MEASURING PROGRESS: A COMPARISON OF THREE OUTCOME MEASURES AS CLASSIFIERS OF TREATMENT RESPONSE AND DIAGNOSTIC REMISSION THROUGHOUT CBT FOR YOUTH ANXIETY

By

CHRISTOPHER M. WYSZYNSKI

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Psychology

Written under the direction of

Brian C. Chu

And approved by

____

____

____

____

New Brunswick, New Jersey

October 2018

ABSTRACT OF THE DISSERTATION

MEASURING PROGRESS: A COMPARISON OF THREE OUTCOME MEASURES AS CLASSIFIERS OF TREATMENT RESPONSE AND DIAGNOSTIC REMISSION THROUGHOUT CBT FOR YOUTH ANXIETY

By CHRISTOPHER M. WYSZYNSKI

Dissertation Director:
Brian C. Chu

The successful proliferation of evidence-based measures has left researchers and practitioners the challenge of sifting through a “dizzying array of measures” (Kazdin, 2005, p. 549) in order to select which measures might be appropriate for tracking client progress. The current study used a receiver operating characteristics (ROC) approach in order to investigate how well three common measures of differing breadth were able to classify the outcomes of treatment response, and diagnostic remission, following a 16-week treatment for youth with anxiety (*n*=165). The three measures included in the study were: a narrow, idiographic measure (Target Problems), a general measure of youth distress (State-Trait Anxiety Inventory for Children-Trait- Child/Parent versions; STAIC-T-C/P), and a broad measure of youth internalizing symptoms (Child Behavior Checklist; CBCL). The measures were first assessed on their ability to classify treatment outcomes over the full course of treatment. Measures that were found to be significant classifiers of each outcome were then compared across three treatment intervals (i.e., intake to week 4, intake to week 8, and intake to week 12) in order to determine how measures performed across time and the earliest point at which each measure was able to classify treatment outcomes. The results of the study found that youth-reported measures were able to classify (1) youth-reported response (2) parent-reported response and (3) diagnostic remission above chance levels. Parent-reported measures classified (1) youth- and (2) parent-reported response above chance levels, but not (3) diagnostic remission. Follow-up ROC analyses revealed that youth-reported STAIC and TP measures performed similarly across treatment when predicting remission but youth-reported STAIC scores outperformed all other measures for the outcome of youth-reported response. Similarly, the outcome of parent-reported response was best classified by parent-reported STAIC scores and the CBCL internalizing scale. The study discusses the advantages and disadvantages of each measure as a classifier of treatment response and diagnostic remission.

.

ACKNOWLEDGEMENT

A sincere “thank you” to Dr. Brian Chu for his mentorship and comfort in letting me pursue projects that required me to develop new skills throughout my doctoral degree (and for having a sports or Star Wars analogy at the ready for nearly everything).

I would also like to express my appreciation for Dr. Oliver Lindhiem, who went above-and-beyond the call of duty for an outside committee member. I hope to extend the same generosity and support to others as you have given me from the very first time I contacted you to be on my dissertation committee.

To Dr. Terry Wilson and Dr. Teresa Leyro, thank you for your willingness to sit on my dissertation committee, for your active participation in the project, and for providing such thoughtful recommendations during the proposal.

To all of the people who read the dissertation or provided much needed support during the drafting stages: Lee Markison, Dr. Danielle McCarthy, Jeremy Engelberg, Nick Kang, Peter Kim, David Bopp, Julie Ann Glaz-Bopp, Christopher Baldi, Carol Baldi, Christopher Aa, David Rosenthal, Jason Becker, Jessica Eckert, Alexander Massie, Vincenzo Averello, Tony Rivera, my parents, and my grandparents: you kept both me and the manuscript together.

Finally, I would like to thank Anne Sokolowski, Donna Tomaselli, and Dianne McMillion, who served as the much needed administrative guides during my training.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

A convergence of recent scientific and political movements has increased the importance of tracking clinical outcomes in psychological services. Research incorporating measurement feedback systems (MFS; e.g., Bickman, 2008) has demonstrated that the act of measuring client progress and providing feedback to clinicians can contribute to positive clinical outcomes outside any therapeutic activities (Bickman et al., 2011). Policy acts, such as the Affordable Care Act, have mandated that publically-funded payment structures (e.g., Medicare) shift from positive payment adjustments (i.e., incentives) for practitioners who document client progress during treatment to a negative payment adjustment (i.e., punishment) for those who do not satisfactorily report data on client progress (Centers for Medicare & Medicaid Services, 2016a, 2016b). An increasing number of states have included language about the use of progress monitoring in their legislation about the use of evidence-based practice (Bruns & Hoagwood, 2008). The majority of research to date involving MFSs has focused on the impact of feedback systems on improving clinical outcomes in a variety of populations across a variety of settings (Anker, Duncan, & Sparks, 2009; Bickman, Kelley, Breda, de Andrade, & Riemer, 2011; Carlier et al., 2012; Lambert, Hansen, & Finch, 2001). Less focus has been paid to the use of systematic outcome measurement as a system to identify clients who are likely to respond or not respond to psychological treatment. The current study reports on efforts to establish the psychometric properties of three types of measures (idiographic target problems, narrow-band symptoms, broad-band symptoms) to predict clinical response and remission amongst youth who received standardized cognitive-behavioral therapy (CBT) for anxiety.

The heightened emphasis on using validated measures (sometimes called "evidence-based assessment;" EBA) to track client progress has helped develop a number of high quality decision tools that can help clinicians manage client progress and adapt treatments, when appropriate, if a client is not responding to treatment (e.g., Chorpita, Daleiden, & Collins, 2014). As with any advancement, new challenges have also arisen. One such problem is how to select which measure/s to include as part of a MFS. One review identified over 90 assessment tools for child mental health (S. Williams, 2008). A second article, which focused on youth anxiety, reviewed 14 measures and represented only a portion of the measures available (Southam-Gerow & Chorpita, 2007). The successful proliferation of evidence-based measures provides a unique challenge to clinicians and policy makers about how to navigate and compare the "dizzying array of measures" available (Kazdin, 2005, p. 549).

Previous research has addressed the issue of measure validation and comparison by analyzing the psychometric properties of measures. The early research on assessment recommended that one way to establish a new measure was to demonstrate that the measure was correlated, but not too highly correlated to be redundant, with existing measures that assess a similar construct (i.e., demonstrate discriminant and convergent validity Campbell, 1960; Campbell & Fiske, 1959). As the number of measures increased, additional psychometric properties were applied in order to improve our understanding of how different measures performed as assessments of specific psychological disorders. For example, information about face validity (does the measure appear to be measuring what it claims to measure?), convergent validity (is the measure related to other measures that measure the same construct?), discriminant validity (is the measure unrelated to measures that assess a different construct?), construct validity (does an assessment actually measure the construct it claims to measure?), and reliability (does the measure seem to provide consistent information over time? Do the items on the measure seem to be related?) has been used to provide empirical support for the utility of a diagnostic or screening tool. Yet, in addition to their usefulness, psychometric properties can be challenging to understand (Kazdin, 2005). One expert noted that the "psychometric evaluation of a measure can be endless because in principle no finite number of studies can exhaust one type of validity" (Kazdin, p. 550). Additional problems arise when the psychometric properties that are important for one measure are unimportant for the other or when the importance of the psychometric properties of a measure is inconsistent across uses. The challenge then becomes not only understanding the psychometric properties but also understanding which property is most important for which use. In order to reduce the burden of making such decisions, alternative solutions for selecting one measure might help with selecting appropriate measures.

