items, frequency of administration, or use of additional forms such as that of a teacher-report) that they value for their particular purposes.

Thus, the primary focus of this study was to evaluate the relative sensitivity to change of the Child Behavior Checklist (CBCL/6-18), the Behavior Assessment System for Children-2 (BASC-2), and the Youth Outcome Questionnaire (Y-OQ-2.01) in order to identify the "best" measure for use in child and adolescent outcome research and to recommend that it becomes the standard outcome assessment tool; such goals have been suggested by researchers as desirable to advance outcome research (Froyd et al., 1996).

Analysis of Therapeutic Change

The statistical methods employed for outcome research are of the utmost importance since different methods can use the same data for a particular client and come to different conclusions about how to qualify the client (e.g., as improved, unchanged, or deteriorated; see, for example Speer & Greenbaum, 1995). The statistical methods for

evaluating change must be able to accommodate repeated measures (or multi-wave) data. To evaluate change that occurs at the individual level (intra-individual change), as well as the group level (inter-individual differences), the statistical method must operate at multiple levels. In other words, a statistical procedure for evaluating change must be able to use all available information while detecting change at the individual and group levels.

Univariate or multivariate analysis of variance (ANOVA or MANOVA) procedures have been commonly used by researchers to measure outcome on a continuous basis (Kazdin, 2003; Raudenbush & Chan, 1993). However, these methods are inappropriate when change studies contain unbalanced designs, missing data, time-varying covariates, or continuous predictors of rates of change (Ware, 1985). Such characteristics are common in large-scale longitudinal studies (Raudenbush & Chan, 1993). Also, within these traditional models, individual variation in change is only accounted for within the interaction of repeated occasions rather than being directly modeled (Byrfk & Raudenbush, 1987, 2002). Therefore, Hierarchical Linear Modeling (HLM) offers a more flexible analytic approach for assessing therapeutic change through repeated measures data (Raudenbush & Chan, 1993).

HLM estimates linear equations that are used to explain outcomes for clients as a function of their own individual characteristics, as well as the characteristics of the group they belong to (i.e., the type of treatment they are receiving; Arnold, 1992). Although HLM has been evolving since the 1970s, its application to the study of therapeutic change is more recent. In the early 1970s, Lindley and Smith (1972) and Smith (1973) developed the Bayesian method for the estimation of linear models with nested data and complex error structures (Arnold, 1992). Later, Dempster, Laird, and Rubin (1977) and

Dempster, Rubin and Tsutakawa (1981) developed the expectation-maximization (EM) algorithm to estimate the covariance components of linear modeling. At each level in the hierarchical model, the EM algorithm produces maximum likelihood estimates of the variance and covariance components (Arnold, 1992). In the late 1980s, HLM became more accessible to researchers with innovations in statistical computer programs (Arnold, 1992). With greater accessibility, many researchers in education and human development began to use HLM in longitudinal studies (Bryk & Raudenbush, 1987). In the 1990s, the mental health field recognized the benefits HLM had to offer and began using HLM in the study of therapy outcome (see, for example: Raudenbush & Chan, 1993; Speer & Greenbaum, 1995).

HLM operates as a two-level hierarchical model: first, analyzing multivariate data by computing individual growth curves, and, second, analyzing individual growth curves as a function of a group (Arnold, 1992; Raudenbush & Chan, 1993). This procedure is often referred to as *nesting* and is most often illustrated in the HLM literature with an example of an educational setting—students nested within their classroom, classrooms within their school, and schools within their school district. Nesting allows HLM to accurately predict change for members of groups while accounting for the attributes of both the member and the group (Arnold, 1992).

At level one, or the within-subject stage, each client's change is represented by an individual growth trajectory plus error that depends on a unique set of individual (person-specific) parameters (Bryk & Raudenbush, 1987; Raudenbush & Chan, 1993). These individual parameters become the dependent variables in a level two analysis, or the between-subjects stage (Raudenbush & Bryk, 2002). In other terms, regressions are

performed at the first level and the results of those regressions become the dependent variables in the regressions performed at the second level (Arnold, 1992). This two-stage model utilizes person-specific parameters, such as background or type of therapeutic treatment, to establish change trajectories which can be used to predict future change, study variation in change, and assess the quality of measurement instruments. In other words, the parameters of the first stage become the outcome variables in the second stage (Bryk & Raudenbush, 1987; 1992; Speer & Greenbaum, 1995), providing a model for analyzing individual and group patterns of change.

HLM has a number of advantages compared to traditional repeated measures techniques. The most important advantage is the greater precision due to the use of Bayesian estimation for assessing individual change in addition to group change (Raudenbush & Chan, 1993; Speer & Greenbaum, 1995). Another advantage, perhaps the most important for data collected within community mental health settings, is that the EM algorithm accounts for missing data, so subjects do not need to be dropped or data discarded due to limitations of the analytic model (Speer & Greenbaum, 1995). HLM also allows for increased flexibility in data requirements due to nesting (Raudenbush & Chan, 1993; Speer & Greenbaum, 1995). Clients may be assessed at different times and on a varied number of occasions because the repeated observations are hierarchical data; all observations are viewed as nested within the individual (Bryk & Raudenbush, 1987, 1992). Repeated observations decrease standard errors and provide consistent estimates of parameter correlations, such as the rate of change correlated with a client's initial status (Speer & Greenbaum, 1995). Thus, HLM makes better estimates of change by using all available information (Speer & Greenbaum, 1995).

There are important considerations for using HLM. Data must be highly reliable and valid because they form the basis for the second-level analysis. Data must be hierarchical, that is, units nested within groups. The groups must have enough within-subjects and between-subjects classifications to provide sufficient degrees of freedom. Large samples are recommended in the literature, but specifications on just how large they should be are essentially nonexistent (Arnold, 1992). The same holds true for the number of waves; typically more data points are considered better (Willett, 1989). Finally, HLM involves performing regression of regressions; therefore, the assumptions of linear regression that apply to causation cannot be applied (Arnold, 1992; Raudenbush & Bryk, 2002).

Hypotheses

Based on the literature review herein presented, the hypotheses of this study were:

1. The Y-OQ-2.01 parent- and self-report versions will be more sensitive to change over time and session number than the BASC-2 parent- and self-report versions for a sample of outpatient children and adolescents.
2. The CBCL/6-18 parent- and self-report versions will be less sensitive to change over time and session number than the BASC-2 parent- and self-report versions or Y-OQ-2.01 parent- and self-report versions for a sample of outpatient children and adolescents.

CHAPTER 3

Method

This study evaluated the relative sensitivity to change of the Child Behavior Checklist/6-18 (CBCL/6-18), the Behavior Assessment System for Children-2 (BASC-2), and the Youth Outcome Questionnaire (Y-OQ-2.01).

Measures

The CBCL/6-18, BASC-2, and Y-OQ-2.01 were each given to study participants, with two versions for each measure: an adult informant measure (parent-report form) and an adolescent self-report measure. The CBCL/6-18, BASC-2-PRS, and Y-OQ-2.01 were administered to adult informants while the YSR, SRP, and Y-OQ SR-2.0 were given to adolescents. Refer to page 21 through page 45 for a detailed description of these measures. For the purposes of this study, the competencies portion of the CBCL/6-18 was not used.

The Test-Taking Survey-Revised (TTS-R; McGrath, 2000) was used to infer validity of the following analyses. The measure is a face-valid 10-item scale with letter responses of N (Never), R (Rarely), S (Sometimes), F (Frequently), and AA (Almost Always) corresponding to the scores 1, 2, 3, 4, and 5 points, respectively, with reverse scoring on six items. The Test-Taking Survey was developed by Durham (1999, 2002) to assess mechanical responding and revised by McGrath (2000) to omit a qualitative section and add two questions relating to response style. Higher numbers on the TTS-R indicate greater conscientiousness and thoughtfulness in filling out the measures.

Setting

Valley Mental Health (VMH) is a community outpatient mental health facility that provides services for over 18,000 people each year in Salt Lake, Summit, and Tooele counties. VMH provides comprehensive services for children, adolescents, adults, and seniors. Services include: inpatient, residential, and outpatient services; substance abuse services, 24-hour crisis services, psychotropic medication management, forensic services, case management, consultation, education services, and prevention services. VMH also provides specialized services, such as for children with autism. The data for this study was collected at VMH's children's outpatient services in Salt Lake City, Utah. Outpatient mental health services included: individual, family, and group therapies; medication evaluation and management; and crisis intervention. Services are provided by psychiatrists, psychologists, clinical social workers, marriage and family therapists, licensed practical nurses, and the like.

Sample

Outcome measures utilize various informants, most frequently the parent and child or adolescent. According to De Los Reyes and Kazdin (2004), reviews of the literature are consistent in showing little agreement between the ratings provided by parents and children, with correlation coefficients often in the .20s. This discrepancy can make it difficult to integrate data from multiple informants and can lead to differences in who is perceived to meet particular criteria (Offord et al., 1996). However, the CBCL/6-18, BASC-2, and Y-OQ-2.01 consider a multi-informant feature important (Achenbach & Rescorla, 2004; Kamphaus et al., 2004; Wells et al., 2003); each measure uses both parent- and self-report versions of their instruments on a regular basis to assess children

and adolescents. In addition, although the parent-report forms may be used more frequently than the self-report versions, Yeh and Weisz (2001) suggest that clinicians who consult with children and adolescents may glean a better sense of shared parent-child goals for treatment than if they consulted with the parent alone. This agrees with earlier research that found that children were better informants than their parents (Herjanic & Reich, 1997). Therefore, both parent- and self-report versions continue to be utilized within research and clinical settings.

Due to this continued use of both parent- and self-report versions of the measures, data was collected from adults and adolescents. Children and adolescents, ages 6 to 17, were included in the study to assess the widest range of ages allowable by the three measures. These subjects were new clients beginning psychotherapeutic treatment at VMH. All parents or significant adult figures of clients meeting the age requirement who agreed to participate in the study were included. Adolescents who had consent of a parent or legal guardian were also included, even if their parent chose not to participate. The CBCL/6-18, BASC-2, and Y-OQ-2.01 were administered to 255 adults representing these VMH clients. Self-report versions of the measures were also given to 100 adolescents ages 12 and above.

Procedures

The researcher received approval to conduct the study with human participants from the Brigham Young University Institutional Review Board and the Utah Department of Human Services Human Subjects Committee. Intake procedures, treatment method, and treatment length were not altered in any way for the purposes of this study. At the intake session, the Y-OQ-2.01 and Y-OQ SR-2.0 were given via

personal digital assistants (PDAs) to all parents and adolescents, respectively, as part of VMH routine procedures. Parents and adolescents were informed of the purpose and procedures of the study after Valley Mental Health's intake session. Those who indicated a desire to participate were given a packet and invited to return to the intake room after their initial meeting with the therapist in order to enjoy snack foods and drinks while they completed the packet. Parents and adolescents who chose to participate and did return to the intake room signed a consent or assent form, respectively. They then filled out the first set of measures which included the BASC-2 and the CBCL/6-18 presented in random order to control for order effects. This initial packet did not include the Y-OQ-2.01 since the Y-OQ-2.01 is given via PDA as part of Valley Mental Health's routine intake procedures before they meet with their therapist. After participants filled out the BASC-2 and CBCL/6-18, the research assistant requested a print-out of the Y-OQ-2.01 from a Valley Mental Health receptionist and then added it to the packet of measures. Of those 678 adults that were present at VMH intake sessions during the course of data collection, 255 (37.6%) adults consented to participate and completed measures. One hundred adolescents also participated. Some of those present at intake were not eligible for the study since their child was younger than age 6.

The goal of the study was to obtain a maximum of five data points from each participant, with the respective measures administered as close to once-per-month as possible. This number of data points was selected because assessing sensitivity to change requires a repeated-measures clinical sample (Burlingame et al., 2001) and researchers have concluded that two time points do not provide adequate data for studying change that may be non-linear (c.f. Bryk & Weisberg, 1977; Rogosa et al., 1982). This study

sought enough data points to capture the changes that can occur during a psychotherapeutic treatment in real time.

The decision to collect the measures once-per-month was due to a number of considerations relating to the various measures. The Y-OQ-2.01 was designed to be administered as frequently as weekly, but can be used at any interval. However, the CBCL/6-18 and the BASC-2 were designed to be used at greater intervals than weekly. For example, Achenbach and Rescorla (2004) indicated the minimum amount of time between assessments for the CBCL/6-18 should be one month, although previously Achenbach (1991) indicated that period should be two months. Achenbach (1991) also indicated that if the interval is reduced much below 6 months, it may reduce scores on some problem items and scales slightly (such as running away or fire-setting), which could make it difficult to interpret scores by comparing them with established norms. Therefore, the appropriate timing of CBCL/6-18 administrations when it is used as an outcome measure is not clear. It was observed in the literature that most studies use the CBCL/6-18 at 3-month intervals. However, Achenbach and Rescorla most recently (2004) indicated that the CBCL/6-18 can be administered every month to examine its ability to detect change at more frequent intervals, so this timing was selected. BASC-2 authors (Kamphaus et al., 2004) have also recommended that it can be used to assess outcome after "brief intervention programs" (p. 349). Although the shortest interval observed for use as an outcome measure thus far appears to be 10 weeks (Merydith, 2000), it was determined to examine the BASC-2 administered on a monthly basis to assess its sensitivity to change as an outcome measure over brief intervals, as well as being consistent in administering it together with the other measures.

Therefore, in an effort to obtain data points as close to one month apart as possible, three weeks after a participant's intake, the research team began to obtain information on the participant's upcoming appointments at Valley Mental Health. Once an appointment was known, the research assistant traveling to Valley Mental Health on the day of the appointment contacted the participant (or the participant's parent if it was an adolescent-only participant) to arrange to meet with them in the lobby of VMH. These meetings were generally 30 minutes before their scheduled appointment, although more time was allowed for those who felt they needed longer to fill out the measures, including adolescents since the self-report form for the BASC-2 (SRP) is longer than the parent-report form. Participants received the CBCL/6-18, BASC-2, and Y-OQ-2.01 in a packet with the measures presented in random order to control for order effects. It required approximately 20- 40 minutes for participants to complete all three measures.

Due to the difficulty of securing repeated-measures data, researchers sought to limit attrition by providing graduated compensation to clients (Kazdin, 2003) upon completion of measures. Compensation was in the form of cash or gift certificates for use at local venues. Participants were offered a five dollar value for a second or third data collection point and a ten dollar value for a fourth or fifth data collection point. Adolescents were also compensated when they filled out the measures.

Furthermore, potential participants were notified in the consent form that if they chose to participate they gave permission to receive the measures by mail regardless of whether they continued with treatment at Valley Mental Health. If a follow-up data point was not obtained within a 3-month period, participants were contacted via email or phone, and measures were mailed upon receiving an affirmative response regarding their

willingness to complete them. This mailing packet included a letter, included herein as Appendix A, a self-addressed envelope with postage, measures, and a gift-certificate to compensate them for their time. This scenario was applicable for those who had discontinued treatment at Valley Mental Health, had transferred to another unit, or had continued to attend therapy but for various reasons the research team was unable to make personal contact to secure measures.

Due to the difficult nature of acquiring repeated-measures, in order to provide debriefing to the maximum number of participants, a debriefing statement was provided to each participant (parent and adolescent) at the third administration of measures, though efforts continued to obtain as many as five data points. Along with this debriefing statement, inquiry was made regarding participant attitudes with which they filled out the measures, and their conscientiousness in filling out the measures in order to infer validity of obtained results. This inquiry was accomplished with the use of the Test-Taking Survey-Revised (TTS-R). The revised Test-Taking Survey and debriefing statement as presented to the research participants is provided in Appendix B.

Scoring of Measures

This study sought to administer and score measures in a naturalistic manner, as would clinicians in routine practice. Therefore, the CBCL/6-18 and the BASC-2 were scored via publisher scoring programs purchased by Brigham Young University's Psychology Department. The Y-OQ-2.01 was scored using a scoring syntax provided by the developers of the Y-OQ-2.01, which scored the Y-OQ-2.01 as it would be scored by clinicians purchasing their software, yet without some of the features their software provides that were not needed for this study (such as graphs tracking individual client

progress). The CBCL/6-18 and the BASC-2 present results to clinicians in standardized T-scores; the Y-OQ-2.01 presents results to clinicians in summative raw scores.

Screening to Obtain Final Analytic Sample

Two discrete screening procedures identified the final analytic sample for this study. First, in order to calculate change, a minimum of two data points is required. Of the original sample of 255, 108 cases were omitted from the analytic sample because they had only one data point, leaving 147 cases. Second, due to the comparative nature and purpose of this study, no cases were included in the analysis that did not show reliable change based on RCI values (in either direction: improved or deteriorated) on any of the three measures, either by parent- and/or self-report versions of a measure. Cases showing reliable change on one or two measures at any point within treatment (i.e. between any data collection points, even if the change from pre- to post-treatment was not reliable) were retained in the analysis. Therefore, the next step in obtaining the analytic sample was to calculate respective RCI scores for the remaining 147 cases.

Burlingame et al. (2005) suggested that for tracking client changes with the Y-OQ-2.01 the Total score should be selected for use, rather than subscale scores. Thus, these analyses are based on changes in client Total scores. The RCI of 13 that has been calculated by authors of the Y-OQ-2.01 was used to evaluate changes in raw scores on the Y-OQ-2.01; the calculated RCI value of 18 was used for the Y-OQ SR-2.0. Since the BASC-2 and the CBCL/6-18 do not employ the use of RCI, these scores were calculated using the following formula (Tingey, Lambert, Burlingame, & Hansen, 1996):

$$RC_{index} = \frac{(pre\text{-}) - (post\text{-}treatment)}{S_{diff}} = 1.96 \text{ (alpha=.05)}$$

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

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

The r_{xx} is the mean of the correlations between all sets of half the items comprising a scale and was determined by the internal consistency estimates in the respective manuals provided as Cronbach's Alpha (r_{xx}). The standard deviation (*SD*), using the CBCL/6-18 and the BASC-2 T-scores, is equal to 10.

The representative scale of the CBCL, Total Problems, is from its empirically based scales. The Total Problems score is the sum of all the problem scales and is the most global index of psychopathology on the CBCL/6-18 (Achenbach & Rescorla, 2004). Achenbach and Rescorla (2001) indicate: "The Total Problems score can also be used as a basis for comparing problems in different groups and for assessing change as a function of time or interventions" (p. 192). For the parent-report of CBCL/6-18 the r_{xx} = .97, and for the adolescent self-report version (YSR) the r_{xx} = .95. Calculations produced an RCI of 4.80 for the CBCL and 6.20 for the YSR.

The representative scale for the parent-report version of the BASC-2-PRS is the Behavioral Symptoms Index (BSI), and for the self-report version of the BASC-2-SRP it is the Emotional Symptoms Index (ESI). Reynolds and Kamphaus (2004) indicated that the BSI and ESI reflect the overall level of a child's or adolescent's problem behavior and recommended that these overall composites be examined first when interpreting results. Chronbach's Alpha for the PRS: BSI and SRP: ESI vary by age of the child. For the PRS ages 6-7 and ages 15-18, r_{xx} = .94; for ages 8-11 and ages 12-14, r_{xx} = .95. For

the SRP ages 12-14, r_{xx} = .95; for ages 15-18, r_{xx} = .94. Calculations produced an RCI of 6.80 for those using r_{xx} = .94, and an RCI of 6.20 for those using r_{xx} = .95.

A case was identified for inclusion in the analytic sample if the absolute value of the calculated change was greater than, or equal to, the RCI criteria for at least one of the given measures, thus identifying cases where children and adolescents reliably improved and reliably deteriorated at any point during psychotherapeutic treatment. Although the RCI of 13 used for the Y-OQ-2.01 (18 for the Y-OQ SR-2.0) was based on raw scores, the analysis used T-scores for the CBCL/6-18 and BASC-2 due to an effort to use the measures as they would be used in naturalistic fashion within a clinical setting. A summary of the RCI's calculated for the CBCL/6-18 and BASC-2 from their T-scores is presented in Table 3.1.

Table 3.1

RCI Values for the CBCL/6-18 and BASC-2

Measure	*Informant*	*Ages*	*Alpha (r_{xx})*	*RCI*
CBCL/6-18	Parent (CBCL)	6-18	.97	4.8
	Youth (YSR)	12-18	.95	6.2
BASC-2	Parent (PRS)	6-7; 15-18	.94	6.8
		8-11; 12-14	.95	6.2
	Youth (SRP)	12-14	.95	6.2
		15-18	.94	6.8

Of the 147 cases, these RCI calculations recognized 136 cases that met change criteria for inclusion in the analytic sample. Of these cases, 92 cases were parent or adult informants only, while 42 had corresponding adolescent informants. In addition, there were two adolescent informants who were retained in the analysis whose parents did not participate.

Twenty-eight adult informants, 21% of the 136 cases retained in the analysis, completed one or more packets of measures via mail at some point during their tenure in the study. Eight of these 28 cases had corresponding self-report measures filled out by an adolescent. A *t*-test on the pair-wise test of differences between the change slopes of these participants and those participants that completed all sets of measures at Valley Mental Health with one of the study's research assistants indicated there were no statistical difference in the two groups ($p = .57$). Therefore, this group was included within the analytic sample for all comparisons, and not treated separately within the analysis.

As presented earlier, the BASC-2 employs the use of three parent-report forms, two of which are used in the present study: the Parent Rating Scales-Child (PRS-C) for ages 6-11 and the Parent Rating Scales-Adolescent (PRS-A) for ages 12-21. The PRS-C has 160 questions, 132 (82.5%) of which are identical to questions in the PRS-A. The PRS-C has 23 questions that are different from the PRS-A (14.38%) and five questions which are similar to questions on the PRS-A (two of them are related to the same PRS-A question; 3.13%). The PRS-A has 150 questions, 132 (88%) of which are identical to questions in the PRS-C, 13 (8.66%) of which are different from those in the PRS-C and five (3.33%) of which are related to questions found in the PRS-C. The PRS-C and PRS-

A have identical composites, primary scales, and content scales (Reynolds & Kamphaus, 2004). A statistical comparison between the slopes created by the PRS-C to the change slopes of the PRS-A indicated there were no statistical differences in the two groups. For cases where the child or adolescent improved during the course of the study, as identified by the Y-OQ-2.01, the *p* value was non-significant at .65; for those cases where the child or adolescent deteriorated during the course of the study the *p* value was non-significant at .79. Therefore, these measures were not treated separately within the analysis.

Analyses

After change was evaluated by calculating the RCI to obtain the analytic sample, the RCI was also used to evaluate pre- to post-treatment change that qualified children and adolescents *improved.* Change was further evaluated by using cut-off scores in tandem with the pre- to post-treatment RCI to evaluate how each of the measures classified changed cases as *recovered*.

Finally, to calculate effect size, Hierarchical Linear Modeling (HLM) was used to determine which of these measures may be most appropriate or most helpful for outcome use. HLM addressed this question by calculating individual slopes for each subject for the CBCL/6-18, the BASC-2 and the Y-OQ-2.01 (i.e., each individual had three slopes); these slopes represented rates of change. The analysis determined if there was a significance difference in the slopes. With statistically significant different slopes, the measure with the steepest negative slope was considered to be the most sensitive to change.

The CBCL/6-18, BASC-2, and Y-OQ-2.01 use different methods for interpreting clients' scores; the CBCL/6-18 uses T-scores, the BASC-2 uses linear T-scores, and the

Y-OQ-2.01 uses summation with raw scores. Therefore, the scores needed to be standardized in order to allow the slopes to be compared relative to each other; this was done by transforming them so they had equivalent scales. Linear T-scores vary from T-scores in that the distribution is not normalized, so the associated percentiles are different from those of T-scores. However, linear T-scores have a mean of 50 and a standard deviation of 10 as do T-scores, so no transformation was necessary for the CBCL/6-18 and BASC-2. The Y-OQ-2.01 data was transformed using the formula $10(Y-\overline{x})/SD+50$ where Y was equal to a participant's raw score on the Y-OQ-2.01, $\overline{x}$ was equal to the mean of Y-OQ-2.01 scores for a normal population, and *SD* was equal to the standard deviation of a normal population. The parent-report Y-OQ-2.01 mean and *SD* were 21.4 and 26.42, respectively (Burlingame et al., 2005) and the Y-OQ SR-2.0 mean and *SD* were 34.37 and 29.42, respectively (Ridge, Warren, Burlingame & Wells, 2007). This transformation gave the Y-OQ-2.01 and the Y-OQ SR-2.0 means of 50 and standard deviations of 10.

Thereafter, the data was analyzed according to the two models. Models were estimated using SAS for Windows (PROC MIXED; maximum likelihood estimation; version 9.1; SAS Institute, Inc., Cary, NC). The first model has a time variable; the second model replaces the time variable with a dosage variable, examining number of therapeutic contacts or sessions:

1. $Y = O + M + O * M + I + O * I + M * I + O * M * I + \log_{(D)} + O * \log_{(D)} + M * \log_{(D)} + O * M * \log_{(D)} + I * \log_{(D)} + M * I * \log_{(D)} + O * I * \log_{(D)} + O * M * I * \log_{(D)}$.

In analytic model number one, Y was equal to a participant's score on each of the measures, O was the outcome of the child's therapy (improved or deteriorated) as

provided by pre- to post-treatment data from the Y-OQ-2.01. Due to the requirements of the HLM analysis, one measure was required to be the standard to which the other measures were compared. The Y-OQ-2.01 was selected as the statistical standard due its use by Valley Mental Health as their intake measure, as well the presence of evidence for sensitivity to change as presented in the literature review. M was the method of measurement (CBCL/6-18, BASC-2, or Y-OQ-2.01) that represented a class variable that produced three different intercepts and slopes for the three different methods, I was the informant providing the data (parent- or self-report), and D was equal to the day number, with intake as day number one. The natural log function of the time variable ($\log_{(D)}$) was chosen, as other research suggests it is a better fit for this type of data than is a simple linear function (e.g., Warren, Nelson, & Burlingame, 2008).

The element O * M represented a two-way interaction examining a measure used to gather the data (CBCL/6-18, BASC-2, and Y-OQ-2.01) relative to a child's final or post-treatment score as improved or deteriorated from intake. O * I is the interaction between outcome and informant. M * I examined how the measures performed relative the informant used to provide the data. O * M * I represented a three-way interaction between a child or adolescent's outcome, the method of measurement, and the informant.

The element O * $\log_{(D)}$ represented a two-way interaction between the outcome of therapy, as established by a decrease or increase in Y-OQ-2.01 scores and day number averaged across the respective outcome classifications of improved and deteriorated. M * $\log_{(D)}$ represented a two-way interaction between the measure and the day number averaged across all three measures. O * M * $\log_{(D)}$ represented a three-way interaction between outcome, measure, and the day number, while the element I * $\log_{(D)}$ represented

a two-way interaction between the informant and the day number averaged across both types of informants. M * I * $\log_{(D)}$ represented a three-way interaction between the measure, the informant, and day number. O * I * $\log_{(D)}$ was the interaction between outcome, informant, and day number. The last element of the model, O * M * I * $\log_{(D)}$, represented a four-way interaction between measure, informant, outcome, and day number. In comparisons not including the youth informants, the informant element (element I) was removed from the model.

2. $Y = O + M + O * M + I + O * I + M * I + O * M * I + \log_{(V)} + O * \log_{(V)} + M * \log_{(V)} + O * M * \log_{(V)} + I * \log_{(V)} + M * I * \log_{(V)} + O * I * \log_{(V)} + O * M * I * \log_{(V)}$. In the second analytic model, V was equal to session or therapeutic contact number (intake counts as session one) in order to also track dose of therapy. A dose of therapy was defined by therapeutic contact at Valley Mental Health including intake, individual therapy sessions, family therapy sessions, medication evaluations, and the like.

The analyses for these two models focused on the interactions that were statistically significant to allow conclusions as to which measure is most sensitive to change and which type of informant is most sensitive to change. These models addressed both of the study's hypotheses regarding the relative sensitivity to change for the three measures. It further examined which type of informant data is most sensitive to change.

In this study the informants provided variable frequencies of data; anywhere from two data points through five data points were collected. However, the analysis proceeded regardless of the number of data points obtained for each measure, since HLM is designed to examine variable frequencies of data.

Level 1 of the analysis was the within-subject stage. HLM first generated individual growth trajectories; these varied by individual participant. This initial analysis took into account covariates that may have affected the trajectory of the slope: initial score (initial severity) and age of client. Initial score has been commonly found to affect treatment trajectory (e.g., Lambert, 2007); age of client was selected as a covariate due to analyses such as those from the Y-OQ-2.01 and the Y-OQ SR-2.0 where researchers (Burlingame et al., 2005; Wells, Burlingame, & Rose, 2003, respectively) found significant differences in their analyses based on the age of the child or adolescent.

These individual growth trajectories then became the dependent variables in level 2 where HLM took the average of the slopes and generated an estimate of the population. Slopes calculated using the estimates from stage 2, were used to draw conclusions regarding the relative sensitivity to change of the CBCL/6-18, BASC-2, and Y-OQ-2.01; these results are presented in Chapter 4.

CHAPTER 4

Results

Sample Demographics

The final analytic sample for comparing the sensitivity to change for the CBCL/6-18, BASC-2, and Y-OQ-2.01 included 136 child or adolescent cases. Of these cases, there were 134 adult informants reporting on 68 (50%) females and 68 (50%) males. Forty-four adolescents (25 females and 19 males) also served as informants by completing self-report measures; 42 of these adolescents filled out measures in addition to the measures filled out by their parent or significant-adult figure, while 2 of these adolescents participated alone. Thus, there were a total of 178 informants in this study for 136 cases.

The average age of the children and adolescents in the study was 10.76 (*SD* = 3.53). The median age was 10.07 and the mode age was 7.32. The average age of males was 10.19 (*SD* = 3.25); the median age was 9.4, and the mode was 8.68. The average age of females was 11.22 (*SD* = 3.73); the median age was 11.5, and the mode was 8.69. An independent sample *t*-test indicated there were no statistically significant differences in the mean ages between males and females ($p = .06$). A *t*-test on the pair-wise test of differences between the HLM change slopes of those adult informants who completed measures for females compared to those completing measures for males indicated there was no statistical difference in the two groups ($p = 0.83$), indicating adults were not reporting change differently for males and females. Therefore, gender was not treated separately within the analysis.

Descriptive Data

This study sought to obtain up to five data points collected from each informant, with adults filling out the CBCL/6-18, BASC-2, and Y-OQ-2.01 at each data collection point and adolescent informants filling out the CBCL/6-18 YSR, BASC-2-SRP, and Y-OQ SR-2.0. The number of data points obtained from this analytic sample is presented in Table 4.1. As can be seen, 79 informants filled out the CBCL/6-18, BASC-2, and Y-OQ-2.01 two times. Thirty eight informants provided completed measures three times, 29 informants provided measures four times, and 32 informants filled out measures five times over the course of their tenure in the study. Thus, there were 548 data points obtained in this study.

Table 4.1

Frequencies and Total Number of Data Points Collection from the 178 Informants for 136 Cases Retained in the Analytic Sample

Number of Data Points Obtained	*Number of Adult Informants*	*Number of Corresponding Adolescent Informants*	*Number of Cases with Only Adolescent Informants*	*Total Informants*	*Total Number of Data Points (Number of Data Points Obtained x Total Informants)*
2	58	19	2	79	158
3	25	13	0	38	114
4	22	7	0	29	116
5	29	3	0	32	160
Total	134	42	2	178	548

Participants were followed as long as they could be reached and measures completed until five data points were completed. The shortest span of time between intake and the last secured data point was 24 days (e.g. Case #315). The longest span of time between intake and the final secured data point was more than a year at 381 days (e.g. Case #168). The average tenure for participants was 122 days, a length of just over 4 months. The median length of time in the study was 112 days; the distribution had a mode of 120 days with five cases.

The average number of days between data collection points was 56.72 with an *SD* of 36.68. Due to the influence of several outliers, the median and mode are considered more accurate representations of the timing involved with data collection: the median is 44 days and the mode is 21 days. The mode is representative of the study's efforts to secure data points beginning at the third week after intake or the last data point in order to obtain data as close to once per month as possible throughout the participant's tenure in the study, as it often required several attempts to secure completed data.

The total number of sessions, or therapeutic contacts, for the 136 cases of the analytic sample was 1,566. These services were provided largely by clinical social workers, social workers, and psychiatrists. Table 4.2 summarizes the frequencies of services provided, while Table 4.3 summarizes provider disciplines and the percentage of services those providers offered to study participants.

Table 4.2

Frequencies of Services Utilized by Study Participants

Service Provided	*Frequency*	*Percentage*
Psychiatric Diagnostic Interview (Assessment/Evaluation)	167	10.7%
Assessment Dictated with Client Present	37	2.4%
Individual Psychotherapy	186	11.9%
Individual Therapeutic Behavioral Services	1	0.1%
Individual Psychotherapy with Medication Management	2	0.1%
Pharmacologic Management (Medication Management)	177	11.3%
Family Psychotherapy with Client Present	785	50.1%
Family Psychotherapy without Client Present	30	1.9%
Family Therapeutic Behavioral Services	2	0.1%
Multiple Family Psychotherapy	46	2.9%
Group Psychotherapy	133	8.5%
Total	1566	100%

Table 4.3

Provider Disciplines with Frequencies of Services Provided by these Disciplines

Provider Discipline	*Frequency of Services*	*Percentage of Services*
Psychiatrist	182	11.6%
Psychologist	90	5.7%
Clinical Social Worker	653	41.7%
Social Worker	467	29.8%
Registered Nurse	4	0.3%
Social Service Worker	2	0.1%
Advanced Practical Nurse	62	4.0%
Licensed Professional Counselor	106	6.8%
Total	1,566	100%

The range of services for participants was from two sessions to 48, with one outlier having 110 therapeutic services (Case #138). Excluding this outlier, the average number of therapeutic contacts, or sessions, was 11 with an *SD* of eight. The median was nine sessions and the mode was four.

Race of the 136 children and adolescents is presented in Table 4.4 with the largest represented populations self-identified as white (n = 111, 81.6%) and of Hispanic origin (n = 14, 10.3%).

Table 4.4

Race of 136 Children and Adolescents in Analytic Sample

Race	*Frequency*	*Percentage*
American Indian-Native Alaska	2	1.5%
Pacific Islander	3	2.2%
Black	2	1.5%
White	111	81.6%
Hispanic origin	14	10.3%
Asian	1	0.7%
Other	3	2.2%
Total	136	100%

One-hundred and eighteen (86.8%) of these children were identified as Severely and Emotionally Disturbed (SED), 14 (10.3%) were not identified as SED, and 4 (2.9%) did not specify status. To illustrate levels of symptom distress for this population, Table 4.5 provides intake score data. Since the Y-OQ-2.01 and Y-OQ SR-2.0 raw scores were transformed into T-scores for these analyses, yet the Y-OQ-2.01 uses raw scores for client tracking, both raw score and T-score variables are included in this table for descriptive purposes.

Table 4.5

Intake Scores from 178 Informants: Adult and Adolescent Measures

Measure	*Informant*	*Mean*	*Median*	*SD*	*Range*
CBCL/6-18	Parent	69.66	71	6.90	49-88
	Youth (YSR)	65.96	65	9.65	42-88
BASC-2	Parent (PRS)	70.92	69	11.81	42-105
	Youth (SRP)	62.64	61	12.66	42-87
Y-OQ-2.01 (T-Scores)	Parent	76.59	75.97	11.23	48-100
	Youth (SR 2.0)	65.96	65.17	10.99	44-89
Y-OQ-2.01 (Raw Scores)	Parent	91.65	90	29.67	15-153
	Youth (SR 2.0)	81.31	79	32.33	17-148

An independent *t*-test between these 136 analytic sample cases and the 119 cases recruited from intake that did not become part of the analytic sample indicated there was no statistical difference between the means of the intake scores for any of the parent- and self-report versions of the CBCL/6-18, BASC-2, and Y-OQ-2.01. Thus, the means of the intake scores of the 108 cases that did not continue in the study, and the 11 cases that did provide two data points but did not show reliable change, were not statistically different from those that continued in the study and showed reliable change at some point in treatment.

Inferred Validity of Data

Validity of the following analyses was inferred from the Test-Taking Survey Revised (TTS-R; McGrath, 2000). A total of 70 adults filled out the TTS-R, 21 of which had corresponding youth that also completed the measure. With a five point Likert scale, higher numbers indicated more thoughtfulness in filling out the measures. The adult mean was 4.41 with an *SD* of 0.36 and the mode was five; the youth mean was 3.61 with an *SD* of 0.74; the mode was four. With these high reports of thoughtfulness and conscientiousness while answering the questions presented in the CBCL/6-18, BASC-2, and Y-OQ-2.01, results of this study were inferred to be valid.