One potential solution is to employ multiple measures. The call to use multi-source, multi-domain, and multi-format assessments is the current gold-standard approach recommended by many in the scientific community (Kazdin, 2005). However, this comprehensive approach comes with multiple costs, including: client (client time, effort), resource (e.g., expense, scoring tools), and expertise (proper selection of measures and interpretation). When using multiple measures, clinicians are tasked with the corresponding difficulty of sifting through the cavernous measurement literature for multiple measures. Reducing the burden on scientists, practitioners, and scholars is a recent goal of psychological science (Kazdin & Blase, 2011); as such investigating the utility of a measure in a less overall cost-intensive way, in addition to psychometrics, might be an important next step in evidence-based assessment.

A first step in evaluating the utility of a measure is to determine its use. As stated previously, specific reliability and validity statistics have different importance across different measure uses. Assessment tools can be employed to numerous valuable ends; however, treatment outcome is one of the most important. Treatment outcome serves as the basis for evidence-based practice (Chambless & Hollon, 1998), MFSs tracking treatment outcome have been shown to boost treatment efficacy (e.g., Bickman, 2008; Bickman et al., 2011; Lambert et al., 2001), and treatment outcome measures can be used to address some of the economic concerns (e.g., cost-effectiveness) that are of interest to policy-makers and health care administrators (Andrews, 1995). Accordingly, the use of a second methodology to compare outcome measures, designed to assess treatment outcomes following CBT for youth anxiety (i.e., treatment response and diagnostic remission), will be the focus of the current study.

Anxiety disorders were selected as the diagnostic category of interest for the current study because they represent the most common condition among youth, and are frequently accompanied by functional impairments. One population-based study estimated that the lifetime prevalence rate for anxiety disorders among youth (ages 13-18) was 32% and that about 8% of youth suffer from severe anxiety (Merikangas et al., 2010). Researchers have indicated that youth with anxiety disorders have functional impairments at school, with peers, and at home (Alfano, 2012; Ginsburg, La Greca, & Silverman, 1998; McCauley, Katon, Russo, Richardson, & Lozano, 2007). A plethora of free and paid-for-use measures of anxiety are also readily available to be included in

EBA studies (e.g., Achenbach & Rescorla, 2001; Beidas et al., 2015; Spielberger, 1983). The prevalence of anxiety disorders in youth and the availability of multiple, well-studied measures make anxiety a well-suited psychological disorder for more detailed investigations of EBA that can then serve as a model for the study of other diagnoses.

A second step in evaluating measure utility is to identify the particular outcome domains of greatest interest. Some measures focus on idiographic treatment goals, others on discrete diagnostic change, and others on broad symptoms change. The proposed paper will examine one measure from each category, including a Target Problems measure (TP; idiographic measure), the State-Trait Anxiety Inventory for Children-Trait (STAIC-T; specific measure of distress) and the Child Behavior Checklist Internalizing subscale (CBCL-I; broad measure of child internalizing symptoms). Each measure has published articles about their psychometric properties and each measure presents different strengths and weaknesses.

Formal research assessing individual treatment goals started about 50 years ago with the development of the Goal Attainment Scale (GAS; Kiresuk & Sherman, 1968). Studies using the GAS as an outcome measure have often found that the effect sizes for the idiographic measure are larger than effect sizes for more standardized symptom checklists (e.g., Berger, Hohl, & Caspar, 2009; Shefler, Dasberg, & Ben-Shakhar, 1995). The findings suggest symptom checklists might underestimate the effectiveness of psychotherapy, and researchers are utilizing idiographic measures at increasing rates (Kolko, Lindhiem, Hart, & Bukstein, 2014; Weisz et al., 2011). A recent meta-analysis also supported the claim that idiographic measures produce larger effect sizes than symptom checklists. Indeed, the meta-analysis found that the effect size for the 12 studies using idiographic measures was .86 compared to.32 for symptom checklists (Lindhiem, Bennett, Orimoto, & Kolko, 2016). The authors also suggested that idiographic measures have been shown to have face validity, temporal stability, and construct validity (Lindhiem et al., 2016).

Psychometric properties of a youth-based idiographic measure were presented in detail in an article describing the development of the Top Problems measure (TP; Weisz et al., 2011). The authors demonstrated that the TP measure had strong test-retest reliability (*r*'s ranged from .69 to .91, all p<.01), demonstrated convergent validity by strong correlations between TPs provided by children and parents with relevant scales from the Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001) and Youth Self Report (YSR), demonstrated discriminant validity by finding no relationship between TP and theoretically unrelated subscales on the CBCL and YSR (e.g., an anxiety-related TP was not related to the CBCL externalizing subscale), and provided reliable estimates of change over time. Idiographic measures also were found to provide information above and beyond symptom checklists (Weisz et al., 2011) , incorporated the client's voice, and directly measured change in the goals that originally motivated the client to seek services (Kazdin, 2005). Importantly, idiographic measures would likely compare well on cost-effectiveness metrics because they do not require copyrighted scales, can be created with limited tools, and require limited expertise to deliver and interpret (individual target goals are usually rated on a simple, single Likert-type rating scale).

Idiographic measures are not without their limitations. Since the target problems are generated by the client and/or caregiver, it is possible that the goals are not defined in measurable, behavioral terms (e.g., "Not be anxious"). Researchers have suggested that treatment goals might best be created during an intake assessment with the guidance of a clinician (Kolko et al., 2014; Weisz et al., 2011). While defining treatment goals with the help of a clinician might help practitioners circumvent the problem of poorly-defined treatment goals, the additional time needed to define the goals collaboratively may increase the length of intake assessments. However, any time burden in creating an idiographic measure is front-loaded and only occurs once. Once individual goals are identified, repeated assessment only requires the client to rate the improvement or deterioration of each goal using a single item per goal. However, researchers and clinicians must be mindful of the up-front efforts required to generate meaningful goals.

Symptom checklists are a more traditional measure of client progress and have demonstrated strong relationships with treatment outcomes in a number of studies. The State-Trait Anxiety Inventory for Children (STAIC; Spielberger, 1983) is a frequently used parent- and youth-report of child distress. Previous research has demonstrated that the STAIC-Trait (STAIC-T) has high concurrent validity with other measures of anxiety (e.g., Children's Manifest Anxiety Scale) and has strong internal consistency and reliability (Kirisci, Clark, & Moss, 1997). Support for the psychometrics of the STAIC-Trait-Parent (STAIC-T-P; Strauss, 1987) has been more mixed (Southam-Gerow, Flannery-Schroeder, & Kendall, 2003). Research has suggested that the STAIC-T-P has high internal consistency and moderate test-retest reliability. The STAIC-T-P performed slightly worse in evaluations of concurrent and discriminant validity. Researchers found that scores on the STAIC-T-P were correlated with items on the CBCL externalizing broadband scale, which was thought to be a subscale unrelated to the STAIC-T-P (Southam-Gerow et al., 2003). The results from the psychometric examinations of the

STAIC-T-P/C suggested that the measure had the greatest predictive validity and reliability when the scores were compared to other measures within the same reporter (Southam-Gerow et al., 2003). While the STAIC-T-P/C have been used extensively throughout the child anxiety literature, the assessment might not be sufficiently specific to tap into the constructs that parents and children believe are the main reasons for seeking treatment. Further, the STAIC-T-P/C may have a higher financial and time cost barrier to entry than the TP measure. Researchers and clinicians need to purchase the scoring manual for $50 (pdf) or $60 (printed) and test administrations are an additional recurring cost of $2 each (Mind Garden, 2016). Clinicians and researchers also need to score each assessment (20 items for self-report and 26 items for parent-report) and compare the scored results to the clinical cutoffs identified in the manual.

The CBCL has been one of the most widely used assessments of child behavior and treatment outcomes for multiple decades (Hatfield & Ogles, 2004; Kazdin, 1994) that has appeal in a variety of disciplines (e.g., social work, education; Early, Gregoire, & McDonald, 2001). Psychometric analyses have suggested that the Internalizing broadband scale of the CBCL (CBCL-I) demonstrates discriminate validity (Seligman, Ollendick, Langley, & Baldacci, 2004) and the items tend to correlate with other measures of internalizing symptoms (Nakamura, Ebesutani, Bernstein, & Chorpita, 2009). The CBCL-I has also demonstrated good test-retest reliability and internal consistency (Achenbach & Rescorla, 2001). However, concern has been expressed regarding the use of the CBCL-I as a tool to help identify anxiety in children because the scale tends to focus more on negative affect and contains fewer anxiety-specific items as compared to depression-specific items (Southam-Gerow et al., 2003). Previous research has also found the CBCL to be less sensitive to change than measures with more response options (McClendon et al., 2011). The CBCL has an even greater initial financial barrier to entry than the STAIC. While the CBCL can be hand-scored, a proprietary scoring software exists to automate the process, which can be purchased online for $395 and still requires a recurring cost for additional forms purchased in packs of 50 for $30 (ASEBA, 2012). The financial cost would increase further if a researcher or clinician also wanted to include the Youth Self-Report and Teacher Report Form (both are available at a recurring cost in packs of 50 for $30 each). The CBCL-I is also the longest measure included in the current study at 42 items.

The above review highlights the difficulty a researcher or practitioner might encounter when deciding how and when to select measures for monitoring a client's progress during treatment. The current study aims to contribute to the growing body of literature on EBA by using receiver operating characteristics (ROC) to compare the performance of each measure as a classifier of treatment response and diagnostic remission. Treatment response and diagnostic remission have been identified as two of the major outcomes of psychological interventions. In treatment contexts, clinical response has been defined as a magnitude change from pre- to post-treatment such that the individual has sufficiently reduced symptoms (frequently identified by reliable change scores) but might continue to endorse some symptoms (Tolin, Abramowitz, & Diefenbach, 2005). Remission refers to a client's movement from meeting diagnostic criteria for a psychological disorder to no longer meeting criteria (Frank et al., 1991). Assessing outcomes based on data from both continuous and discrete data is important as continuous outcomes provide a more sensitive dimension of change while discrete outcomes have the potential to speak to clinically significant change (Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999).

While ROC analysis has been primarily used as a tool to assess the diagnostic accuracy of measures as screeners for the presence or absence of a psychological disorder (Dierker et al., 2001; J. R. Williams et al., 2012), recent studies have utilized ROC analysis as a methodology to identify treatment non-responders (Steidtmann et al., 2013). Steidtmann and colleagues (2013) derived hypotheses from the literature on rapid response, which stated the clinical change should be noticeable by the fourth treatment session (Ilardi & Craighead, 1994). The researchers examined if the percentage of symptom reduction on the Inventory of Depressive Symptoms-Self-Report (IDS-SR; Rush, Gullion, Basco, Jarrett, & Trivedi, 1996) at weeks 4, 6, or 8 predicted remission on the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1967) at week 12 (Steidtmann et al., 2013). The results from the first step of their analysis found that percentage decrease on the IDS-SR at week 6 and week 8 (but not week 4) was related to HRSD status at the end of treatment (Steidtmann et al., 2013).

After identifying percentage reduction on the IDS-SR at week 6 and week 8 as predictors of HRSD status at week 12, experimenters then investigated the “performance” of each predictor (Steidtmann et al., 2013). Performance was defined as the sensitivity (correctly identify remitters at week 12) and specificity (correctly identify non-remitters at week 12) of the ROC curve. The results found that, for individuals in the psychotherapy-only condition, week 6 was more sensitive (χ^2=8.10, p<.01) and less specific (χ^2=5.14, p<.05) when compared to the sensitivity and specificity at week 8 (Steidtmann et al., 2013). This suggested that non-remitters might be identifiable, albeit with high false positive rates, as early was week 6 (of 12) for individuals receiving psychotherapy. Waiting until session 8 (of 12) to make decisions about the likelihood a patient might not remit during the current course of treatment will provide a better predictor of the patient's outcome. Taken together, the results suggested the IDS-SR might be better suited to longer treatment intervals. The paper provided an important step in using ROC analysis as a method to identify treatment non-responders during treatment.

The current paper aimed to use a similar methodology to test different models using the TPs, STAIC-T-P/C, and CBCL-I measures as classifiers of treatment response and diagnostic remission following a 16-week EBP for anxiety. The current study conducted a step-wise approach. The study first investigated the overall performance of each measure as a classifier of response and remission by comparing the difference between pre-treatment and post-treatment scores as a classifier of treatment response and diagnostic remission. Each measure that outperformed chance during the first step underwent an additional step similar to the approach used in previous research (Steidtmann et al., 2013) to identify the earliest point at which each measure could reliably classify response and remission. Each significant classifier was compared to the other significant classifiers in order to determine which, if any, of the measures performed best. It is important to note the study used outcomes reported by youth and their parents as a large body of literature suggests that youth and parents are discrepant in their report on psychological measures (De Los Reyes & Kazdin, 2005). The current study included youth and parent outcomes in order to provide information about each measure's ability to classify outcomes both within and across informants. The following hypotheses were generated:

1) As found in the previous literature on using ROC analysis as a tool to determine the diagnostic accuracy of screening instruments (e.g., J. R. Williams et al., 2012), it was hypothesized that all of the measures would significantly outperform chance when the difference between pre-treatment scores and post-treatment scores were used to classify treatment response and diagnostic remission.
    a. Similarly, it was hypothesized that within-informant performance of each measure would be superior to cross-informant performance (i.e., youth predictors would perform better than parent predictors for youth-reported outcomes and likewise for parent predictors and parent-reported outcomes).
2) For comparisons among the significant predictors, it was hypothesized that the narrower, idiographic measure (TPs) would respond to change earlier during treatment than the broader measure of distress (STAIC-P/C) and symptom inventory (CBCL-I). The idiographic measure provided youth and parents with the opportunity to track a unique problem that motivated their seeking treatment and, therefore, might have been more sensitive to change than the larger and broader measure of general distress and symptom checklists.