Although the study sought to obtain one TTS-R from each informant, 17 adults and three adolescents completed the TTS-R a second time. A matched-sample, or paired samples, *t*-test of this adult group compared scores from the first administration to the second and found the first administration had a mean of 4.51 with an *SD* of 0.31, and the second administration had a mean of 4.45 and an *SD* of 0.31. However, the paired differences were not significantly different from each other (2-tailed, $p = 0.36$) indicating that parents had continued conscientiousness in attending to the questions as they completed the three measures.

Descriptive Analysis of Change

RCI Change at Any Point in Treatment

The manner by which the three measures evaluated change was first examined using the reliable change index (RCI). A crosstabulation comparing the RCI variables created for the CBCL/6-18, BASC-2, and Y-OQ-2.01 identified the cases that were recognized by each measure as showing reliable change in either direction (improved or

deteriorated). Of the 136 cases classified as showing reliable change at some point within treatment (by parent- and/or self-report versions of a measure), each measure identified cases the other two also identified; each measure identified cases that one of the other measures also identified, but the third measure did not; and each measure identified cases that the other two measures did not identify. Due to missing data, 133 of the study's cases are accounted for within this crosstabulation. All three measures recognized 63 of the cases as meeting RCI inclusion criteria while there were 70 cases recognized for inclusion in the analytic sample that were identified by only one or two of the measures. These 70 cases identified by a single measure or a combination of two measures are presented in Table 4.6.

Table 4.6

Number of Cases Identified as Meeting RCI Change Criteria by a Single Measure or a Combination of Two Measures for Cases in which there was Not Agreement

	CBCL/6-18	*BASC-2*	*Y-OQ-2.01*	*Total*
CBCL/6-18	6	7	14	27
BASC-2	7	6	22	35
Y-OQ-2.01	14	22	15	51

As shown in Table 4.6, the CBCL/6-18 identified a total of 27 cases alone or in combination with another measure; the BASC-2 identified 35 cases, and the Y-OQ-2.01 identified 51 cases as meeting RCI change criteria either alone or in combination with another measure was 51. Thus, it appears the Y-OQ-2.01 is most sensitive to changes

that occur in client scores by identifying the greatest number of cases for inclusion in the analytic sample, while the BASC-2 is more sensitive to identifying changes that occur than the CBCL/6-18.

By presenting cases recognized for RCI inclusion in the analytic sample, Table 4.6 implicitly provides information regarding cases that were not identified as meeting RCI change criteria by a measure or combination of measures. To further describe the performances of the CBCL/6-18, BASC-2, and Y-OQ-2.01, Table 4.7 presents the total number of cases that have been included in the analytic sample (i.e. they were identified as meeting RCI criteria by at least one measure) that were not identified by a particular measure or its combination with another measure as meeting RCI change criteria.

Table 4.7

Number of Total Cases Not Identified by a Measure or Combination of Two Measures as Meeting RCI Change Criteria for Study Inclusion

	CBCL/6-18	*BASC-2*	*Y-OQ-2.01*	*Total*
CBCL/6-18	22	15	6	43
BASC-2	15	14	6	35
Y-OQ-2.01	6	6	7	19

As shown in Table 4.7, the CBCL/6-18 identified a total of 43 cases as showing no reliable change that were identified by either the BASC-2 alone, the Y-OQ-2.01 alone, or both of those measures as meeting RCI change criteria for inclusion in the analytic sample; the BASC-2 identified 35 cases, while the Y-OQ-2.01 identified a total of 19

cases. Thus, the CBCL/6-18 was the least sensitive to change by having the greatest number of cases showing no reliable change, while the Y-OQ-2.01 was the most sensitive to change by having the least number of cases identified as such.

Beyond this, a purer test of whether a participant's score changed was when two of the three measures identified said change. According to this criterion, as presented in Table 4.7, of those cases included in the analytic sample due to their identification by two of the measures for reliable change, the CBCL/6-18 had 22 cases showing no reliable change, while the BASC-2 had 14 cases and the Y-OQ-2.01 had seven cases. Thus, in regards to these frequencies, the Y-OQ-2.01 was most sensitive to identifying change, in that it did not identify the least amount of cases corroborated for reliable change by two measures, while the BASC-2 was more sensitive to identifying changes than the CBCL/6-18.

Pre- to Post-treatment RCI Change

The 136 cases of the analytic sample included all cases from the original sample that met RCI criteria according to the parent- and/or self-report versions of one of the measures at any point within their tenure in the study. An examination of initial intake scores compared to scores from the final data collection point (i.e. pre- to post-treatment) indicated that 81 (59.5%) met RCI criteria according to the CBCL/6-18, 86 (63.2%) met RCI criteria according to the BASC-2, while 95 (69.9%) met pre- to post-treatment RCI criteria according to the Y-OQ-2.01. Table 4.8 indicates the number of cases improved and deteriorated in these RCI totals.

Table 4.8

Cases Identified by the CBCL/6-18, BASC-2, and Y-OQ-2.01 as Meeting RCI Change Criteria Pre- to Post-treatment from 136 Cases in the Analytic Sample

	Improved	*Deteriorated*	*Total*
CBCL/6-18	60	21	81
BASC-2	57	29	86
Y-OQ-2.01	74	21	95

Relative to the manner in which each measure recognized change at any point in treatment for determining inclusion in the analytic sample, the measures performed in like fashion indentifying change from pre- to post-treatment. The Y-OQ-2.01 identified the greatest number of cases as exhibiting reliable change, while the BASC-2 identified the next greatest number and the CBCL/6-18 identified the fewest number of cases.

A crosstabulation comparing the RCI variables from pre- to post-treatment for the CBCL/6-18, BASC-2, and Y-OQ-2.01 showed how the cases that were recognized by each measure as showing reliable change in either direction (improved or deteriorated) were identified by the other measures. When data was combined there were 127 valid cases within this crosstabulation. There was RCI agreement among all three measures' parent- and/or self-report versions for 52 cases: 47 of the cases were identified as meeting RCI pre- to post-treatment criteria and can be classified as reliably improved or reliably deteriorated, while five cases were identified as not meeting RCI pre- to post-treatment criteria. Of the 47 cases showing reliable change, 35 cases were identified by all three measures as reliably improved and 6 cases were identified by all three measures as

reliably deteriorated. Interestingly, there were six cases that were identified by all three measures as showing reliable pre- to post-treatment change for which the measures had disagreement as to the nature of the change (i.e. one measure classified a case as reliably improved while another measure classified the same case as reliably deteriorated).

Thus, 75 of the 127 cases compared in this crosstabulation did not have pre- to post-treatment agreement between the three measures. These 75 cases, as recognized by a single measure or combination of measures, are presented in Table 4.9.

Table 4.9

Number of Cases Identified as Meeting Pre- to Post-treatment RCI Change Criteria by a Measure or Combination of Two Measures for Cases in which there was Not Agreement

	CBCL/6-18	*BASC-2*	*Y-OQ-2.01*	*Total*
CBCL/6-18	9	9	14	32
BASC-2	9	13	15	37
Y-OQ-2.01	14	15	15	31

As shown in Table 4.9, each of the three measures identified cases as reliably changed that the other measures did not identify. The CBCL/6-18 identified nine cases that the BASC-2 and Y-OQ-2.01 did not identify. The BASC-2 identified 13 cases that the CBCL/6-18 and Y-OQ-2.01 did not identify. The Y-OQ-2.01 recognized 15 cases as meeting RCI change criteria that the CBCL/6-18 and the BASC-2 did not recognize.

Table 4.10 presents each measure's individual total number for cases that were identified as meeting pre- to post-treatment RCI criteria that were not identified by it or its combination with another measure.

Table 4.10

Number of Total Cases Not Identified by a Measure or Combination of as Meeting RCI Pre- to Post-treatment Change Criteria

	CBCL/6-18	*BASC-2*	*Y-OQ-2.01*	*Total*
CBCL/6-18	28	15	13	56
BASC-2	15	14	9	38
Y-OQ-2.01	13	9	9	31

In addition to the total number of cases, Table 4.10 highlights the purer test of change by identifying the number of cases not identified for change when the two other measures did so. The CBCL/6-18 did not identify 28 cases that were identified by both of the other measures, while the BASC-2 did not identify 14 cases, and the Y-OQ-2.01 did not identify 9 cases. This suggests that the Y-OQ-2.01 had the greatest level of corroboration with the other measures.

Cut-off Score Analysis

Cut-off scores are discrete cut-off points that represent the point a score must fall below to classify cases from clinical and sub-clinical, or normal, populations. Due to the measurement error inherent in the use of these points, the CBCL/6-18 has provided a *borderline* range (T-scores of 60-63) and the BASC-2 has provided an *at risk* range (T-

scores of 60-69) to alert clinicians to cases that may still need clinical attention though no longer considered within the clinical range. In order to compare the CBCL/6-18 and BASC-2 to the Y-OQ-2.01 that does not provide borderline descriptors, the *borderline* and *at risk* ranges needed to be collapsed into dichotomous categories of normal and clinical populations. There are three ways in which this could be done with various justifications; thus all three options were calculated and are presented herein.

First, the borderline descriptors were collapsed into the normal population, as CBCL/6-18 T-scores of 64 and above are classified as clinical and the BASC-2 classified T-scores of 70 and above as clinical. This, however, leaves all measures with discrepant standards in regards to the cut-off scores: the CBCL/6-18 approximately one-and-a-half *SD* above the mean, the BASC-2 is two *SD* above the mean, and the Y-OQ-2.01 is one *SD* above the mean.

Second, the CBCL/6-18 *borderline* range was collapsed with T-scores of 60 and above identified as clinical. This is justified by Achenbach and Rescorla (2001): "For efficient dichotomous discrimination between deviant and nondeviant scores, the borderline clinical range can be combined with the clinical range by classifying T-scores ≥60 as deviant on the Internalizing, Externalizing, and Total Problems scales" (p. 96). This equalized the CBCL/6-18 with the Y-OQ-2.01 with cut-off scores that resided one *SD* above the mean.

Third, with the shift in the CBCL/6-18 to a cut-off T-score of 60, the BASC-2 remained at a cut-off score of 70 (two *SD* above the mean) though the two measures are both on standardized scales; with intention to equalize all three measures to the same

standard of cut-off scores at one *SD* above the mean, the BASC-2 cut-off score was lowered from a T-score of 70 to 60.

A comparison of the pre- to post-treatment scores of each measure produced four categories of change as reported by a parent- and/or self-report version of a measure:

1. Clients that started in the clinical range and crossed the cut-off score into the normal range (i.e. *recovered*);
2. Clients that started in the normal range and crossed the cut-off score into the clinical range (i.e. *entered clinical*);
3. Clients that started in the clinical range and remained in the clinical range (i.e. *remained clinical*);
4. Clients that started in the normal range and remained in the normal range (i.e. *remained normal*).

In the following tables, categories 1 and 2 are presented together as they include all cases that crossed cut-off scores, while in categories 3 and 4 are presented together as they include all cases that did not cross cut-off scores. A summary of the findings for the first variation in collapsing three descriptor categories into two, including borderline descriptors in the normal range, are presented in Table 4.11 and Table 4.12.

Table 4.11

Number of Cases Crossing Cut-off Scores using Pre- to Post-treatment Data: Borderline Descriptors for CBCL/6-18 and BASC-2 Collapsed into Normal Range

	Recovered	*Entered Clinical*	*Total Cases Crossing Cut-off*
CBCL/6-18	30	4	34
BASC-2	34	13	47
Y-OQ-2.01	21	1	22

A chi-squared (χ^2) analysis of these findings, indicated that there was no significant difference [(2, N = 85) = 3.13, p = .21] between the measures in regards to frequencies of cases that recovered, but significant difference [(2, N =18) = 13, p = 0.01] between the measures for those cases which entered the clinical range.

Table 4.12

Number of Cases <u>Not</u> Crossing Cut-off Scores using Pre- to Post-treatment Data: Borderline Descriptors for CBCL/6-18 and BASC-2 Collapsed into Normal Range

	Remained Clinical	*Remained Normal*	*Total Cases Not Crossing Cut-off*
CBCL/6-18	88	12	100
BASC-2	42	47	89
Y-OQ-2.01	108	5	113

A χ^2 analysis, indicated that there was statistically significant difference between the measures in regards to frequencies of cases they categorized as *remained clinical* or *remained normal* [$(2, N = 238) = 28.87, p < .01$; $(2, N = 64) = 47.47, p < .01$, respectively].

The BASC-2 had the greatest total number of cases crossing the cut-off while the Y-OQ-2.01 had the least number of cases crossing the cut-off score. This first variation, in collapsing borderline descriptors into the two categories of *normal* and *clinical,* produced results that are not consistent with the general findings of this study, and due to differences in criteria for establishing cut-off scores among the three measures, these results should be interpreted with caution.

A crosstabulation of these pre- to post-treatment cut-off scores for the CBCL/6-18, BASC-2, and Y-OQ-2.01 support this assessment. The analysis yielded 133 valid cases of the 136 in the analytic sample. Of these, there was a relatively low level of agreement between the measures. Fifty-five cases (41%) were classified within the same category by all three measures, with nine cases ranked as *recovered*, 41 cases ranked into *remained clinical*, and five cases ranked into *remained normal*. Interestingly, this comparison reveals there was no agreement between the three measures regarding the category *entered clinical*, cases where a child or adolescent had begun in the normal range and deteriorated to the clinical range.

Results for the findings for the second variation, three borderline descriptor categories collapsed into two by changing the CBCL/6-18 cut-off score from 64 to 60 while leaving the BASC-2 *at risk* borderline descriptor collapsed into the normal range, are presented in Table 4.13, and Table 4.14.

Table 4.13

Number of Cases Crossing Cut-off Scores using Pre- to Post-treatment Data: CBCL/6-18 Cut-off Score Adjusted from 64 to 60 with BASC-2 'At Risk' Category Collapsed into Normal Range

	Recovered	*Entered Clinical*	*Total Cases Crossing Cut-off*
CBCL/6-18	20	3	23
BASC-2	34	13	47
Y-OQ-2.01	21	1	22

A χ^2 analysis, indicated that there was no significant difference in how the measures preformed in regards to identifying cases that recovered [$(2, N = 75) = 4.88, p = .08$], but significant difference between the measures for those cases which entered the clinical range [$(2, N = 17) = 14.59, p < .01$].

Table 4.14

Number of Cases <u>Not</u> Crossing Cut-off Scores using Pre- to Post-treatment Data: CBCL/6-18 Cut-off Score Adjusted from 64 to 60 with BASC-2 'At Risk' Category Collapsed into Normal Range

	Remained Clinical	*Remained Normal*	*Total Cases Not Crossing Cut-off*
CBCL/6-18	105	6	111
BASC-2	42	47	89
Y-OQ-2.01	108	5	113

A χ^2 analysis of these findings, indicated that there was significant difference between the measures in regards to frequencies of cases that remained in the clinical range [(2, N = 255) = 32.68, $p < .01$], as well as those cases that remained in the normal range [(2, N = 58) = 59.41, $p < .01$].

This second variation, collapsing three borderline descriptor categories into two, changed the CBCL/6-18 cut-off score from 64 to 60 while leaving the BASC-2 *at risk* borderline descriptor collapsed into the normal range. As can be seen, the BASC-2 has the greatest number of cases out-performs the CBCL/6-18 and the Y-OQ-2.01 as it did in the first variation, however the CBCL/6-18 and Y-OQ-2.01 are now performing in relatively like manner. There is now no statistically significant difference in how the measures are reporting change for those cases that recovered. These results appear to support the change in the CBCL/6-18 cut-off score.

A crosstabulation of these pre- to post-treatment cut-off scores for this second variation yielded 133 valid cases of the 136 in the analytic sample. There was a lower level of agreement between the measures in this comparison than was seen in the first variation. Fifty cases (38%) were classified within the same category by all three measures, with five cases ranked as *recovered*, 41 cases ranked as *remained clinical*, and four cases ranked as *remained normal*. This comparison reveals there was no agreement between the three measures regarding the category *entered clinical*, as was also the case in the first cut-off score variation.

A third variation in the use of cut-off scores for the CBCL/6-18, BASC-2, and Y-OQ-2.01 involved collapsing three borderline descriptor categories into two by changing

the CBCL/6-18 cut-off score from 64 to 60 and the BASC-2 cut-off score from 70 to 60. Results are presented in Tables 4.15 and 4.16.

Table 4.15

Number of Cases Crossing Cut-off Scores using Pre- to Post-treatment Data: CBCL/6-18 Cut-off Score Adjusted from 64 to 60 and BASC-2 Cut-off Score Adjusted from 70 to 60

	Recovered	*Entered Clinical*	*Total Cases Crossing Cut-off*
CBCL/6-18	20	3	23
BASC-2	22	10	32
Y-OQ-2.01	21	1	22

A χ^2 analysis, indicated that there was no significant difference between the measures in regards to frequencies of cases that recovered [$(2, N = 63) = 0.1, p = .95$]. However, the measures were still performing differently [$(2, N = 14) = 9.57, p < .01$] for those cases that entered the clinical range.

Table 4.16

Number of Cases Not Crossing Cut-off Scores using Pre- to Post-treatment Data: CBCL/6-18 Cut-off Score Adjusted from 64 to 60 and BASC-2 Cut-off Score Adjusted from 70 to 60

	Remained Clinical	*Remained Normal*	*Total Cases Not Crossing Cut-off*
CBCL/6-18	105	6	111
BASC-2	94	10	104
Y-OQ-2.01	108	5	113

A χ^2analysis, indicated that there was no significant difference between the measures in regards to frequencies of cases that remained in the clinical range [(2, N = 307) = 1.06, p = .59] or for those that remained in the normal range [(2, N = 21) = 2, p = .37].

Thus, these results indicated the three measures were performing in like manner in regards to three categories: recovered, remained clinical, and remained normal. The measures were only performing differently in one category: entered clinical. These results were most consistent with expected performance as the standards for cut-off score determination were equalized to one *SD* above the mean. Thus, for the purposes of this study, this variation is considered to be the most appropriate when comparing the CBCL/6-18, BASC-2, and Y-OQ-2.01 in regards to their performance classifying cases according to cut-off scores using dichotomous categories. The three measures had similar number of cases that *recovered*. The BASC-2 had the greatest number of cases that *entered clinical*, showing deterioration, with the CBCL/6-18 and Y-OQ-2.01 showing similar results. The BASC-2 had the least number of cases that *remained*

clinical and the greatest number of cases that *remained normal,* with the CBCL/6-18 and Y-OQ-2.01 showing similar results, though the performance of the BASC-2 is not statistically different in this category from the other two measures.