Method

Participants

Participants included youth (N = 165; ages 7-17; *M*=11.48, *SD*=2.42) and their parents recruited from an ongoing study at a university-based specialty clinic for youth with anxiety and depression disorders. Slightly over 50% of the youth were female

(84/165) and the majority were Caucasian (136/165, 84.50%). Specifically, youth who identified as minority were African American (11/165; 6.90%), Asian American (13/165; 8.10%), Latinx (14/165; 8.8%), or Other (4/165; 2.50%). The majority of mothers (81.10%) and fathers (75.80%) identified as Caucasian. Mothers who identified as a minority identified as African American (6.10%), Asian American (7.60%), Latina (3.80%), or Other (3.20%). Fathers who identified as a minority identified as African American (4.50%), Asian American (7.10%), Latino (7.80%), and Other (4.60%). The majority of parents reported that they were currently married (123/165; 74.50%). Information about family social status was determined using the Hollingshead criteria (Hollingshead, 1975). Both the average (43.76) and median (45.00) Hollingshead score suggested that families from the current sample were within the "Medium business, minor professional, technical" social strata (Hollingshead, 1975). Exclusion criteria were minimized to maximize external validity and included parent report of mental retardation, autism, psychotic disorder, or bipolar disorder. Inclusion criteria were (1) youth met diagnostic criteria for one or more anxiety disorders and (2) youth and parents responded to questionnaires at both pre-treatment and post-treatment.

Measures

Anxiety Disorders Interview Schedule for Children Parent/Child Interview (ADIS-P/C). The ADIS-P/C (Silverman & Albano, 2000) is a semi-structured interview that assesses the presence and severity of childhood disorders outlined by the revised fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR; American Psychiatric Association, 2000). Diagnostic profiles include parent, youth, and consensus diagnosis. Interference (Clinician's Severity Rating; CSR) is rated per disorder on a 0 (not at all) to 8 (debilitating) scale, where 4 represents clinical threshold. Interviewers were considered reliable when they matched expert ratings of diagnoses and CSRs (Cohen's κ ≥0.80). Actual mean interrater reliability was κ=0.94 (range=0.85–0.99). Thirty percent of study interviews were randomly selected and coded for study adherence. Reliability remained strong (mean κ=0.91, range=0.78–1.00).

State-Trait Anxiety Inventory for Children–Trait–Parent/Child Version (STAIC-T-P/C). The STAIC-T-P/C is a 26-item (parent) and 20-item (child) report form of anxiety symptoms (Strauss, 1987). Items are rated on a 1 (hardly ever) to 3 (often) scale (youth range=20–60; parent range=26–78). Strong psychometric properties including internal consistency, test-retest reliability, and convergent validity have been reported (Southam-Gerow & Chorpita, 2007). Cronbach's alpha was strong for the current study (range = .89-.97).

Child Behavior Checklist. The Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001) is a 112-item parent report measure that assesses a youth's behavior in various problem areas. The 42-item *internalizing* (CBCL-I; which assesses a youth's anxious, depressive, and withdrawn behavior) and 35-item *externalizing* (CBCL-E; which assess a youth's delinquent and aggressive behavior) broadband subscales were used for the current study. The CBCL-E was used as a control variable to verify that the hypothesized predictor variables were related significantly to the outcome variable when compared to an unhypothesized predictor variable. *T* scores for the internalizing and externalizing subscales range from 50 to 100 (*M*=50, SD=10) with scores between 65 and 69 representing the borderline clinical cases (3rd to 7th percentile) and scores of 70 and above representing clinical cases (2nd percentile). Previous research has indicated that the CBCL-I and CBCL-E have good psychometrics (Achenbach & Rescorla, 2001).

Target Problems (TP). Target problems (TP) is an idiographic measure derived from the Top Problems measure described in previous research (Weisz et al., 2011). TPs have been shown to provide information above and beyond what is collected in standardized measures such as the CBCL (Weisz et al., 2011). In the current study, the TP measure was administered separately to youth and parents immediately following the intake assessment in order to increase the likelihood that parents and youth would provide problems related to the youth's diagnoses. Youth and parents were asked to list the "problem they would most like to see addressed during treatment." The diagnostic interviewer wrote down the problems in the language provided by the respondent (e.g., "Not being able to sleep in my own room every night"). Therapists encouraged parents and children to provide specific, behavioral descriptions of TPs. Respondents were then asked to provide a second and then third problem not previously listed. Youth and parents then rated on a scale from 0 "not a problem" to 8 "a very big problem." Previous research has suggested that similar measures have demonstrated good test-retest reliability, convergent validity, discriminant validity, and slope reliability (Weisz et al., 2011). TPs could also be assessed individually or as a "Total Problem" by calculating the mean of all problems (Weisz et al., 2011). The current study used the "Total Problem" approach (i.e., mean of all target problems) as not all participants had three TPs. It is important to note, the TP measure used in the previous research had response options that ranged from 0 to 10 and no psychometrics have been calculated for the TP measure in the current study.

Revised Child Anxiety and Depression Scales–Child/Parent Versions (RCADS-C/P). The RCADS-C/P (Chorpita, Yim, Moffitt, Umemoto, & Francis, 2000) is a 47-item child- and parent-report questionnaire of youth symptoms consistent with DSM–IV anxiety and mood disorders. Items are rated on a 0 (never) to 3 (always) scale. The RCADS has demonstrated good internal consistency and strong convergent and discriminant validity. The 37-item Total Anxiety subscale (range=0–111) was used as the independent measure of treatment response for the current study. Good internal consistency (alpha = .77) was found for the current study.

Procedure

Treatment-seeking families contacted a university-based specialty clinic and completed a phone screen. Participants who were screened as possibly eligible were then mailed home an assessment battery that included the CBCL and STAIC-P/C and were scheduled for a no-cost intake assessment (ADIS-P/C Silverman & Albano, 2000). ADIS- P/C interviews were conducted by psychology doctoral students trained to criterion (see ADIS-IV reliabilities above). At the completion of the diagnostic interview, parents and youth each provided a list of up to three target problems to track during treatment. Participants who met diagnostic criteria for a primary anxiety disorder, based on interference scores from the ADIS-P/C, were then enrolled in a 16-week treatment using the Coping Cat (Kendall & Hedtke, 2006). Parents and youth completed symptom, behavior, and target problem forms every four weeks during treatment. Youth and parents also completed a second ADIS-P/C interview at the end of treatment to identify diagnostic remission. All procedures were approved by the institution’s Institutional Review Board.

Statistical Analysis

Prior to analysis, data were screened for univariate and multivariate outliers. Any datum that was three or more standard deviations above the mean was brought to the fence (median + 2 interquartile ranges). Missing data pattern analysis was conducted in SPSS 23 (SPSS IBM, New York, U.S.A.) using Little's Missing Completely at Random (MCAR) test. Missing data was replaced using expectation maximization (EM) SPSS 23 (SPSS IBM, New York, U.S.A.), which produces unbiased and efficient estimates (Graham, Cumsille, & Elek-Fisk, 2003).