A crosstabulation of these pre- to post-treatment cut-off scores for this third cut-off score variation yielded 133 valid cases of the 136 in the analytic sample. Of these, there was high level of agreement between the measures. Ninety-five cases (71%) were classified within the same category by all three measures, with eight cases identified as *recovered*, one case identified as *entered clinical,* 83 cases identified as *remained clinical*, and three cases identified as *remained normal*. Thus, the higher level of agreement indicates that this variation used to demarcate cut-off scores may be the most appropriate for those seeking to make comparisons between the CBCL/6-18, BASC-2, and Y-OQ-2.01.

Clinically Significant Change via RCI and Cut-off Scores

Although descriptive for this study's purpose of examining the manner in which these three measures evaluate change, cut-off scores are not highly clinically useful alone (Jacobson & Truax, 1991) due to the "measurement error inherent in the use of such cutoff points" (p.16). Cut-off scores work in tandem with the pre- to post-treatment RCI change data in order to evaluate clinically significant change. Each of the three measures performed differently when evaluating clinically significant change via RCI and cut-off scores. Those that met the RCI criteria with scores diminishing are referred to as *improved,* while those whose scores increased are referred to as *deteriorated.* Those participants that had scores fall below a cut-off score are referred to as *recovered,* while

those who had scores that rose to meet or exceed the cut-off are referred to as *entered clinical*. A summary is presented in Table 4.17.

Table 4.17

Number of Cases Classified for Change via RCI Classifications (Improved or Deteriorated) and Cut-off Score Classifications (Recovered or Entered Clinical) using Pre- to Post-treatment Data

	Improved & Recovered	*Deteriorated & Entered Clinical*	*Total Cases Crossing Cut-off*
CBCL/6-18	27	4	31
BASC-2	31	8	39
Y-OQ-2.01	20	1	21

Of the 81 cases that the CBCL/6-18 identified as meeting pre- to post-treatment RCI change criteria, 31 cases crossed the cut-off score of 64. For the 86 cases that met RCI criteria according to the BASC-2, 39 cases crossed the T-score cut-off of 70. For the 95 cases the Y-OQ-2.01 identified as having met the RCI change criteria, 21 also crossed the raw score cut-off of 46 (47 for the Y-OQ SR-2.0), thus meeting classification requirements to be *improved* and *recovered* or *deteriorated* and *entered clinical*.

As with the original cut-off score comparisons, this clinically significant change analysis showed that the BASC-2 identified the greatest number of cases showing change according to the RCI and cut-off score combination, the CBCL/6-18 showed the next greatest number of cases, and the Y-OQ-2.01 identified the least number of cases.

Table 4.18 indicates the total number of cases that did not show clinically significant change as evidenced by the RCI and movement across the cut-off score. This table includes all cases that reliably improved and those that reliably deteriorated yet did not cross the cut-off score threshold. It separates cases according to their identification within the clinical or normal populations: *remained clinical* and *remained normal*, respectively based on their scores from pre- to post-treatment.

Table 4.18

Number of Cases Not Showing Clinically Significant Change via Meeting RCI Criteria but Not Crossing Cut-off Scores using Pre- to Post-treatment Data

	Remained Clinical	*Remained Normal*	*Total Cases Not Crossing Cut-off*
CBCL/6-18	40	10	50
BASC-2	24	22	46
Y-OQ-2.01	70	4	74

According to the CBCL/6-18, 50 cases met RCI change criteria in either direction; thus they remained in the clinical range from pre- to post-treatment according to cut-off criteria (i.e., they were reliably changed but not recovered) or remained in the normal range from pre- to post-treatment (i.e., they were reliably changed but remained sub-clinical throughout treatment). The BASC-2 identified 46 such cases, while the Y-OQ-2.01, identified 74. Thus, the BASC-2 had the least number of cases that did not show clinically significant change and the CBCL/6-18 identified the next least. The Y-OQ-2.01 had the greatest number of cases that did not show clinically significant change.

These results indicate how each measure would classify cases according to clinically significant change criteria as they are designed to operate. However, as identified in the cut-off score variations, comparisons using the cut-off scores should be interpreted with caution due to the differences in criteria among the three measures for establishing the cut-off scores (Achenbach, 1991, Burlingame et al., 2001, Reynolds & Kamphaus, 1992, 2004). Yet, by altering the cut-off scores to illustrate consistent standards for establishing the cut-off score, as was illustrated in the third cut-off score variation, where the CBCL/6-18 cut-off score was lowered from a T-score of 64 to 60 and the BASC-2 cut-off score was lowered from a T-score of 70 to 60 these comparisons are more equitable. Using these criteria in another clinically significant change analysis, the measures evaluated clinically significant change in the following manner, with Table 4.19 and Table 4.20 summarizing the results:

Table 4.19

Number of Cases Showing Clinically Significant Change using Pre- to Post-treatment Data and Adjusted Cut-off Scores for the CBCL/6-18 and BASC-2

	Improved & Recovered	*Deteriorated & Entered Clinical*	*Total Cases Crossing Cut-off*
CBCL/6-18	19	3	21
BASC-2	21	8	29
Y-OQ-2.01	20	1	21

Table 4.19 shows that of the 81 cases that the CBCL/6-18 identified as meeting pre- to post-treatment RCI change criteria, 21 cases crossed the adjusted cut-off score of 60,

meeting classification requirements to be either *improved* and *recovered* or *deteriorated* and *entered clinical*. For the 86 cases that met RCI criteria according to the BASC-2, 29 cases crossed the adjusted T-score cut-off of 60. The Y-OQ data remained the same as in the original analysis; 21 of 95 cases crossed the raw score cut-off of 46 (47 for the Y-OQ SR-2.0). This clinically significant change analysis showed that the BASC-2 identified the greatest number of cases showing change according to the RCI and cut-off score combination, while the CBCL/6-18 and the Y-OQ-2.01 operated in like-fashion.

Table 4.20 indicates the total number of cases that did not show clinically significant change as evidenced by the RCI and movement across the cut-off score. As with Table 4.18, this table includes all cases that reliably improved and those that reliably deteriorated yet did not cross the cut-off score threshold.

Table 4.20

Number of Cases Not Showing Clinically Significant Change via Meeting RCI Criteria but Not Crossing Cut-off Scores using Pre- to Post-treatment Data and Adjusted Cut-off Scores for the CBCL/6-18 and BASC-2

	Remained Clinical	*Remained Normal*	*Total Cases Not Crossing Cut-off*
CBCL/6-18	54	5	59
BASC-2	54	3	57
Y-OQ-2.01	70	4	74

According to the CBCL/6-18, 59 cases met RCI change criteria in either direction and the the BASC-2 identified 57 such cases; thus they remained in the clinical range from pre- to post-treatment according to adjusted cut-off score criteria or remained in the normal range from pre- to post-treatment. The Y-OQ-2.01 data remained as in the original analysis, classifying 74 cases as unchanged. Thus, with adjusted cut-off score criteria, the BASC-2 still had the least number of cases that did not show clinically significant change and the CBCL/6-18 identified the next least. The Y-OQ-2.01 had the greatest number of cases that did not show clinically significant change.

It appears that when evaluating clinically significant change via RCI and cut-off scores, as was found in the previous comparison that included the cut-off scores as well as the original clinically significant change comparison, the BASC-2 is the most sensitive to change, the CBCL/6-18 has the next greatest sensitivity to change, and the Y-OQ-2.01 has the least sensitivity to change. Since there was no statistical difference between the cut-off score findings for the three measures, these results for clinically significant change should be interpreted with caution.

HLM Analyses with Time Variable

Change was lastly examined through HLM inferential statistical methods for three major comparisons: adult comparison, adult and adolescent dyad comparison, and adolescent comparison.

Adult Informant Comparison

The first HLM comparison involved only the parent-report versions of the CBCL/6-18, BASC-2, and Y-OQ-2.01. Therefore the first analytic model was tested omitting the I variable representing informant, which produced the following analytic

model: $Y = O + M + O * M + \log_{(D)} + O * \log_{(D)} + M * \log_{(D)} + O * M * \log_{(D)}$. Initial Y-OQ-2.01 scores and age at intake were covariates in the analysis, and remained as such in all subsequent analyses. Results as measured by the 134 parents or significant-adult figures in the analytic sample are presented in Table 4.21.

Table 4.21

134 Adult Informant Cases: F values and Significance Levels

Effect	*Num DF*	*Den DF*	*F Value*	*Significance Level*
Initial Y-OQ-2.01	1	265	506.33	< .01
Age At Intake	1	265	6.66	.01
Outcome (O)	1	265	6.39	.01
Measure (M)	3	265	63.89	< .01
*Outcome * Measure (O * M)*	2	265	12.02	< .01
Logdays ($log_{(D)}$)	1	132	82.09	< .01
*Outcome * Logdays ($O * log_{(D)}$)*	1	265	49.81	< .01
*Measure * Logdays ($M * log_{(D)}$)*	2	253	14.68	< .01
*Outcome * Measure * Logdays ($O * M * log_{(D)}$)*	2	265	11.82	< .01

The covariates, initial Y-OQ-2.01 and child's age at intake were both significant in this analysis ($p < .01$ and $p = .01$, respectively), which indicated their significance in the analysis; results were different for those with more extreme intake scores and those of differing ages. For this adult informant comparison, every element within the analytic model produced significant results. Element O in the analytic model represented outcome, which was classified into two groups: *improved* and *deteriorated.* These labels

identified cases according to pre- to post-treatment data from the Y-OQ-2.01, as established previously. If the final Y-OQ-2.01 score decreased they were classified as *improved*, while if the score increased they were classified as *deteriorated*. Results indicated a significant difference relative to outcome ($p = .01$), indicating there was a difference between those who improved and deteriorated, averaged across the three measures.

As an explanatory note, these classifications included both those that met pre- to post-treatment RCI criteria and those that showed RCI change at some point during the study but did not meet RCI criteria for pre- to post-treatment. For example, over a 2-month period, Case #230 showed non-RCI changes, yet when filling out the fifth data collection point the mother reported to a research assistant that her daughter was currently in a suicidal crisis. Her Y-OQ-2.01 converted T-scores reflect this (65, 69, 56, 63, and 90) and provide illustration for scores that met pre- to post-treatment RCI criteria. Over a 7-month period, Case #204 showed RCI changes between various data collection points, but the change between the initial Y-OQ-2.01 and the final Y-OQ-2.01 did not meet RCI criteria. His converted T-scores illustrated this (74, 81, 60, 95, and 77) and provide an example of scores that met RCI criteria at some point in the study but not for pre- to post-treatment.

Results were significant between the intercepts of the three different instruments or measures ($p < .01$), as indicated by the M element of the statistical model, though the difference in slopes, rather than intercepts, is the primary focus of this study. The two-way interaction of M * O was also significant, indicating an interaction between method of measurement and the therapeutic outcome of the child or adolescent. This indicates

that relative to outcome, there was a significant difference in how the instruments tracked change. The CBCL/6-18 was significantly different from the Y-OQ-2.01 ($p < .01$). The BASC-2 was also significantly different from the Y-OQ-2.01 ($p = .01$). A *t*-test on the pair-wise test of differences indicated that the CBCL/6-18 and the BASC-2 were not significantly different from each other ($p = .36$). The element $\log_{(D)}$ in the analytic model was also significant ($p < .01$).

The interaction of outcome by time (O * $\log_{(D)}$) was significant ($p < .01$), as well as the interaction of the measures and time (M * $\log_{(D)}$; $p < .01$). Again, however, in the pair-wise test of differences, the CBCL/6-18 and the BASC-2 were not significantly different from each other ($p = .73$). These findings are best described by the significant three-way interaction of outcome (improved or deteriorated), measures (the CBCL/6-18, BASC-2, and Y-OQ-2.01), and time (O * M * $\log_{(D)}$; $p < .01$) in which CBCL/6-18 and the BASC-2 were not significantly different from each other ($p = .55$). Slopes that resulted from this interaction, as measured by the 134 adults in the analytic sample are presented in Table 4.22. The SAS 9.1 estimates from the final analytic models were combined to calculate these slopes. Negative slopes indicate that the measures showed a reduction of negative symptoms and behaviors (i.e. the child or adolescent had improved as reported by the parents); while a positive slope indicates the measures showed an increase of negative symptoms and behaviors, indicating the child or adolescent had deteriorated.

Table 4.22

Slopes for CBCL/6-18, BASC-2, and Y-OQ-2.01 from 134 Adult Informants, Relative to Outcome

	Improved	*Deteriorated*
CBCL/6-18	-1.02	0.35
BASC-2	-1.11	0.57
Y-OQ-2.01	-2.3	1.39

The slopes produced by the CBCL/6-18 and BASC-2 relative to therapeutic outcome (*improved*, n = 99, and *deteriorated*, n=35) averaged across days are significantly different from the slopes produced by the Y-OQ-2.01 relative to outcome averaged across days. The Y-OQ-2.01 slope for *improved* (*m* = -2.23) was 2 and 2.19 times steeper than the corresponding slopes for the CBCL/6-18 and the BASC-2 (*m* = -1.11 and -1.02, respectively). The Y-OQ-2.01 slope for *deteriorated* (*m* = 1.39) was 3.97 and 2.44 times greater than the slopes for the corresponding classifications of the CBCL/6-18 and BASC-2 (*m* = 0.35 and 0.57, respectively). The slope of the BASC-2 was steeper than the slope of the CBCL/6-18 in both *improved* and *deteriorated* classifications, but they were not statistically different.

HLM slopes are represented graphically with the following considerations. The statistical model calculates change via days, yet for graphical necessity days have been grouped within weeks. For representation purposes, the graphs represent change slopes up to 259 days (37 weeks). There were 7 participants that were followed beyond that

period; 381 days (1.04 years) was the longest span of time a participant remained in the study.

Figure 4.1 presents these findings graphically for those cases where the child or adolescent had an improved outcome.

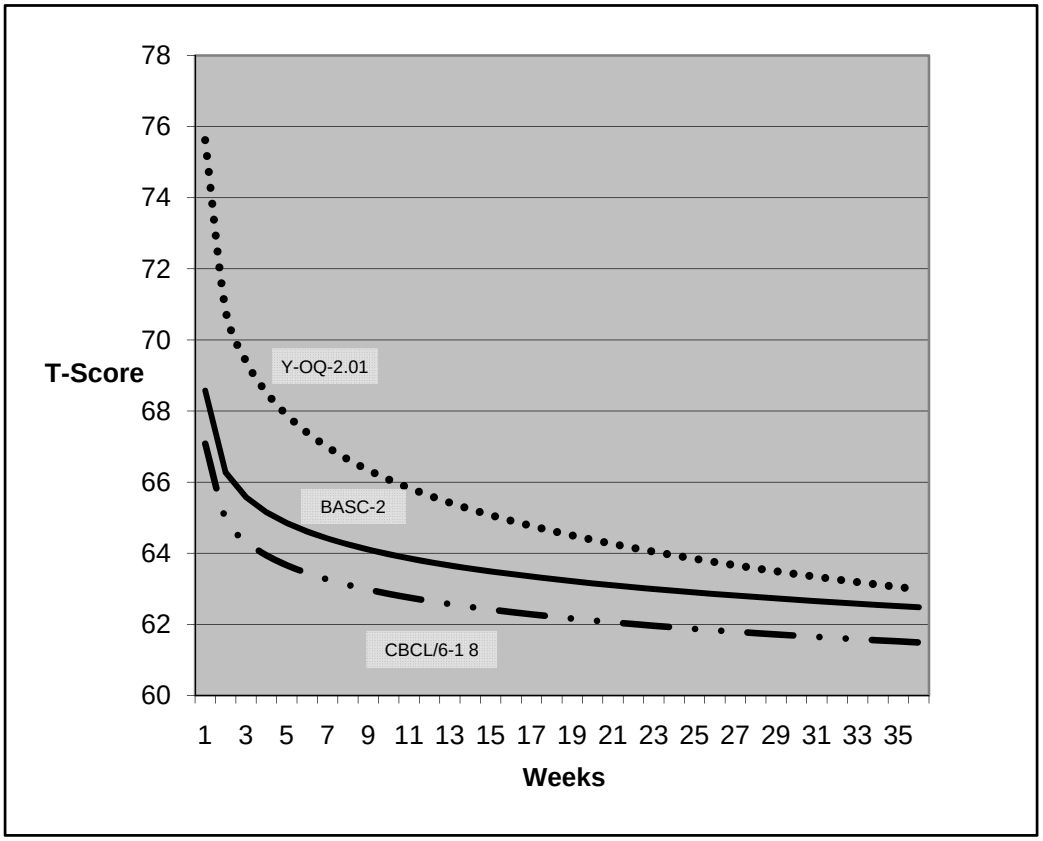

Figure 4.1 Slopes for the CBCL/6-18, BASC-2, and Y-OQ-2.01 calculated from 99 adult informants, relative to children or adolescents with an improved outcome.

As shown, the Y-OQ-2.01 was most sensitive to symptom distress at intake and showed the steepest slope. Although a difference appears in the BASC-2 and CBCL/6-18 slopes, there is no statistical difference and graphically there appears to be little practical difference as they follow an almost parallel course with approximately one T-score

between them. The CBCL/6-18 reported the lowest levels of distress throughout the entire course of change. Figure 4.2 presents these findings graphically for 35 cases where the child or adolescent had a deteriorated outcome.