Identifying treatment response and diagnostic remission. The first step in the analytic plan was to operationalize treatment "response" and diagnostic "remission." Commensurate with previous research (Beidas et al., 2014; Frank et al., 1991; Steidtmann et al., 2013; Tolin et al., 2005), "response" was operationalized as making a "reliable change" from pre- to post-treatment on an independent measure of youth anxiety (RCADS total anxiety subscale). The reliable change index (RCI; Jacobson & Truax, 1991) was used in order to identify reliable change cut-points for each measure included in the study. "Remission" was operationalized as the absence of the youth's principal diagnosis at post-treatment identified by the ADIS-P/C interview.

Comparing the effectiveness of each measure as a classifier or response and remission. In order to compare the effectiveness of the STAIC-T-P/C, CBCL-I, CBCL-E, and TP as predictors of response and remission, ROC curves were compared. ROC curves are a technique developed in the 1950's as a method to detect radio signal from noise (Green & Swets, 1966). ROC has since developed into a useful methodology for comparing the effectiveness of different diagnostic tools. ROC plots the "sensitivity" (true positive rate – correctly identify a diagnosis) against “1-specificity” (false positive rate – incorrectly identify a diagnosis) as well as provides information about true negatives (“specificity”) and false negatives (“1-sensitivity”). A diagnostic measure is said to be a perfect classifier (100% sensitivity and 100% specificity) at the point (0,1) of the ROC curve with 100% of the area underneath the curve. ROC curves are usually first compared to the line of non-discrimination, a diagonal line with a slope of 1 emanating from the origin and terminating at the point (1,1). The line of non-discrimination bisects the coordinate plan, meaning that 50% of the area in above the line and 50% of the area is below the line. Therefore, the line of non-discrimination represents “random guessing.” See Figure 1 for an example of different ROC curves. The areas under the curve (AUC) of different ROC curves can be compared in order to identify which diagnostic measure is a superior classifier. Measures with a significantly greater AUC are said to have a greater probability of distinguishing between a hit (correct classification) and a miss (incorrect classification). AUCs will be compared using the pROC package (Robin et al., 2011) for R (R Core Team, 2017).

In the first step of the analytic plan for the current study, AUC comparisons were made among the STAIC-T-P/C, CBCL-I/E, and TP for parent-reported and youth-reported outcomes separately. AUCs for each measure were first compared using the difference between baseline STAIC, CBCL, or TP scores and post-treatment scores as a predictor of response and remission at post-treatment. In the second step, the AUCs of the measures that were found to be significant in step 1 were compared at different intervals during treatment. ROC curves were generated for each measure during various intervals of treatment (i.e., baseline to week 4, baseline to week 8, baseline to week 12). In order to determine the optimal cut-point (i.e., change score that separates participants who are likely to respond or remit from those who are less likely to respond or remit) to generate the ROC curve, the current study calculated Youden's J (Youden, 1950). Youden's J identifies the cut-point that maximizes sensitivity and specificity. The cut-point was then used to determine which clients would be predicted to respond to treatment or achieve diagnostic remission. The predicted classification and true classification were then compared to generate the ROC curve and AUC.

Power Analysis

Estimated power was calculated using MedCalc 16.8 (MedCalc Software bvba, Ostend, Belgium, 2016), which follows the standard procedures outlined in previous research (Hanley & McNeil, 1982). "True" sample sizes for participants in the diagnosed (true positive) and undiagnosed group (true negative) at post-treatment were determined using the clinician consensus diagnosis from the ADIS-P/C. The power analysis estimated that the current study was sufficiently powered (≥.80) to detect an AUC as small as .631 for the remission outcome, .625 for the child response outcome, and .624 for the parent response outcome.

Results

Analyses followed a decision tree in order to determine (a) which measures were significantly better than chance at classifying remission or response and (b) which of the significant measures identified in step (a) were significantly different from each other. In order to determine which measures performed better than chance (i.e., an AUC significantly different from 50%), the difference between intake and post-treatment scores on youth- or parent-reported measures was used as a predictor of remission or response using the ROC option in SPSS version 23 (SPSS IBM, New York, U.S.A.). The decision to use the difference between intake and post-treatment scores as the initial analysis to identify significant predictors was made because the difference between intake and post-treatment is often used as a main outcome in clinical research trials and many monitoring feedback systems (Lambert et al. 2001). Measures that were identified as significantly different from chance were then compared across multiple time points using the pROC (Robin et al., 2011) for R (R Core Team, 2017).

Table 1 shows the results from the initial ROC analysis that compared each predictor as a classifier of response or remission to chance using data collected from pre- to post-treatment. Youth-reported measures were predictive at a rate greater than chance for all three outcomes (remission, youth-reported response, and parent-reported response) whereas parent-reported measures were predictive at a rate greater than chance for response outcomes only. For example, rows 2 and 3 show that youth-reported STAIC scores (*p*<.001) and youth-report TPs (*p*<.05) classified remitters above chance. None of the parent-reported change measures predicted remission above chance. Both parent- and child- reported measures predicted child response and parent response, but child measures did so more consistently. Effect sizes comparing the predicted diagnostic status to observed diagnostic status were mostly in the fair to moderate range, using Landis and Koch's (1977) guidelines where the ranges represent the following: [.00-.20 = slight agreement], [.21-.40 = fair], [.41-.60 = moderate], [.61-.80 = substantial], and [.81-1.00 =almost perfect]. The table also shows that symptom inventories (i.e., STAIC, CBCL) tended to have the greatest effect sizes. The last column of Table 1 reports on the comparisons among the significant predictors identified in column 3. For example, the final column of the “Child STAIC” row under the “Outcome: Child Response” section shows that Child STAIC scores were significantly better than Child TPs and Parent STAIC at predicting the outcome of child response. Columns 4 and 5 report on the specific cut-point (in change scores) that maximized sensitivity and specificity for each measure and outcome. Finally, Columns 6 and 7 present the positive predictive value (PPV) and negative predictive value (NPV).

The measures identified as significantly different from chance in the above analysis were further analyzed across treatment intervals of increasing 4-week increments (i.e., intake to treatment week 4, intake to treatment week 8, and intake to treatment week 12). First, the measures were, once again, compared to chance. Second, if both measures were significantly different from chance, DeLong’s method for comparing AUCs was used to compare differences among the significant AUCs (DeLong, DeLong, & Clarke-Pearson, 1988).

Tables 2-4 show the results from the comparison of the significant predictors identified in Table 1 across three, progressively longer treatment intervals in predicting remission, youth response, and parent response. For example, the first 3 rows under the “Outcome: Child Response” section in Table 3 show the results for youth-reported STAIC, youth-reported TPs, and parent-reported STAIC during the 0-4 week treatment interval. The results presented in the table identify the AUC (third column), if the difference between pre-treatment and the post-treatment interval predicted the identified outcome above chance (column 4), the raw score cut-point for each measure used in the ROC analysis as the predictor of treatment outcomes (column 5), the effect size of the comparison (column 6), the PPV (column 7), the NPV (column 8), and if the measures were significantly different from each other (column 9). For the example rows, the differences in the scores between pre-treatment and week 4 of treatment for youth-reported STAIC was the only measure that outperformed chance during the weeks 0-4 interval. The difference had a fair effect size (.38) and both the youth-reported TPs and parent-reported STAIC were significantly worse than youth-reported STAIC at predicting youth-reported treatment response after 4 weeks of treatment.

As seen in table 2, youth-reported STAIC scores and TPs were able to classify remission at above chance levels after 8 weeks of treatment (column 4) and neither measure outperformed the other (column 7). Response outcomes were able to be classified above chance levels as early as 4 weeks into treatment. For example, youth-reported STAIC scores were able to classify both youth-reported response and parent-reported response after 4 weeks of treatment. Parent- and youth-reported STAIC scores and the CBCL-I were able to predict parent-reported response after 4 weeks of treatment. Table 2 also shows the expected trend that measures became better predictors of outcomes compared to chance during larger treatment intervals.

Discussion

The current study aimed to investigate the performance of three measures of differing breadth as classifiers of diagnostic remission (i.e., no longer meeting diagnostic criteria at post-treatment) and treatment response (i.e., reliable change at post-treatment). The study examined change scores on each measure, between pre-treatment and post-treatment, as a predictor of remission and response. The results of the initial ROC analysis found a difference between the outcomes of youth- and parent-reported measures. Youth-reported measures classified (1) youth-reported response (2) parent-reported response and (3) diagnostic remission above chance levels. Parent-reported measures classified (1) youth- and (2) parent-reported response, but not (3) diagnostic remission above chance. Follow-up ROC analyses revealed that youth-reported STAIC and TP measures performed similarly across treatment when predicting remission but youth-reported STAIC scores outperformed all other measures for the outcome of youth-reported response. Similarly, the outcome of parent-reported response was best classified by parent-reported STAIC scores and the CBCL-I. Psychometric strengths of the reporting group, youth or parent, should be considered along with cost and participant burden when deciding which measures to use in measurement feedback systems.

The first finding of note was that youth-reported measures were able to classify diagnostic remission status above chance while parent-reported measures were not able to predict remission. Discrepancies between youth and parent report of outcomes are common in the literature (De Los Reyes & Kazdin, 2005), and it is possible that youth were able to detect subtle predictive changes better than their parents. This is particularly true in youth with internalizing disorders who may have greater awareness of their concerns (e.g., symptoms, functional impairments) than parents (De Los Reyes & Kazdin, 2005). Likewise, remission has been described as a more conservative outcome than response (Caporino et al., 2013) and larger differences on measures may be needed in order to predict treatment remission than needed to predict treatment response (Caporino et al., 2013; Steidtmann et al., 2013). Thus, youth might be more reliable in predicting remission as a diagnostic outcome for internalizing disorders as they might be more sensitive to changes in symptomatology and functioning.

It is worth noting that both youth-reported measures (i.e., STAIC, TPs) performed similarly in their ability to classify diagnostic remission and the pattern of results held throughout the different treatment intervals. Both measures were able to classify remission by treatment midpoint (week 8) with no significant differences between their performances for any treatment interval. When few statistical differences exist between measures, it is important to consider other metrics that might help inform the decision to implement a single tool. For example, the TP measure might compare favorably against the STAIC in a cost-benefit analysis in that the TP measure is an easily scalable and comprehendible measure that is extremely brief and written in the client's language. The measure can be created on-the-fly and does not need any sophisticated scoring aids or interpretation guides. The finding contributes to the growing body of literature detailing the potential of idiographic measures as useful tools for monitoring and predicting treatment outcomes (Lindhiem, Bennett, Orimoto, & Kolko, 2016). At the same time, standardized tools, like the STAIC, may retain value to the extent that it provides useful screening information for diagnostic disorders or provides subscales that hint at underlying mechanisms. The choice will depend on the prioritized goal the clinician wishes to achieve.

The second finding of note was that youth-reported response status was classified by youth-reported TPs, youth-reported STAIC (STAIC-C), and parent-reported STAIC (STAIC-P). The finding provided evidence for the consistency of youth-reported measures in that they were able to predict diagnostic remission status and youth-reported response to treatment status by post-treatment. The STAIC also received support as it was robust across reporters in its ability to classify response to treatment based on youth-report. The follow-up analyses that investigated the performance of each measure across treatment intervals (sessions 0-4, 0-8, and 0-12) found that the STAIC-C tended to outperform the other measures. Specifically, the STAIC-C was able to classify youth-reported response outcomes by treatment week 4 and consistently outperformed youth-reported TPs and the STAIC-P across all treatment intervals.

The third finding of note was that the outcome of parent-reported response status was classified by all youth-reported measures (TPs, STAIC-C) and parent-reported measures (TPs, STAIC-P, CBCL-I, CBCL-E). Once again, the results found that youth-reported measures were consistent in their ability to classify outcomes. Parents, on the other hand, demonstrated less consistency in their report of youth symptomatology – with parent-reported measures only predicting the outcome of treatment response, not remission. Relatedly, both the CBCL-I and CBCL-E were able to classify parent-reported response status, which suggests that parents might have difficulty discriminating between factors that might contribute to their child's distress. It might also be the case that parents see an increase in externalizing symptoms in their child who is diagnosed with an internalizing disorder. Despite the lack of discernment in predicting parent-reported response, the follow-up analyses revealed that the STAIC-C, STAIC-P, and CBCL-I tended to outperform the other measures across the various treatment intervals. Each measure predicted parent-reported response by week 4. However, only the STAIC-P and CBCL-I consistently outperformed other measures across each treatment interval (i.e., 0-4, 0-8, 0-12). The findings paralleled those for classifying response based on youth report: symptom inventories tended to have better diagnostic metrics than idiographic measures earlier in treatment that remained better throughout treatment.

Combining the findings from all three outcomes provides insight into selecting a measure to be used in a MFS. First, the findings suggested that youth-report was more robust than parent-report. Youth-reported measures were able to predict all three outcomes included in the current study (i.e., diagnostic remission, youth-reported response, and parent-reported response). However, despite the robustness of youth-reported measures, the findings suggested that measures tended to perform best as classifiers of treatment outcomes when the predictor and outcome were completed by the same informant (e.g., parent-reported STAIC scores better predicted parent-reported response). Second, symptom inventories appeared to be robust in their classification of all three outcomes. Specifically, if a researcher or practitioner elected to use a single measure for tracking the progress of a client diagnosed with an anxiety disorder, the STAIC might be the tool of choice as the measure was able to classify remission, youth-reported response, and parent-reported response. However, idiographic target problem instruments can offer a cost-effective alternative. Third, the results contributed additional support to the concept that treatment response can be predicted early during treatment (Ilardi & Craighead, 1994). Diagnostic remission was predicted as early as treatment midpoint and both youth- and parent-reported response were predicted by treatment week 4. Knowing that treatment outcomes can be predicted early in treatment should encourage practitioners to use MFSs as a tool to help improve psychological interventions. If a client does not achieve symptom reduction (or exhibits symptom increase) equal-to-or-greater than the empirically-determined cut-point associated with predicting or classifying a positive outcome, then it might be worthwhile to shift the course of treatment or implement new strategies to help improve client functioning. The current study expands on previous research that used empirically-derived cut-points to use as a guide for predicting which youth have a good chance of being treatment responders or remitters early in treatment (e.g., Steidtmann et al., 2013) by including multiple measures of differing breadth.