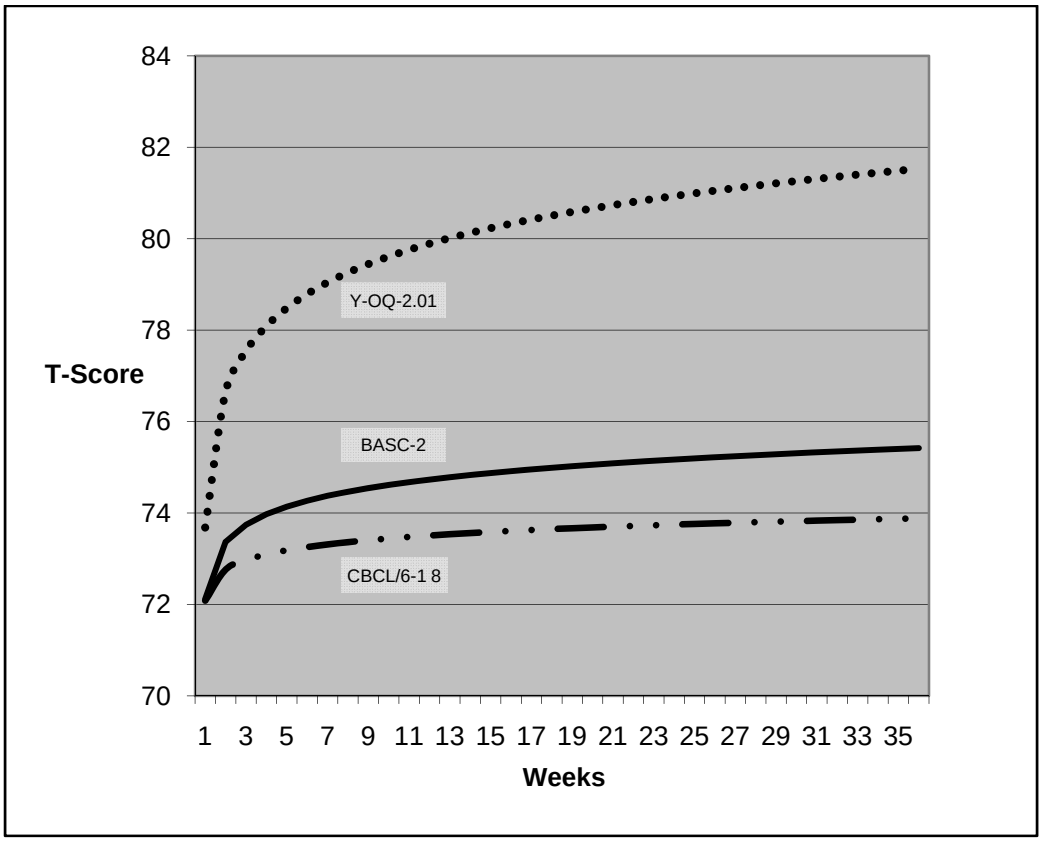

Figure 4.2 Slopes for the CBCL/6-18, BASC-2, and Y-OQ-2.01 calculated from 35 adult informants, relative to children or adolescents with a *deteriorated* outcome.

As presented in Table 4.16 and Figure 4.2, the slope of the CBCL/6-18 for those that deteriorated was 0.35 while the slope of the BASC-2 was 0.57, representing an average shift of approximately two and three T-score points, respectively. The slope of the Y-OQ-2.01 equaled 1.39, representing an average shift of approximately seven T-score points. Although the average deterioration shown by each of these measures did

not meet RCI criteria, the magnitude of the Y-OQ-2.01 slope relative to the slopes of the CBCL/6-18 and BASC-2 indicates that for a client that begins to show signs of deterioration through increasing scores, a clinician would best be able to detect those changes quickly by using the Y-OQ-2.01.

Adult and Adolescent Dyad Comparison

The first analytic model was also examined with the informant variable in place to examine any differences between the parent- and self-report versions of the measures: $Y = O + M + O * M + I + O * I + M * I + O * M * I + \log_{(D)} + O * \log_{(D)} + M * \log_{(D)} + O * M * \log_{(D)} + I * \log_{(D)} + M * I * \log_{(D)} + O * I * \log_{(D)} + O * M * I * \log_{(D)}$. Forty-two of the 44 adolescent informants of the analytic sample had a corresponding significant-adult figure who also served as an informant. Results as measured by these adult and adolescent dyads are presented in Table 4.23.

Table 4.23

42 Adult and Adolescent Informant Dyad Comparisons: F values and Significance Levels

Effect	*Num DF*	*Den DF*	*F Value*	*Significance Levels*
Initial Y-OQ-2.01	1	156	29.99	< .01
Age At Intake	1	156	1.53	.22
Outcome (O)	1	156	3.19	.08
Measure (M)	2	156	41.32	< .01
*Outcome * Measure (O * M)*	2	156	4.61	.01
Informant (I)	1	156	14.33	<.01
*Outcome * Informant (O * I)*	1	154	1.05	.31^{+}
*Measure * Informant (M * I)*	2	156	18.21	< .01
*Outcome * Measure * Informant (O * M * I)*	2	154	2.54	.08^{+}
Logdays ($log_{(D)}$)	1	76	39.81	< .01
*Outcome * Logdays (O * $log_{(D)}$)*	1	156	16.90	< .01
*Measure * Logdays (M * $log_{(D)}$)*	2	147	6.94	< .01
*Outcome * Measure * Logdays (O * M * $log_{(D)}$)*	2	156	3.48	< .05
*Informant * Logdays (I * $log_{(D)}$)*	1	156	3.37	.07
*Measure * Informant * Logdays (M * I * $log_{(D)}$)*	2	154	0.44	.64^{+}
*Outcome * Informant * Logdays (O * I * $log_{(D)}$)*	1	156	4.23	< .05
*Outcome * Measure * Informant * Logdays (O * M * I * $log_{(D)}$)*	2	154	1.96	.15^{+}

($^{+}$) denotes this element was dropped from the final model.

The results of HLM analyses indicate that for the covariates, the intake score from the Y-OQ-2.01 was significant ($p < .01$), but the age at intake was not ($p = .22$), possibly due to the restricted range of ages (e.g., only ages 12 and above) included in this comparison. Outcome alone was not significant (O; $p = .08$); this may be due to the reduced sample size from the sample size that was used in the adult informant analysis. The CBCL/6-18, BASC-2, and Y-OQ-2.01 each performed significantly differently (M; $p < .01$), and did so also when the measures interacted with outcome (O* M; $p = .01$). Informant was significant in this model (I; $p < .01$) indicating there was a difference between the intake scores, or intercepts, of parent- and self-report informants. O * I, the interaction between outcome and informant, was not significant ($p = .31$) and was dropped from the final model. The interaction between informant and measure was significant (M * I; $p < .01$) indicating the parent- and self-report versions of the measures were performing differently. The interaction of outcome, measure, and informant was not significant (O * M * I; $p = 0.08$) and was dropped from the final model.

The time element, $\log_{(D),}$ was significant ($p < .01$), as were the interactions of outcome by time (O* $\log_{(D)}$; $p = < .01$), and measure by time (M * $\log_{(D)}$; $p < .01$). Thus, the interaction of outcome by measure by time (O * M * $\log_{(D)}$) was also significant ($p < .05$) indicating that when averaged across time and outcome the measures performed differently. Each measure will be illustrated in turn respective to this interaction. The interaction of informant by time (I * $\log_{(D)}$) was not significant ($p = .07$), nor was the measure by informant by time interaction (M * I * $\log_{(D)}$; $p = .64$) and due its high p value it was dropped from the final model. The outcome by informant by time interaction was significant ($O * I * log_{(D);}$ $p < .05$). The four-way interaction of outcome

by measure by informant by time was non-significant (O * M * I * $\log_{(D);}$ $p = .15$) and dropped from the final model due to its high value.

Table 4.24 and Table 4.25 present the slopes for adult and adolescent informant dyads, 33 of whom improved and nine of whom deteriorated, respectively.

Table 4.24

Slopes for CBCL/6-18, BASC-2, and Y-OQ-2.01 from 33 Adult and Adolescent Dyads Relative to an Improved Outcome

	Adult	*Adolescent*
CBCL/6-18	-1.25	-0.49
BASC-2	-1.43	-0.66
Y-OQ-2.01	-2.42	-1.65

Table 4.25

Slopes for CBCL/6-18, BASC-2, and Y-OQ-2.01 from 9 Adult and Adolescent Dyads Relative to a Deteriorated Outcome

	Adult	*Adolescent*
CBCL/6-18	0.8	-0.2
BASC-2	0.85	-0.15
Y-OQ-2.01	1.27	0.26

Figure 4.3 presents the CBCL/6-18 Youth Self-Report (YSR) slopes compared to those from their corresponding adult-figure CBCL/6-18 forms, calculated from 33

adolescent informants, relative to an *improved* outcome. The CBCL/6-18 adult informants produced a slope of -1.25, 2.55 times greater than the slope of -0.49 produced by the adolescent informants.

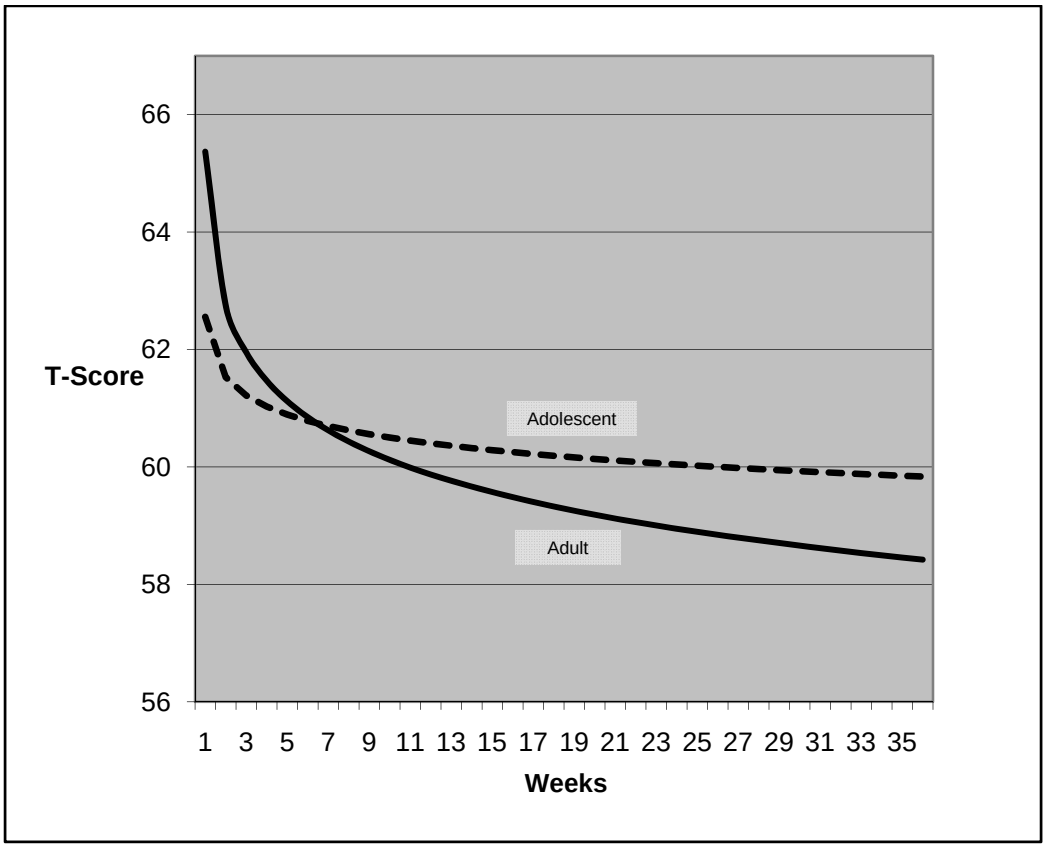

Figure 4.3 The CBCL/6-18 Youth Self-Report (YSR) slopes compared to those from their corresponding adult CBCL/6-18 forms, calculated from 33 adult and adolescent informant dyads, relative to an *improved* outcome.

As can be seen, the CBCL/6-18 adult form is more sensitive to change throughout the treatment process, representing higher levels of distress at intake, as well as lower levels of distress at the 37 week mark. Thus, it would be recommended that clinician's select the parent-report of the CBCL/6-18 over the YSR if data from only one informant can be obtained.

Figure 4.4 presents the CBCL/6-18 Youth Self-Report (YSR) slopes compared to those from their corresponding adult CBCL/6-18 forms, calculated from nine adult and adolescent informant dyads, relative to a *deteriorated* outcome as previously established by the Y-OQ-2.01. As presented in Table 4.25, the CBCL/6-18 adult informants produced a slope of 0.8, 6 times that of the slope produced by the adolescent informants (m = -0.2). The adolescents represented themselves as slightly improved while their parent or significant-adult figure showed deterioration of approximately four T-score points.

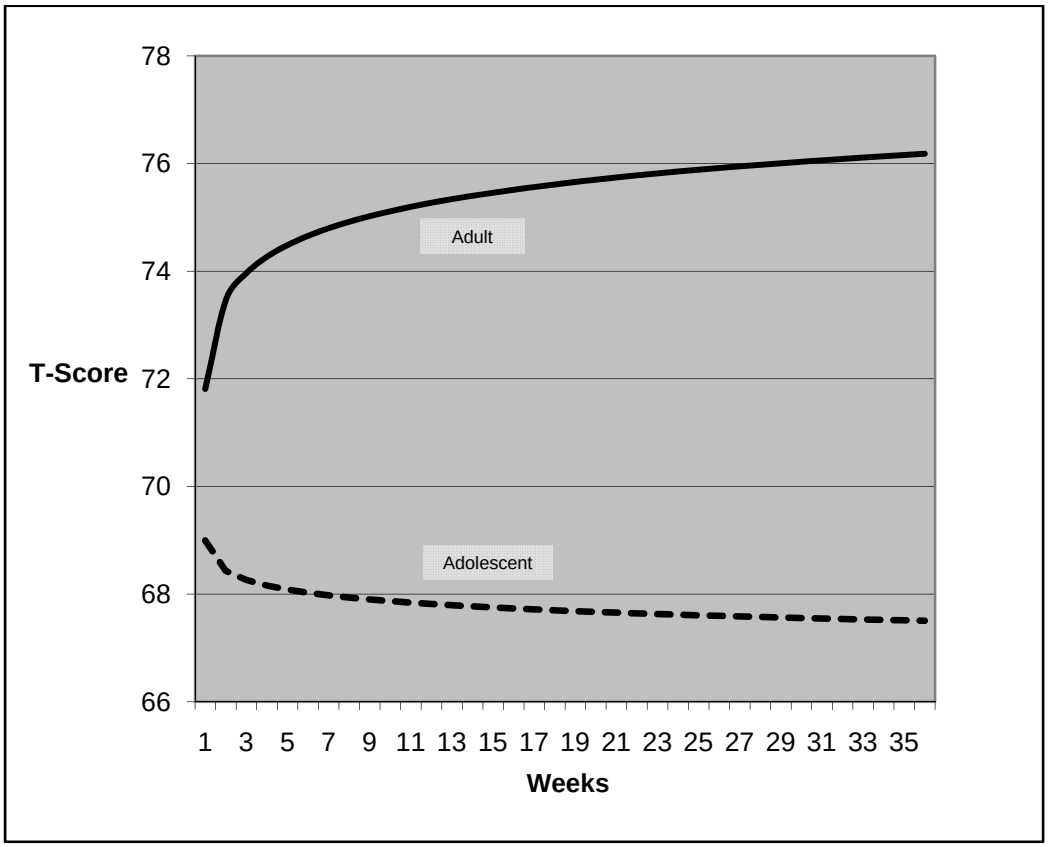

Figure 4.4 The CBCL/6-18 Youth Self-Report (YSR) slopes compared to those from their corresponding adult CBCL/6-18 forms, calculated from nine adult and adolescent informant dyads, relative to a *deteriorated* outcome.

Figure 4.5 presents the BASC Self-Report of Personality (SRP) slopes compared to those from their corresponding adult BASC-PRS-A forms, calculated from 33 adult and adolescent informant dyads, relative to an *improved* outcome. As presented in Table 4.24, the BASC-2 adult informants produced a slope of -1.43, 2.17 times greater than the slope of -0.66 produced by the adolescent informants. As can be seen graphically, adolescents report their distress at a lower level than do their corresponding adult informants.

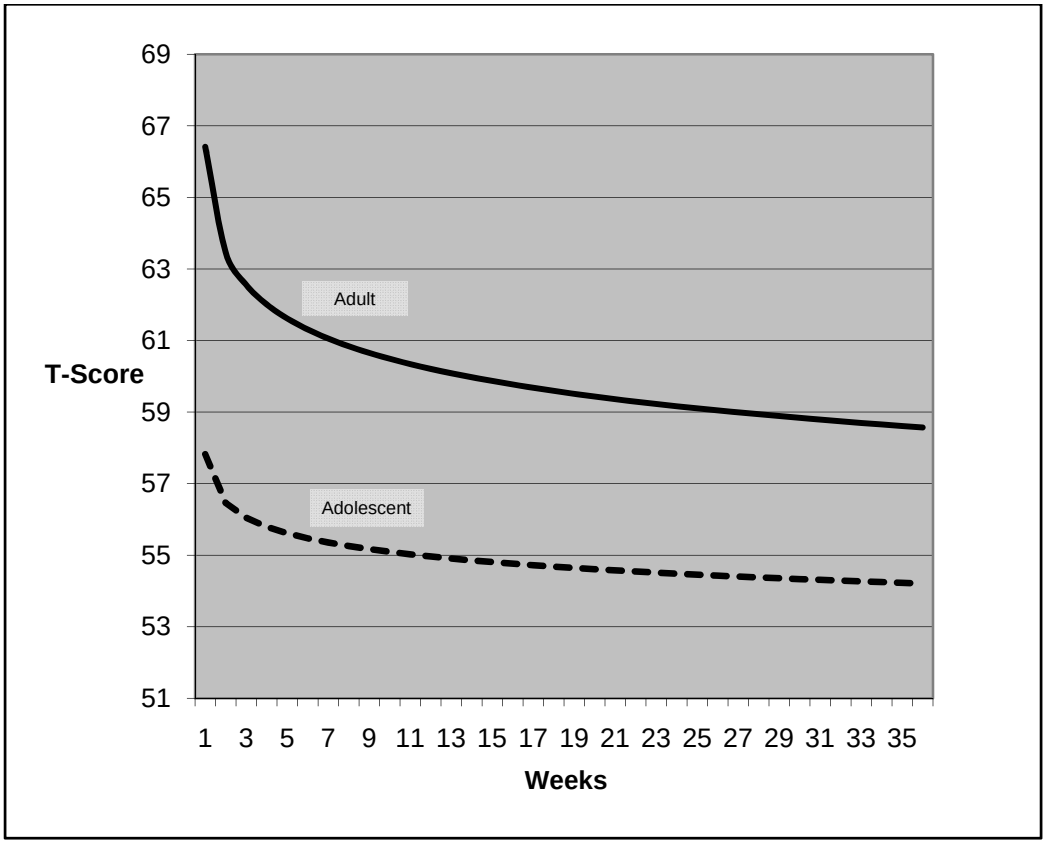

Figure 4.5 The BASC-2 Self-Report of Personality (SRP) slopes compared to those from their corresponding adult-figure BASC-PRS-A forms, calculated from 33 adolescent informants, relative to an *improved* outcome.