The study has a number of limitations. First, the sample used to identify the ROC cut-points was the same as the sample that was used to test the identified cut-points as thresholds for identifying treatment response and diagnostic remission. Previous research has suggested that using independent samples, or a split sample, yields increased power and increased confidence in the obtained results (Steidtmann et al., 2013). As such, caution should be used when generalizing the results from the current study. Second, session intervals were limited to every four sessions. While collecting information on a weekly basis might increase client burden, the additional information would help fine tune the earliest point at which treatment responders or remitters could be identified. Third, the results do not provide a crystal ball for predicting who will and who will not respond and/or remit following CBT for anxiety. The results simply provide an additional metric that can be used to inform researchers and practitioners about the classification accuracy of a measure given a specific difference score (e.g., between intake and week 4). Fourth, the measure used for youth- and parent-reported response (RCADS) shared some similarity to the symptom inventories used as predictors in the current study (STAIC, CBCL). Some of the superior performance of symptom inventories as a classifier of treatment response might have been due to the similarity between the measures. Future research should include multiple measures of treatment response that assess various domains of client progress (e.g., functional impairment). Finally, the study cannot comment on additional metrics of measure utility (e.g., scalability, cost-effectiveness) because two of the three measures required upfront and continued costs rendering the third, free measure (TP) the default winner in all comparisons. Future research looking to contribute to the growing knowledge about EBA and progress monitoring might consider using one or more of the high quality measures that are available for free (Beidas et al., 2015). Using free measures would remove some of the inherent advantages held by the TP measure in the current study and allow for a fairer investigation of the utility of each measure as a classifier of (non)remission or (non)response.

The current study also has a number of strengths. First, the study followed the suggestions of previous research to investigate the utility of different diagnostic and progress monitoring measures in a setting that more closely resembles community clinical practice (Steidtmann et al., 2013) by conducting the study using an open trial. The study also used separate measures for the predictors and outcomes. Finally, though two of the measures used in the current study require a fee, the use of commonly used measures increases the transportability of the findings to clinical practices.

Future research can build upon the findings from the current project by first testing the cut-points in a separate sample. If additional studies using the same cut-points find similar results, then we could be more confident in our understanding of how and when it might be best to utilize measures of different breadth. Future investigators might also consider tracking client progress with multiple measures on a weekly basis. While there is a large body of research that suggests patients who respond to treatment are identifiable by session four (Ilardi & Craighead, 1994), there is room for more research about the first three sessions. Perhaps we might be able to make data-informed probability predictions about a patient's likelihood of response or remission earlier than session four.

The current study aimed to investigate the utility of three measures of differing breadth as classifiers of diagnostic remission and response. Identification of responders or remitters earlier in treatment would have the benefit of helping clinicians identify which clients are at risk for not responding and either reevaluate the case conceptualization, change the treatment approach, or refer the patient to a more appropriate clinic. The results found that youth report was robust in that youth reported STAIC and TPs were able to predict both remission and response. While the STAIC was a superior measure for classifying response, both measures performed equally well when predicting remission, suggesting that free, idiographic measures hold great promise as a tool to track and predict patient progress during treatment. Parent report only predicted response outcomes, with symptom inventories tending to outperform idiographic measures. The difference in classifying remission and response suggests that remission is a more complex outcome that includes information from both symptoms and functional impairments and can be classified by measures that take either or both facets into consideration whereas response is a more unidimensional outcome that is best predicted by more symptom-focused measures.

References

Achenbach, T. M., & Rescorla, L. A. (2001). Manual for the ASEBA school-age forms & profiles: an integrated system of multi-informant assessment Burlington, VT: University of Vermont. *Research Center for Children, Youth, & Families*.

Alfano, C. A. (2012). Are children with "pure" generalized anxiety disorder impaired? A comparison with comorbid and healthy children. *Journal of Clinical Child and Adolescent Psychology*, *41*(6), 739–745. https://doi.org/10.1080/15374416.2012.715367

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR®*. Arlington, VA: American Psychiatric Publishing.

Andrews, G. (1995). *The most widely used outcomes measures: A user's guide and critical review* [Cassette Recording No. BHG95-20]. Portola Valley, CA: Institute for Behavioral Healthcare.

Anker, M. G., Duncan, B. L., & Sparks, J. A. (2009). Using client feedback to improve couple therapy outcomes: A randomized clinical trial in a naturalistic setting. *Journal of Consulting and Clinical Psychology*, *77*(4), 693.

ASEBA. (2012). Child Behavior Checklist 6-18-store.aseba.org. Retrieved September 3, 2016, from http://store.aseba.org/Child-Behavior-Checklist-6-18/products/19/

Beidas, R. S., Lindhiem, O., Brodman, D. M., Swan, A., Carper, M., Cummings, C., … others. (2014). A probabilistic and individualized approach for predicting treatment gains: An extension and application to anxiety disordered youth. *Behavior Therapy*, *45*(1), 126–136.

Beidas, R. S., Stewart, R. E., Walsh, L., Lucas, S., Downey, M. M., Jackson, K., … Mandell, D. S. (2015). Free, brief, and validated: Standardized instruments for low-resource mental health settings. *Cognitive and Behavioral Practice*, *22*(1), 5–19.

Berger, T., Hohl, E., & Caspar, F. (2009). Internet-based treatment for social phobia: a randomized controlled trial. *Journal of Clinical Psychology*, *65*(10), 1021–1035.

Bickman, L. (2008). A Measurement Feedback System (MFS) Is Necessary to Improve Mental Health Outcomes. *Journal of the American Academy of Child and Adolescent Psychiatry*, *47*(10), 1114–1119. https://doi.org/10.1097/CHI.0b013e3181825af8

Bickman, L., Kelley, S. D., Breda, C., de Andrade, A. R., & Riemer, M. (2011). Effects of routine feedback to clinicians on mental health outcomes of youths: Results of a randomized trial. *Psychiatric Services*, *62*(12), 1423–1429.

Bruns, E. J., & Hoagwood, K. E. (2008). State Implementation of Evidence-Based Practice for Youths, Part I: Responses to the State of the Evidence. *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*(4), 369–373. https://doi.org/10.1097/CHI.0b013e31816485f4

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist*, *15*(8), 546.

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

Caporino, N. E., Brodman, D. M., Kendall, P. C., Albano, A. M., Sherrill, J., Piacentini, J., … Walkup, J. T. (2013). Defining Treatment Response and Remission in Child Anxiety: Signal Detection Analysis Using the Pediatric Anxiety Rating Scale. *Journal of the American Academy of Child and Adolescent Psychiatry*, *52*(1), 57–67. https://doi.org/10.1016/j.jaac.2012.10.006

Carlier, I. V., Meuldijk, D., Van Vliet, I. M., Van Fenema, E., Van der Wee, N. J., & Zitman, F. G. (2012). Routine outcome monitoring and feedback on physical or mental health status: evidence and theory. *Journal of Evaluation in Clinical Practice*, *18*(1), 104–110.

Centers for Medicare & Medicaid Services. (2016a, May 24). Payment Adjustment Information. Retrieved June 17, 2016, from https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/Payment-Adjustment-Information.html

Centers for Medicare & Medicaid Services. (2016b, May 26). Physician quality reporting system: Overview. Retrieved June 17, 2016, from https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/index.html?redirect=/pqri/

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*(1), 7.