Figure 4.6 presents the BASC Self-Report of Personality (SRP) slopes compared to those from their corresponding adult BASC-PRS-A forms, calculated from nine adult and adolescent informant dyads, relative to a *deteriorated* outcome. The BASC-2 adult informants produced a slope of 0.85, 5.66 times greater than the slope of -0.15 produced by the adolescent informants, as presented in Table 4.25. In addition to a large discrepancy in rate of change as illustrated by these differences in slope, as can be seen graphically, adolescents report their distress by week eight an average of 13 T-score points lower than do their corresponding adult informants.

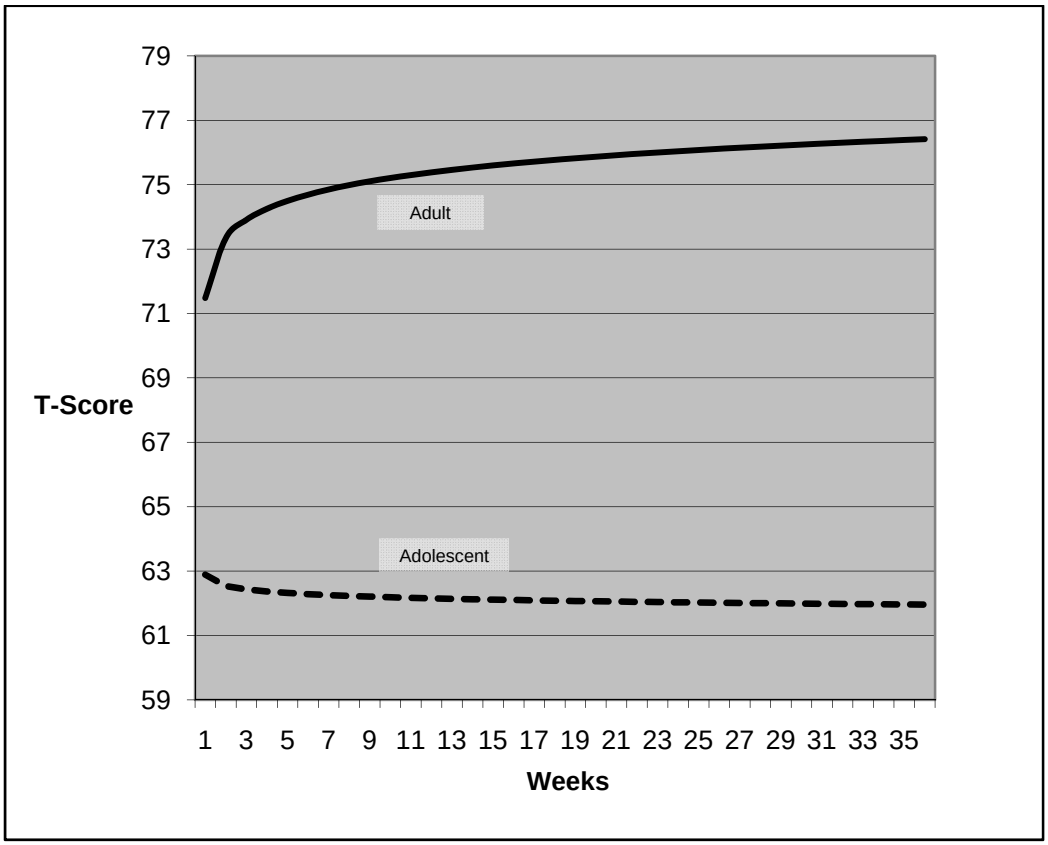

Figure 4.6 The BASC-2 Self-Report of Personality (SRP) slopes compared to those from their corresponding adult-figure BASC-PRS-A forms, calculated from 9 adolescent informants, relative to a *deteriorated* outcome.

Figure 4.7 presents the Y-OQ SR-2.0 slopes compared to those from their corresponding adult- Y-OQ-2.01 forms, calculated from 33 adult and adolescent informant dyads, relative to an *improved* outcome. The Y-OQ-2.01 adult informants produced a slope of -2.42, 1.47 times greater than the slope of -1.65 produced by the adolescent informants. As can be seen, adolescents report their distress at lower levels than do their corresponding adults.

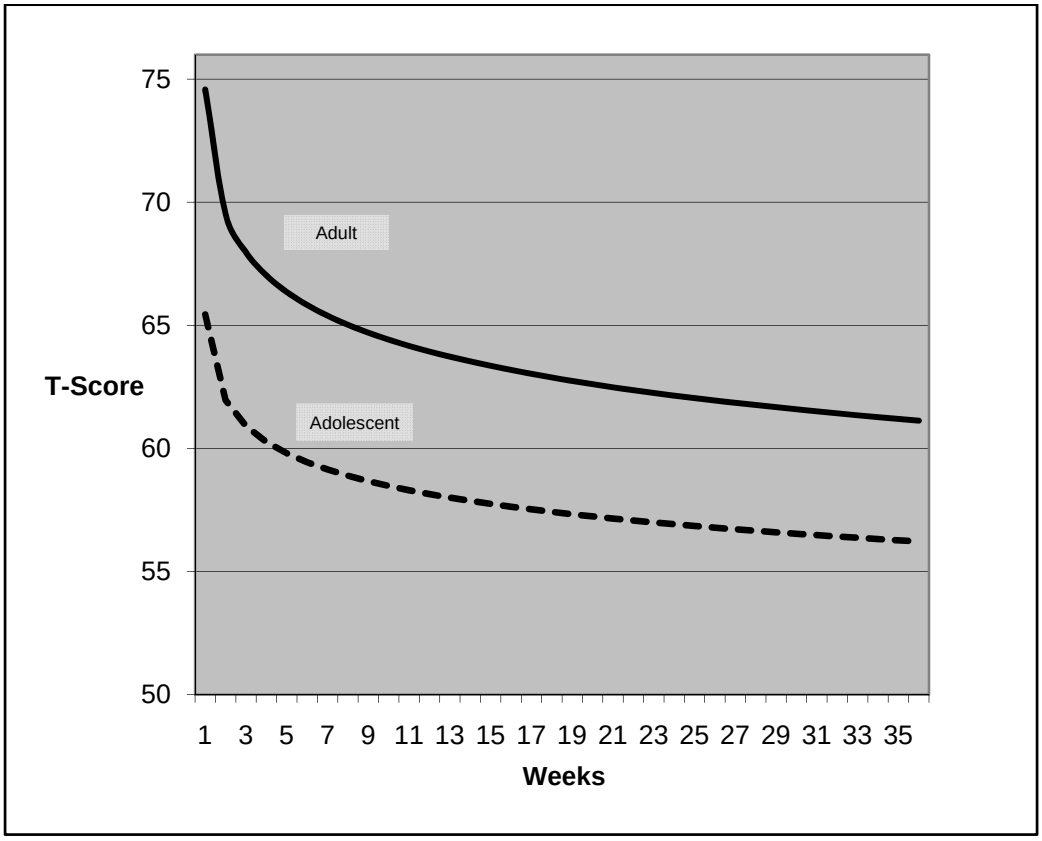

Figure 4.7 The Y-OQ SR-2.0 slopes compared to those from their corresponding adult-figure Y-OQ-2.01 forms, calculated from 33 adolescent informants, relative to an *improved* outcome.

Figure 4.8 presents the Y-OQ SR-2.0 slopes compared to those from their corresponding adult Y-OQ-2.01 forms, calculated from nine adult and adolescent

informant dyads, relative to a *deteriorated* outcome. The Y-OQ-2.01 adult informants produced a slope of 1.27, 4.88 times greater compared than the slope of 0.26 produced by the adolescent informants. This discrepancy among the slopes of youth and parent informants for the Y-OQ-2.01 is substantial, indicating that the Y-OQ SR-2.0 is not detecting client deterioration as is the Y-OQ-2.01.

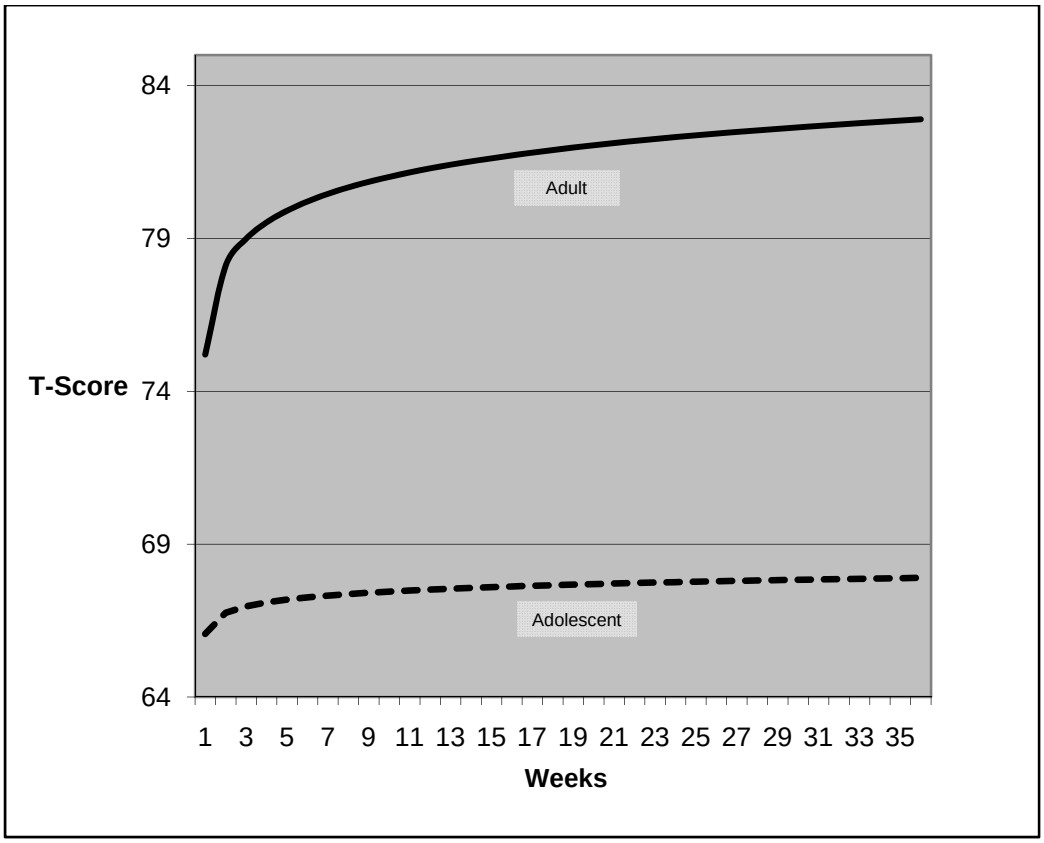

Figure 4.8 The Y-OQ SR-2.0 slopes compared to those from their corresponding adult-figure Y-OQ-2.01 forms, calculated from nine adolescent informants, relative to a *deteriorated* outcome.

Adolescent Informant Comparison

The analytic sample of 136 contained 44 adolescent informants that participated in the study; this sample contained 19 males and 25 females. Thirty-five of these

adolescents were classified by the Y-OQ-2.01 pre- to post-treatment change as *improved*, while nine were classified as *deteriorated*. Results of a comparison between the slopes of the CBCL/6-18 YSR, BASC-2-SRP, and Y-OQ SR-2.0 as measured by these 44 youth are presented in Table 4.26.

Table 4.26

44 Adolescent Informant Cases: F values and Significance Levels

Effect	*Num DF*	*Den DF*	*F Value*	*Significance Level*
Initial Y-OQ-2.01	1	85	1.84	.18
Age At Intake	1	85	3.35	.07
Outcome (O)	1	85	2.79	.09
Measure (M)	3	85	6.11	< .01
*Outcome * Measure (O * M)*	2	85	0.31	.74
Logdays ($log_{(D)}$)	1	43	7.85	< .01
*Outcome * Logdays (O * $log_{(D)}$)*	1	85	2.58	.11
*Measure * Logdays (M * $log_{(D)}$)*	2	83	1.30	.28
*Outcome * Measure * Logdays * (O * M * $log_{(D)}$)*	2	85	0.27	.77

For this adolescent informant comparison, the covariates of initial score and age were not significant ($p = 0.18$; $p = 0.07$, respectively). Element O of the analytic model, based on adolescent outcome classification, did not produce significant results ($p = 0.09$). The M element, representing type of measure, was significant ($p < .01$), indicating the intercepts of the measures were different. The element $log_{(D)}$ in the analytic model was also significant ($p < .01$), indicating adolescents made progress as they continued through

therapy. No interactive elements of the analytic model were significant in this analysis, indicating the therapeutic progress reported by adolescents was not different relative to instrument. Thus, the CBCL/6-18 YSR, BASC-2-PRS, and Y-OQ SR-2.0 were not distinguishable relative to sensitivity to change in the adolescent-only sample.

Although no significance was found for the three-way interaction of outcome, measures, and time (O * M * $\log_{(D)}$), indicating no significant differences between how the measures performed in tracking changes over time for adolescent self-informants, since the results for those that improved are consistent with the findings of the adult informant comparison as well as for the adult and adolescent dyad comparison, the tables and figures for this comparison are included for descriptive illustration. Table 4.27 provides the slopes for the CBCL/6-18 YSR, BASC-2-SRP, and Y-OQ SR-2.0 from 44 adolescent informants, relative to outcome.

Table 4.27

Slopes for CBCL/6-18 YSR, BASC-2-SRP, and Y-OQ SR-2.0 from 44 Adolescent Informants, Relative to Outcome (Based on Non-significant Results)

	Improved	*Deteriorated*
CBCL/6-18 YSR	-0.43	0.2
BASC-2 SRP	-0.82	0.03
Y-OQ SR-2.0	-1.32	0.08

As seen in Table 4.27, although not statistically significant in this analytic comparison, the slopes produced by the Y-OQ SR-2.0 (m = -1.32) relative to an improved therapeutic outcome (*improved*, n = 35) was steeper by 3.06 and 1.61 times than the slopes for the corresponding classifications of the CBCL/6-18 and BASC-2 (m = -0.43 and -0.82, respectively). The slope of the BASC-2 is steeper than the slope of the CBCL/6-18 by 1.91 times for those that improved.

Figure 4.9 presents these findings graphically for those 35 youth that were classified as *improved.* The slopes in this figure are similar to those in the adult informant and adult and adolescent dyad analyses. However, statistical significance was not reached.

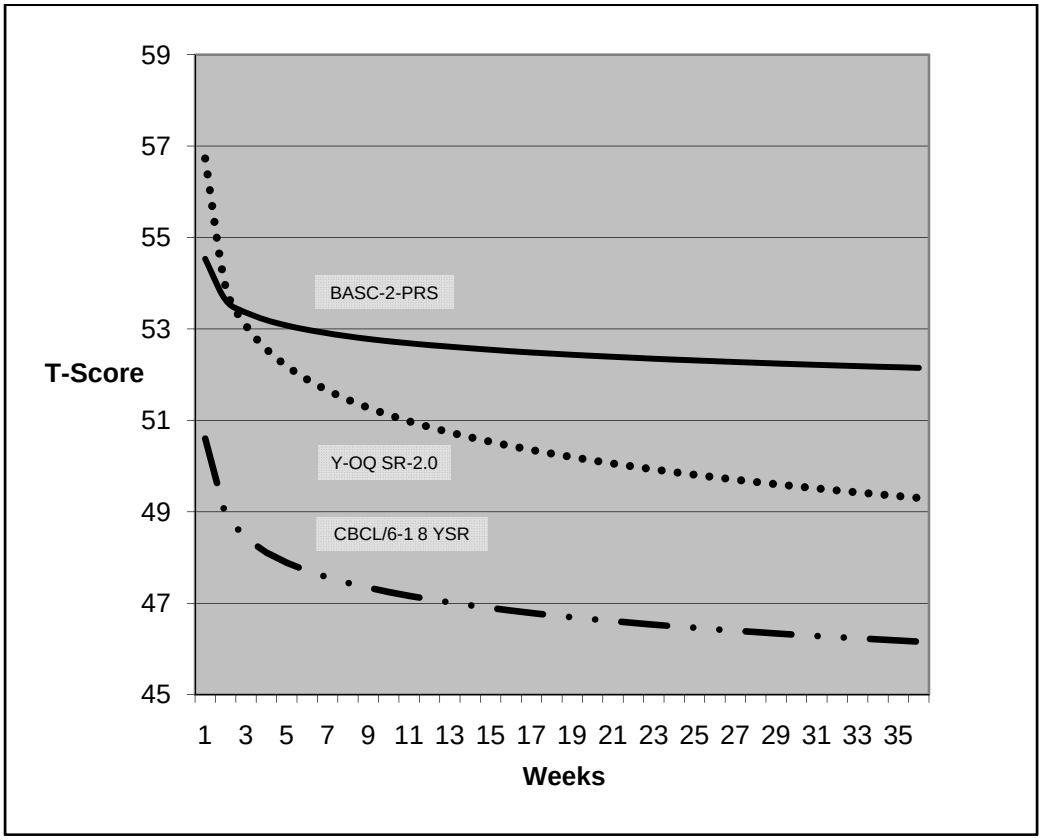

Figure 4.9 Slopes for the CBCL/6-18 YSR, BASC-2-SRP, and Y-OQ SR-2.0 calculated from 35 adolescent informants, relative to those with an *improved* outcome.

Figure 4.10 presents these non-significant findings graphically for those nine youth classified as *deteriorated.* Relative to a deteriorated therapeutic outcome (*deteriorated*, n = 9) the slope of the CBCL/6-18 (m = 0.2) is steeper than the slope of the BASC-2 (m = 0.03) and the slope of the Y-OQ-2.01 (m = 0.08) by 6.66 and 2.5 times, respectively.

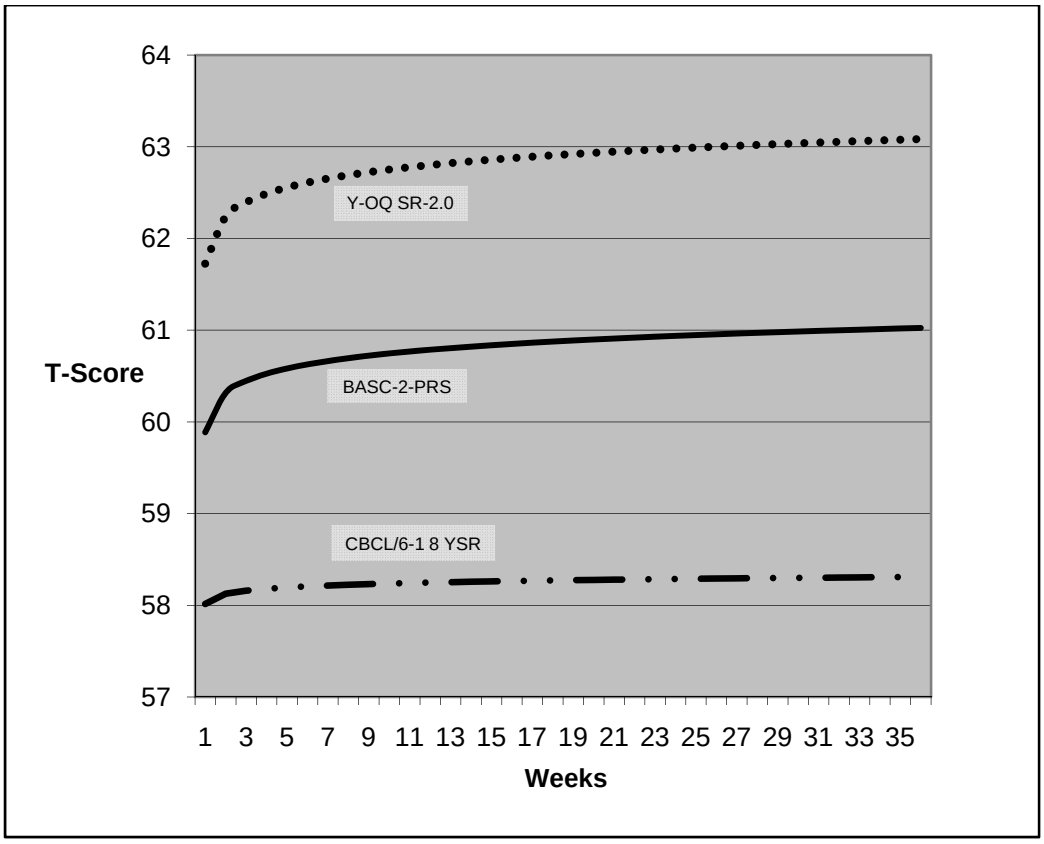

Figure 4.10 Slopes for the CBCL/6-18 YSR, BASC-2-SRP, and Y-OQ SR-2.0 calculated from 9 adolescent informants, relative to those with a *deteriorated* outcome.

HLM Analyses with Dosage Variable

The second analytic model for HLM analyses was as follows: $Y = O + M + O * M + I + M * I + O * M * I + \log_{(V)} + O * \log_{(V)} + M * \log_{(V)} + O * M * \log_{(V)} + I * \log_{(V)} + M * I * \log_{(V)} + O * M * I * \log_{(V)}$ with V equal to session or therapeutic contact number (intake counts as session one) in order to also track dose of therapy. This model is the same as the first analytic model, only with session number substituted for the time variable. One outlier was excluded from theses analyses due to an excessive number of sessions (Case #138 had 110 therapeutic services during his tenure in the study while the range of all other cases was 1-48 sessions). Therefore, the sample for the adult informant

analysis contained 133 cases, the adult and adolescent dyad analysis contained 41 cases, and the adolescent informant analysis contained 41 cases. HLM results regarding sensitivity to change for the CBCL/6-18, BASC-2, and Y-OQ-2.01 using the session variable showed statistical significance in identical model elements and interactions at approximately the same levels to those from the first model using the time variable, indicating no significant statistical differences between the two models. Thus, it appears that session number and time are surrogates for each other in this study. Furthermore, the analyses using days in treatment as the time variable were emphasized given previous literature that found no dose-effect relationship (Salzer, Bickman, & Lambert, 1999) or even a "reverse dose-effect trend" when dose was defined solely by the number of sessions attended (Reardon, Cukrowicz, Reeves, & Joiner, 2002, p. 280).

CHAPTER 5

Discussion

This study examined the relative sensitivity to change of the CBCL/6-18, BASC-2, and Y-OQ-2.01 parent- and self-report versions. Using the CBCL/6-18 and BASC-2, measures that were not designed for tracking change and that are less change sensitive, for tracking change is potentially problematic since their consumers attempt to assess the quality of treatments and calibrate the needs of individual clients based on the data these measures produce. Though these measures are the standards in the fields for which they were created (i.e., assigning diagnoses), the adopted use of tracking treatment and outcome is discouraged based on the results of this study. Significant differences found between these measures and the Y-OQ-2.01 regarding sensitivity to change allow the work of researchers, clinicians, and health care corporations to be maximized by discontinuing the use of the CBCL/6-18 and BASC-2 as tracking instruments in favor of the Y-OQ-2.01.

Change was examined through the use of the reliable change index (RCI), cut-off scores, and hierarchical linear modeling (HLM) slopes. Overall, the results across RCI and HLM analytical methods confirmed the study hypothesis that the parent-report version of the Y-OQ-2.01 was more sensitive to change over time and session number than the BASC-2. Yet, the hypothesis that the CBCL/6-18 was less sensitive to change over time and session number than the BASC-2 or Y-OQ-2.01 parent-report versions was not confirmed in this study. Cut-off score analyses indicated there were no statistically significant differences between the three measures in the manner in which they classified

cases. Also, HLM analyses found no statistical differences between the CBCL/6-18 and the BASC-2.

Though these analyses created a landscape of average change results by examining the study's sample on a nomothetic level, it should be noted that these average changes are of import only in that they represent the idiographic changes the measures captured; these measures document change in severity of, and frequency of, psychopathological symptoms on an individual, per-client basis. The sensitivity with which these measures capture change, and the manner in which the changes are then acknowledged and incorporated into treatment by researchers, clinicians, and health care corporations, may have significant impact on the course of therapy and eventual therapeutic outcome for each child and adolescent seeking psychotherapeutic services.

The hypothesis that this pattern of change sensitivity would also be found for the self-report versions of the measures was not confirmed in this study; although trends were similar to those seen in the adult informant comparison, results were not statistically significant.

This chapter highlights these important findings and their implications for the clinical treatment of children and adolescents seeking psychotherapeutic services. This chapter also briefly discusses limitations concurrently with recommendations for future research.

Major Findings and Implications

The CBCL/6-18, BASC-2, and Y-OQ-2.01 were compared relative to nine characteristics identified by researchers and clinicians to aid consumers in choosing an ideal outcome measure:

1. Excellent reliability (Lambert & Hill, 1994; Vermillion & Pfeiffer, 1993; Weber, 1997);
2. Excellent validity, including validity for change (Lambert & Hill, 1994; Lipsey, 1990; Vermillion & Pfeiffer, 1993; Weber, 1997);
3. Sensitivity to change (Achenbach & Rescorla, 2004; Lambert & Hill, 1994; Vermillion & Pfeiffer, 1993);
4. Normed, with use of score cut-off scores and RCI norms (Durlak et al., 1995; Jacobson et al., 1984; Jacobson & Truax, 1991; Vermillion & Pfeiffer, 1993);
5. Brief; to be completed quickly (Burlingame et al., 1995; Lambert & Hill, 1994; Kazdin, 2005; Lipsey, 1990; Hatfield & Ogles, 2004);
6. Scored and interpreted easily and quickly (Burlingame at al., 1995; Kazdin, 2005; Lambert & Hill, 1994, Hatfield & Ogles, 2004);
7. Relevant client information regarding change provided (e.g., Lambert & Hill, 1994; Vermillion & Pfeiffer, 1993);
8. Can be used frequently to track progress (i.e., more than two data points; Bryk & Raudenbush, 1987; Bryk & Weisberg, 1977; Burlingame et al., 1995; Hatfield & Ogles, 2004; Kazdin, 2005; Rogosa et al., 1982);
9. Cost efficient (Burlingame et al., 1995; Hatfield & Ogles, 2004; Weber 1997).

As shown in Table 2.1, adequate information was available to compare the measures in seven of these nine characteristics from the extant literature. Each had excellent reliability and validity, including validity for change. The Y-OQ-2.01 is the shortest of the measures, with the others taking at least twice the time for clients to complete. Scoring appears to be relatively identical, as all use computer scoring

programs. Each measure also provides relevant information regarding client change, though these domains vary depending on the purposes for which the measures were designed. The Y-OQ-2.01 is designed to be used the most frequently and is the most cost-effective. If each of these seven characteristics were considered equal, the Y-OQ-2.01 appears to be the better choice for use as an outcome measure, regarding its brevity, ability for frequent use, and cost. To complete the comparison between the measures in regards to the recommended characteristics of outcome measures, this study focused on sensitivity to change and the use of cut-off scores and RCI.

HLM Change Slopes for the CBCL/6-18, BASC-2, and Y-OQ-2.01 Illustrating Sensitivity to Change

The CBCL/6-18, BASC-2, and Y-OQ-2.01 performed differently relative to their sensitivity to changes that occurred in children and adolescents receiving outpatient psychotherapeutic services from Valley Mental Health, a community clinic. The most significant findings in this study were produced through the HLM analyses evaluating the parent-report versions of the CBCL/6-18, BASC-2, and Y-OQ-2.01 with a time variable and a session or dosage variable. These analyses produced change slopes that illustrated differential sensitivity to change; the Y-OQ-2.01 was most change sensitive, while the BASC-2 and the CBCL/6-18 were not statistically different from each other.

The Y-OQ-2.01 was "designed cooperatively by clinicians, researchers, and managed care administrators in order to meet the needs of all three" (Burlingame et al., 2001, p. 361). The Y-OQ-2.01 satisfies the recommendations previous researchers have suggested for outcome measures; indeed, it was presented by Kazdin (2005) as an illustration for a child and adolescent measurement option. In addition, the Y-OQ-2.01

showed the greatest sensitivity to change relative to the BASC-2 and CBCL/6-18 using RCI analyses and HLM change slopes. In RCI analyses that included examinations of RCI change at any point in treatment and RCI change pre- to post-treatment, the Y-OQ-2.01 identified the most cases as showing reliable change in both analyses, followed by the BASC-2 and CBCL/6-18 respectively. In the HLM analyses, change slopes for the 134 adult informants using the days in treatment, or time, variable, indicated that the Y-OQ-2.01 slope for cases that *improved,* was more than two times steeper than the slopes for the corresponding classification of the BASC-2 and CBCL/6-18. In the adolescent informant comparison the interactions were non-significant, yet the HLM slopes for 35 cases that improved are consistent with the findings of the adult informant comparison suggesting that the Y-OQ SR-2.0 may be more sensitive to change than the BASC-2-PRS and the CBCL/6-18 YSR. This finding regarding improved cases is substantial and has implications for researchers, clinicians, and managed care administrators by showing the Y-OQ-2.01 has the change sensitivity to answer the questions relevant to their domain. Current issues, such as efforts to establish or work within session limits or evaluate progress and final outcome of treatment, can be facilitated by using the measure with the greatest established change sensitivity.

Perhaps the most important finding, relative to its potential impact on the therapeutic outcome of the client, is the manner in which the measures performed relative to clients that showed an increase in symptoms. The Y-OQ-2.01 slope for cases that *deteriorated* was 3.97 and 2.44 times steeper than the slopes for the CBCL/6-18 and BASC-2, respectively. The intercepts of the three measures for those that deteriorated were similar; yet within an average of the first three to five weeks, the Y-OQ-2.01

identified the steepest amount of deterioration, and continued to show deterioration over time. Thus, using the Y-OQ-2.01 would allow therapists to detect deterioration early which may serve to prevent treatment failure or identify potentially harmful treatments (Lambert, 2007; Lilienfeld, 2007).

Evidence for the manner in which the Y-OQ SR-2.0 detects deterioration is not compelling in this study. In the adolescent informant comparison, the slopes for the nine cases that deteriorated indicated the Y-OQ SR-2.0 slope was the steepest of the three measures, followed by the CBCL/6-18 YSR and the BASC-2-SRP. Due to the relatively flat slopes and non-significant results these results should not be interpreted alone; however, they are presented in support of the general findings.

When the HLM analytic comparisons were replicated using the therapeutic dosage or session variable results were identical in statistical significance levels to those results from the analytic comparisons using the time variable. Thus, the findings of the HLM analyses have supported the first hypothesis of this study. It is recommended that clinicians select the Y-OQ-2.01 over use of the BASC-2 or CBCL/6-18 as an outcome measure or for tracking client changes in child and adolescent symptoms and behaviors.

The adult and adolescent dyad comparison results were done to further examine which informant produced results that were most sensitive to change. Results of this comparison indicated that the informant variable was significant. In every comparison, the adult informants produced a higher intercept than did the adolescent informants. The interaction between informant and measure was also significant, indicating the parent- and self-report versions of the measures were performing differently as they tracked change. HLM slopes for the CBCL/6-18, BASC-2, and Y-OQ-2.01 for both improved

and deteriorated conditions were steeper for parent versions than for self-report versions. Thus, the parent versions of these measures are more sensitive to change throughout the treatment process and are recommended for use over the self-report versions of the measures when data from only one informant can be obtained or when considering policy. The discrepant results in the parent versus youth self-report analyses suggest that the self-report measures (the CBCL/6-18 YSR, BASC-2-SRP, and Y-OQ SR-2.0) did not operate as parallel forms to the parent versions, that the adolescent data contained greater variability, or that the adolescent informant sample size was not large enough to detect significance differences in the slopes.

The CBCL/6-18, BASC-2, and Y-OQ-2.01 as Screening Instruments for Reliable Change

When evaluating sensitivity to change based on RCI criteria for study inclusion, (i.e., reliable change at any point in treatment) the Y-OQ-2.01 identified the greatest number of cases as having met the criteria for reliable change, followed by the BASC-2 and the CBCL/6-18. A purer test of whether reliable change had occurred defined cases for which at least two of the measures indicated RCI criteria was satisfied. At this stringency, the Y-OQ-2.01 identified the least number of cases, followed by the BASC-2, and the CBCL/6-18, that showed no reliable change but were identified as meeting reliable change criteria by the other two measures. Relative to screening cases for meeting RCI criteria pre- to post-treatment, the results were consistent. According to these findings the Y-OQ-2.01 would be the first choice for selecting an outcome measure, followed by the BASC-2 and CBCL/6-18.

Identifying reliable change is important because the RCI establishes a confidence-interval based on error variance that must be exceeded in order to label a client's change

as “reliable” (Burlingame et al., 2005). The Y-OQ-2.01 will give a more accurate depiction of client change due to its higher sensitivity. Thus, consumers will make client-care decisions based on actual change in functioning rather than change that occurred due to error or daily fluctuations in mood/functioning. This finding supports the findings of the HLM analyses. Thus, the Y-OQ-2.01’s sensitivity to change can aid researchers in their attempts to make more refined inquires to advance the current literature, such as in attempts to identify a dose-effect relationship (Reardon et al., 2002).

The CBCL/6-18, BASC-2, and Y-OQ-2.01 as Outcome Measures for Clinically Significant Change

In examining clinically significant change—which includes meeting RCI criteria and crossing the cut-off score into the normal range—the results indicated the BASC-2 recognized the greatest number of cases as *improved* and *recovered*, followed by the CBCL/6-18, and the Y-OQ-2.01. This ordering remained consistent for those that *deteriorated* and *entered clinical* and for cases showing no movement across the cut-off threshold.

Since the criteria used to establish the cut-off scores for the various measures were dissimilar (Achenbach, 1991, Burlingame et al., 2001, Reynolds & Kamphaus, 1992, 2004) cut-off scores were adjusted to one *SD* above the mean. With these adjustments, the same ordering of the measures was found, yet in regards to the cut-off score analysis the performance of the BASC-2 was not statistically different from others, thus this finding should be interpreted with caution. In addition, this finding is not consistent with the general findings of this study regarding the measures’ relative sensitivity to change, as established by the RCI and HLM. It is suggested that the HLM

change slopes and the RCI analyses are most accurate in representing the relative performance of the CBCL/6-18, BASC-2, and Y-OQ-2.01 in tracking changes in clients.

Limitations and Recommendations

VMH is a community-based outpatient child and adolescent clinic that generally serves those with lower incomes, including those who are on the United States health program Medicaid. Within this naturalistic setting, treatment method and treatment length were not altered in any way for the purposes of this study. Data collection involved gathering repeated data measures, up to five times for each participant; 548 data points were obtained from 178 participants. Thus, external validity is strong for the findings and implications. Yet, there are four main limitations that elicit recommendations for future research in this area.

The first limitation of this study is generalizability. This population was not a random sample and was limited to those attending VMH in Salt Lake City, UT. Although findings may likely be generalized to other community clinics in the United State due to the general nature of community clinics and the workings of the federal Medicaid program, caution should be used in generalizing results to those who may be from more diverse cultures or countries. Findings should not be generalized for populations which were not included in this study, such as chronic inpatients, those in acute care facilities, or normal populations.

A second limitation is sample size. Although statistical significance was achieved for the adult informants with the sample size obtained, there were only a small number of cases that were shown to deteriorate. Outcome was taken into account within the statistical model to statistically utilize degrees of freedom for the entire sample, yet the

individual graphs of deterioration were based on relatively small numbers and may lack the sensitivity of the graphs based on larger numbers. In addition, since significance was not achieved with the adolescent informant sample, yet the results followed similar trends found in the statistically significant comparisons, it is possible that the adolescent informant sample size was not large enough to detect significant interactions among these observed differences, there may be greater variability in the data, or the self-report measures do not track changes as do the adult forms.

Third, the omission of a control group is a significant limitation. The original study was conceived with a control group consisting of those clients that had gone through intake, agreed to participate in the study, did not return for therapy after that first session at VMH, and yet followed through by returning additional measures for additional data points completed through the mail. However, due to the number of participants recruited, this was not realized. In this study researchers attempted to contact all participants via mail that did not return to VMH, or that were unable to meet research assistants personally at VMH to fill out measures. There were 32 cases from which researchers were able to obtain completed measures via mail; 28 of these cases were retained in the analytic sample due to RCI criteria and none of these cases met the requirements for the control group (i.e., they all had continued services beyond intake). This limitation does not allow confidence in the conclusions that the change observed in this study was achieved due to psychotherapeutic intervention, as was originally conceived. Instead, sensitivity to change results have been explored in a more general sense, without claim that psychotherapy was the mechanism of the observed changes.

Lastly, a procedural limitation is to be noted. This naturalistic study did not alter VMH intake procedures in any way. Thus, the Y-OQ-2.01 was given prior to meeting with a therapist, while the CBCL/6-18 and BASC-2 were given afterwards when clients who chose to become study participants returned to the intake room to fill out consent/assent forms and complete the other measures. Future researchers should attempt to control for this limitation by assuring all measures are completed without interruption.

REFERENCES

Achenbach, T. M. (1985). *Assessment and taxonomy of child and adolescent psychopathology*. Beverly Hills, CA: Sage.

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 Profile.* Burlington, VT: University of Vermont Department of Psychiatry.

Achenbach, T. M., & Edelbrock, C. (1978). The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological Bulletin, 85,* 1275-1301.

Achenbach, T. M. & Edelbrock, C. S. (1983). Manual for the Child Behaviour Checklist and Revised Child Behaviour Profile. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., Howell, C. T., Quay, H. C., & Conners, C. K. (1991). National survey of competencies and problems among 4- to 16-year-olds: Parents' reports for normative and clinical samples. *Monographs of the Society for Research in Child Development, 56,* 1-120.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles.* Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Achenbach, T. M., & Rescorla, L.A. (2004). The Achenbach System of Empirically Based Assessment (ASEBA) for ages 1.5 to 18 years. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (3rd ed., Vol. 2). Mahwah, NJ: Lawrence Erlbaum Associates.

AGS Publishing (2005). BASC-2: Behavior Assessment System for Children: Second edition. Retrieved August 7, 2008, http://ags.pearsonassessments.com/Group.asp?nGroupInfoID=a30000.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Montery, CA: Brooks/Cole Publishing Company.

Aneshensel, C. S., Estrada, A. L., Hansell, M. J., & Clark, V. A. (1987). Social psychological aspects of reporting behavior: Lifetime depressive episode reports. *Journal of Health and Social Behavior, 28,* 232- 246.

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

Behrens, E. & Satterfield, K. (2006, August). Report of findings from a multi-center study of youth outcomes in private residential treatment. Paper presented at the meeting of the American Psychological Association, New Orleans, LA.

Berrett, K. M. S. (2000). Youth Outcome Questionnaire (Y-OQ): Item sensitivity to change (Doctoral Dissertation, Brigham Young University, 2000). *Dissertation Abstracts International, 60*, 4876.

Bickman, L., Rosof-Williams, J., Salzer, M. S., Summerfelt, W. T., Noser, K., Wilson, S. J., et al. (2000). What information do clinicians value for monitoring adolescent client progress and outcomes? *Professional Psychology: Research and Practice, 31,* 70-74.

Brown, G. S., Burlingame, G. M., Lambert, M. J., Jones, E., & Vaccaro, J. (2001). Pushing the quality envelope: A new outcome management system. *Psychiatric Services, 52,* 925-934.

Brown, J. (1987). A review of meta-analyses conducted on psychotherapy outcome research. *Clinical Psychology Review, 7,* 1-23.

Brown, L. L., & Hammill, D. D. (1983). *Behavior Rating Profile: An ecological approach to behavioral assessment*. Austin, TX: Pro-Ed.

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

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park: Sage Publications.

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

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

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

Burlingame, G. M., Wells, M. G., Cox, J. C., Lambert, M. J., Latkowski, M., & Ferre, R. (2005). *Administration and scoring manual for the Y-OQ (Youth Outcome Questionnaire).* Stevenson, MD: American Professional Credentialing Services.

Burlingame, G. M., Wells, M. G., Hoag, M. J., Hope, C. A., Nebeker, R. S., Konkel, K., et al. (1996). *Administration and scoring manual for the Y-OQ.1.* Stevenson, MD: American Professional Credentialing Services.

Burlingame, G. M., Wells, M. G., Lambert, M. J., & Cox, J. C. (2004). Youth Outcome Questionnaire (Y-OQ). In M. E. Maruish (Ed.). *The use of psychological testing for treatment planning and outcome assessment (*3rd ed., Vol. 2). Mahwah, NJ: Lawrence Erlbaum Associates.

Casey, R. J., & Berman, J. S. (1985). The outcome of psychotherapy with children. *Psychological Bulletin, 98,* 388-400.

Clark, J. P. (2002). *The effects of wilderness therapy on perceived psychosocial stressors, defense styles, dysfunctional personality patterns, clinical syndromes, and maladaptive behaviors of troubled adolescents.* George Fox University, Newberg, Oregon.

Clement, P. W. (1994). Quantitative evaluation of 26 years of private practice. *Professional Psychology: Research and Practice, 25,* 173-176.

Cohen, J. (1998). *Statistical power analysis for the behavior sciences* (2nd ed.). New York: Academic Press.

Conners, C. K. (1973). Rating scales for use in drug studies with children.*Psychopharmacology Bulletin, 9,* 24-42.

Conners, C. K. (1989). *Manual for Conners' Rating Scales*. N. Tonawanda, NY: Multi-Health Systems.

Conners, C. K. (1990). *Conners' Rating Scales Manual.* North Towanda, NY: Multi-Health Systems.

Crawford, S. G., Field, C. J., Fisher, J. E., Kaplan, B. J., & Kolb, B. (2004). Improved mood and behavior during the treatment with a mineral-vitamin supplement: An open-label case series of children. *Journal of Child and Adolescent Psychopharmacology, 14,* 115-122.

De Los Reyes, A., & Kazdin, A. E. (2004). Measuring Informant Discrepancies in Clinical Child Research. *Psychological Assessment, 16*, 330-334.

Deyo, R. A., & Inui, T. S. (1984). Toward clinical applications of health status measures: Sensitivity of scales to clinically important changes. *Health Services Research, 19,* 275-289.

Doyle, A., Ostrander, R., Skare, S., Crosby, R. D., & August, G. J. (1997). Convergent and criterion-related validity of the Behavior Assessment System for Children – Parent Rating Scale. *Journal of Clinical Child Psychology, 26,* 276-284.

Drotar, D., Stein, R. E. K., & Perrin, E. C. (1995). Methodological issues in using the Child Behavior Checklist and its related instruments in clinical child psychology research. *Journal of Clinical Child Psychology, 24,* 184-192.

Durham, C. (1999). Outcome Questionnaire: Repeated administrations, mechanical responding, and social desirability. (Doctoral Dissertation, Brigham Young University, Provo, UT, 1998). *Dissertation Abstracts International, 59*, 6112.

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

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

Evans, S. W., Axelrod, J., & Langberg, J. M. (2004). Efficacy of a school-based treatment program for middle school youth with ADHD: Pilot data. *Behavior Modification, 28,* 528-547.

Fitzpatrick, R., Fletcher, A., Gore, S., Jones, D., Spiegelhalter, D., & Cox, D. (1992). Quality of life measures in health care: I. Applications and issues in assessment. *British Medical Journal, 305,* 1074-1077.

Froyd, J. E., Lambert, M. J., & Froyd, J. D. (1996). A review of practices of psychotherapy outcome measurement. *Journal of Mental Health (UK), 5.*

Gironda, M. A. (2000). A validity study of the Youth-Outcome Questionnaire and the Ohio Scales. *Dissertation Abstracts International: Section B: The Sciences & Engineering, 61*(5-B).

Gladman, M., & Lancaster, S. (2003). A review of the Behaviour Assessment System for Children. *School Psychology International, 24,* 276-291.

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

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

Hathaway, S. R. & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory*. University of Minnesota Press.

Herjanic, B., & Reich, W. (1997). Development of a structured psychiatric interview for children: Agreement between child and parent on individual symptoms. *Journal of Abnormal Child Psychology, 25,* 25-31.

Howell, C. T. (2002). *Statistical methods for psychology.* Pacific Grove, CA: Wadsworth Group.

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

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

Jorm, A. F., Duncan-Jones, P., & Scott, R. (1989). An analysis of the re-test artifact in longitudinal studies of psychiatric symptoms and personality. *Psychological Medicine, 19,* 487-493.

Kamphaus, R. W., Reynolds, C. R., Hatcher, N. M., & Kim, S. (2004). Treatment planning and evaluation with the Behavior Assessment System for Children (BASC). In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (3 rd ed., Vol. 2). Mahwah, NJ: Lawrence Erlbaum Associates.

Kazdin, A. E. (1991). Treatment research: The investigation and evaluation of psychotherapy. In M. Hersen, A. E. Kazdin & A. S. Bellack (Eds.), *The clinical psychology handbook* (2nd ed., pp. 293-312). New York: Pergamon.

Kazdin, A. E. (1993). Psychotherapy for children and adolescent psychotherapy research: Limited sampling of dysfunctions, treatments, and client characteristics: *Journal of Clinical Child Psychology, 24,* 125-140.

Kazdin, A. E. (1994). Psychotherapy for children and adolescents. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 543-594). Oxford: John Wiley & Sons.

Kazdin, A. E. (1995). Child, parent, and family dysfunction as predictors of outcome in cognitive-behavioral treatment of antisocial children. *Behavior Research and Therapy, 33,* 271-281.

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston, MA: Allyn & Bacon.

Kazdin, A. E. (2004). Psychotherapy for children and adolescents. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change.* New York: Wiley.

Kazdin, A. E. (2005). Evidence-based assessment for children and adolescents: Issues in measurement development and clinical application. *Journal of Clinical Child and Adolescent Psychology, 34,* 548-558.

Koss & Shiang. (1994). Assessing psychotherapy outcomes and processes. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 664-700). New York: John Wiley & Sons, Inc.

Lambert, M. J. (2007). Presidential address: What we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research, 17,* 1-14.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69,* 159-172.

Lambert, M. J., Hansen, N. B., Umphress, V., Lunnen, K., Okiishi, J., Burlingame, G. M., et al. (1996). *Administration and scoring manual for the Outcome Questionnaire (Q-45.2).* Wilmington, DL: American Professional Credentialing Services.

Lambert, M. L., & Hill, C. E. (1994). Assessing psychotherapy outcomes and processes. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 72-113). New York: Wiley.

Lambert, M. J. & Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 139-193). New York: Wiley.

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

Larson, J. D. (1998). Aggression management with disruptive adolescents in the residential setting: Integration of a cognitive-behavioral component. *Residential Treatment for Children & Youth, 15,* 1-9.

Lehner-Dua, L. L. (2002). *The effectiveness of Russell A. Barkley's Parent Training Program on parents with school-aged children who have ADHD on their perceived severity of ADHD, stress, and sense of competence.* Hofstra University, Hempstead, NY.

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science, 2,* 52-70.

Linden, W., & Wen, F. K. (1990). Therapy outcome research, health care policy, and the continuing lack of accumulated knowledge. *Professional Psychology: Research and Practice, 21,* 482-488.

Lipsey, M. W. (1983). A scheme for assessing measurement sensitivity in program evaluation and other applied research. *Psychological Bulletin, 94,* 152-165.

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

McGrath, L. D. (2000). Youth Outcome Questionnaire: A multiwave instruments' susceptibility to retest artifacts (Doctoral Dissertation, Brigham Young University, 2000). *Dissertation Abstracts International, 60,* 4896.

Meier, S. T. (1997). Nomothetic item selection rules for tests of psychological interventions. *Psychotherapy Research, 7,* 419-427.

Merenda, P. F. (1996). BASC: Behavior Assessment System for Children. *Measurement & Evaluation in Counseling & Development, 28,* 229-232.

Merydith. (2000). Aggression intervention training: Moral reasoning and moral emotions. *NASP Communique, 28* 6-8.

Millon, T., Millon, C., & Davis, R. (1993). *The Millon Adolescent Clinical Inventory.* Minneapolis, MN: NCS Assessments.

Mosier, J. I. (1998). *The predictive validity of the Youth Outcome Questionnaire: Prognostic assessment.* University of Utah, Salt Lake City.

Mosier, J. I. (2001). *Predicting treatment outcome using psychosocial characteristics: A re-examination of the Youth Outcome Questionnaire Prognostic Assessment.* Brigham Young University, Provo, UT.

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 Associates.

Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., et al. (1996). Integrating assessment data from multiple informants. *American Academy of Child and Adolescent Psychiatry, 35,* 1078-1085.

Ogles, B. M., Lambert, M. J., & Fields, S. A. (2002). *Essentials of outcome assessment.* New York: Wiley.

OQ Measures, LLC, (2005). How to obtain an OQ paper & pencil product license. Retrieved August 7, 2008, http://www.oqmeasures.com/LicenceAggrement2005.pdf.

Ostrander, R., Weinfurt, K. P., Yarnold, P. R., & August, G. J. (1998). Diagnosing Attention Deficit Disorders with the Behaviour Assessment System for Children and the Child Behaviour Checklist: Test and construct validity analyses using optimal discriminant classification trees. *Journal of Consulting and Clinical Psychology, 66,* 660-672.

Packman, J. (2002). *Group activity therapy with learning disabled preadolescents exhibiting behavior problems,* University of North Texas.

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

Psychological Assessment Resources, Inc. (2005). Child Behavior Checklist for Ages 6-18, Teacher's Report Form for Ages 6-18, and Youth Self-Report for Ages 11-18 (CBCL 6-18, TRF 6-18, YSR 11-18). Retrieved August 7, 2008, http://www3.parinc.com/products/product.aspx?Productid=CBCL-S.

Quay, H. C., & Peterson, D. (1983). *Quay-Peterson Revised Behavior Problem Checklist*, University of Miami: Miami, FL.

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

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

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

Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior Assessment System for Children.* Circle Pines, MN: American Guidance Service.

Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavioral Assessment System for Children, Second Edition (BASC-2)* (2nd ed.). Circle Pines, MN: AGS Publishing.

Reynolds, C. R., & Kamphaus, R. W. (2005). *Behavioral Assessment System for Children, Second Edition (BASC-2).* Paper presented at the NASP 2005 Workshop.

Richardson, L. M., & Austad, C. S. (1991). Realities of mental health practice in managed care settings. *Professional Psychology: Research and Practice, 22,* 52-59.

Ridge, N. W., Warren, J. S., Burlingame, G. M., Wells, M. G. (2007). *The reliability, concurrent validity, and factor structure of the Youth Outcome Questionnaire Self-Report.* Unpublished manuscript.

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

Salzer, M. S., Bickman, L, & Lambert, E. W., (1999). Dose-effect relationship in children's psychotherapy services. *Journal of Consulting and Clinical Psychology, 67,* 228-238.

Seligman, L. D., Ollendick, T. H., Langley, A. K., & Baldacci, H. B. (2004). The utility of measures of child and adolescent anxiety: A meta-analytic review of the Revised Children's Manifest Anxiety Scale, the State-Trait Anxiety Inventory for Children, and the Child Behavior Checklist. *Journal of Clinical Child & Adolescent Psychology, 33,* 557-565.

Speer, D. C., & Greenbaum, P. E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an Individual Growth Curve approach. *Journal of Consulting and Clinical Psychology, 63,* 1044-1048.

Tingey, R. C., Lambert, M. L., Burlingame, G. M., & Hansen, N. B. (1996). Assessing clinical significance: Proposed extension to the method. *Psychotherapy Research, 6,* 109-123.

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

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome Questionnaire: Item Sensitivity to Change. *Journal of personality assessment, 74,* 242-261.

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

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

Warren, J. S., Nelson, P. L., & Burlingame, G. M. (2008). *Identifying youth at risk for treatment failure in outpatient community mental health services.* Manuscript submitted for publication.

Waters, E., Stewart-Brown, S., & Fitzpatrick, R. (2003). Agreement between adolescent self-report and parent reports of health and well-being: Results of an epidemiological study. *Child: Care, Health & Development, 29,* 501-509.

Weber, D. O. (1997). A field in its infancy: Measuring outcomes for children and adolescents. In K. M. McCoughlin (Ed.), *The 1998 Behavioral Outcomes & Guidelines Sourcebook* (pp. 201-205). New York: Faulkner & Gray, Inc.

Webster-Stratton, C. (1984). Randomized trial of two parent-training programs for families with conduct-disordered children. *Journal of Consulting and Clinical Psychology, 52,* 666-678.

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

Wells, M. G., Burlingame, G. M., & Rose, P. M. (2003). *Administration and scoring manual for the Y-OQ SR-2.0 (Youth Outcome Questionnaire-Self Report).* Wilmington, DE: American Professional Credentialing Services.

Yeh, M., & Weisz, J. R. (2001). Why are we here at the clinic? Parent-child (dis)agreement on referral problems at outpatient treatment entry. *Journal of Consulting and Clinical Psychology, 69,* 1018-1025.

APPENDIX A: MAILING LETTER

Dear Parent:

We appreciate your participation in our research study at Valley Mental Health. We have not been able to connect with you to fill out the next set of measures. Whether or not you are still receiving services at Valley Mental Health, we invite you to continue to participate in our study.

We have included a self-addressed, stamped envelope, as well as a gift certificate as a "Thank You" for your time and participation.

This is all you have to do:

1) Complete the enclosed forms
2) If your child is over 12, and is participating in the study, please have them complete their set that has the SRP-A (blue) and the tan questionnaire
3) Return all the completed forms in the self-addressed stamped envelope
4) Enjoy the enclosed gift certificate (2 are included if your child is over 12 and also participating)

Thank you in advance for your participation! This study will provide valuable information to aid therapists that treat children and adolescents. If you have any questions, please call Debra Theobald McClendon at [phone number included here] or email at [email address included here].

Sincerely,

Debra Theobald McClendon, MA
CEPICA Research Group
Brigham Young University

APPENDIX B: TEST TAKING SURVEY

TEST TAKING SURVEY

Study ID# _____ Please circle one: Parent/Guardian Youth

Please respond to the following 10 questions by circling the answer that best describes your experience filling out the measures over the course of this study. We are interested in how you really felt about filling out the same measures a number of times. For example, did you get bored, did you not mind doing it, did you feel you observed behavior more carefully, etc. Please use the following scale:

N=Never R=Rarely S=Sometimes F=Frequently AA=Almost Always

	1	2	3	4	5
1. I carefully completed the test each time I took it.	N	R	S	F	AA
2. I got tired of taking the test and just marked the answers.	N	R	S	F	AA
3. I took time to think about my answers.	N	R	S	F	AA
4. I didn't read the questions thoroughly before answering.	N	R	S	F	AA
5. I marked answers just to get done quicker.	N	R	S	F	AA
6. I didn't mind re-taking the test.	N	R	S	F	AA
7. I got better at observing my child's behavior by taking the test more than once.	N	R	S	F	AA
8. I skimmed the questions instead of reading them through.	N	R	S	F	AA
9. I tried to answer each question like I had answered it before.	N	R	S	F	AA
10. Sometimes I got bored and lost interest in finishing it.	N	R	S	F	AA

Debriefing Statement

Thank you for filling out this additional questionnaire and for your participation in this study. This study has sought to learn more about how to best track changes in children and adolescents that occur due to therapy services. This study will compare results of the three measures you filled out in order to learn which measure is the best instrument for use in tracking these changes. These results will ultimately provide more helpful treatment to children and adolescents because the information provided by the best measure will help therapists adjust treatment when necessary and help health care organizations make better decisions regarding access to therapy services. Your time and attention has been greatly appreciated.