Chorpita, B. F., Yim, L., Moffitt, C., Umemoto, L. A., & Francis, S. E. (2000). Assessment of symptoms of DSM-IV anxiety and depression in children: A revised child anxiety and depression scale. *Behaviour Research and Therapy*, *38*(8), 835–855.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*(3), 837–845.

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: a critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*(4), 483.

Dierker, L. C., Albano, A. M., Clarke, G. N., Heimberg, R. G., Kendall, P. C., Merikangas, K. R., … Kupfer, D. J. (2001). Screening for anxiety and depression in early adolescence. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(8), 929–936.

Early, T. J., Gregoire, T. K., & McDonald, T. P. (2001). An assessment of the utility of the Child Behavior Checklist/4-18 for social work practice. *Research on Social Work Practice*, *11*(5), 597–612. https://doi.org/10.1177/104973150101100504

Frank, E., Prien, R. F., Jarrett, R. B., Keller, M. B., Kupfer, D. J., Lavori, P. W., … Weissman, M. M. (1991). Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Remission, recovery, relapse, and recurrence. *Archives of General Psychiatry*, *48*(9), 851–855.

Ginsburg, G. S., La Greca, A. M., & Silverman, W. K. (1998). Social anxiety in children with anxiety disorders: Relation with social and emotional functioning. *Journal of Abnormal Child Psychology*, *26*(3), 175–185.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. *Handbook of Psychology*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/0471264385.wei0204/full

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. *Society*, *1*, 521.

Hamilton, M. A. X. (1967). Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology*, *6*(4), 278–296.

Hollingshead, A. B. (1975). *Four factor index of social status*. Unpublished manuscript, Yale University, New Haven, CT. Retrieved from http://www.academia.edu/download/30754699/yjs_fall_2011.pdf#page=21

Ilardi, S. S., & Craighead, W. E. (1994). The role of nonspecific factors in cognitive-behavior therapy for depression. *Clinical Psychology: Science and Practice*, *1*(2), 138–155.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12.

Kazdin, A. E. (2005). Evidence-Based Assessment for Children and Adolescents: Issues in Measurement Development and Clinical Application. *Journal of Clinical Child & Adolescent Psychology*, *34*(3), 548–558. https://doi.org/10.1207/s15374424jccp3403_10

Kazdin, A. E., & Blase, S. L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on Psychological Science*, *6*(1), 21–37.

Kendall, P. C., & Hedtke, K. A. (2006). *Cognitive-behavioral therapy for anxious children: Therapist manual*. Workbook Pub.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*(3), 285–299. https://doi.org/10.1037/0022-006X.67.3.285

Kiresuk, T. J., & Sherman, M. R. E. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal*, *4*(6), 443–453.

Kirisci, L., Clark, D. B., & Moss, H. B. (1997). Reliability and validity of the State-Trait Anxiety Inventory for Children in adolescent substance abusers: Confirmatory factor analysis and item response theory. *Journal of Child & Adolescent Substance Abuse*, *5*(3), 57–70.

Kolko, D. J., Lindhiem, O., Hart, J., & Bukstein, O. G. (2014). Evaluation of a Booster Intervention Three Years After Acute Treatment for Early-Onset Disruptive Behavior Disorders. *Journal of Abnormal Child Psychology*, *42*(3), 383–398. https://doi.org/10.1007/s10802-013-9724-1

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*(2), 159.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lindhiem, O., Bennett, C. B., Orimoto, T. E., & Kolko, D. J. (2016). A Meta-Analysis of Personalized Treatment Goals in Psychotherapy: A Preliminary Report and Call for More Studies. *Clinical Psychology: Science and Practice*, *23*(2), 165–176. https://doi.org/10.1111/cpsp.12153

McCauley, E., Katon, W., Russo, J., Richardson, L., & Lozano, P. (2007). Impact of anxiety and depression on functional impairment in adolescents with asthma. *General Hospital Psychiatry*, *29*(3), 214–222. https://doi.org/10.1016/j.genhosppsych.2007.02.003

McClendon, D. T., Warren, J. S., M. Green, K., Burlingame, G. M., Eggett, D. L., & McClendon, R. J. (2011). Sensitivity to change of youth treatment outcome measures: a comparison of the CBCL, BASC-2, and Y-OQ. *Journal of Clinical Psychology*, *67*(1), 111–125. https://doi.org/10.1002/jclp.20746

Merikangas, K. R., He, J., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., … Swendsen, J. (2010). Lifetime Prevalence of Mental Disorders in US Adolescents: Results from the National Comorbidity Study-Adolescent Supplement (NCS-A). *Journal of the American Academy of Child and Adolescent Psychiatry*, *49*(10), 980–989. https://doi.org/10.1016/j.jaac.2010.05.017

Mind Garden. (2016). State-Trait Anxiety Inventory for Children - Mind Garden. Retrieved September 3, 2016, from http://www.mindgarden.com/146-state-trait-anxiety-inventory-for-children

Nakamura, B. J., Ebesutani, C., Bernstein, A., & Chorpita, B. F. (2009). A psychometric analysis of the child behavior checklist DSM-oriented scales. *Journal of Psychopathology and Behavioral Assessment*, *31*(3), 178–189.

R Core Team, R. C. (2017). *R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017*. ISBN3-900051-07-0 https://www. R-project. org.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77.

Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B., & Trivedi, M. H. (1996). The inventory of depressive symptomatology (IDS): psychometric properties. *Psychological Medicine*, *26*(03), 477–486.

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 and Adolescent Psychology*, *33*(3), 557–565.

Shefler, G., Dasberg, H., & Ben-Shakhar, G. (1995). A randomized controlled outcome and follow-up study of Mann's time-limited psychotherapy. *Journal of Consulting and Clinical Psychology*, *63*(4), 585.

Silverman, W. K., & Albano, A. M. (2000). *Anxiety Disorders Interview Schedule for Children (ADIS-IV) Child and Parent Interviews*. New York, NY: Oxford University Press.

Southam-Gerow, M. A., & Chorpita, B. F. (2007). Anxiety in children and adolescents. *Assessment of Childhood Disorders*, *4*, 347–397.

Southam-Gerow, M. A., Flannery-Schroeder, E. C., & Kendall, P. C. (2003). A psychometric evaluation of the parent report form of the State-Trait Anxiety Inventory for Children—Trait Version. *Journal of Anxiety Disorders*, *17*(4), 427–446. https://doi.org/10.1016/S0887-6185(02)00223-2

Spielberger, C. D. (1983). Manual for the State-Trait Anxiety Inventory STAI (form Y)(“ self-evaluation questionnaire”). Retrieved from https://ubir.buffalo.edu/xmlui/handle/10477/1873

Steidtmann, D., Manber, R., Blasey, C., Markowitz, J. C., Klein, D. N., Rothbaum, B. O., … Arnow, B. A. (2013). Detecting critical decision points in psychotherapy and psychotherapy + medication for chronic depression. *Journal of Consulting and Clinical Psychology*, *81*(5), 783–792. https://doi.org/10.1037/a0033250

Strauss, C. (1987). Modification of trait portion of State-Trait Anxiety Inventory for Children-parent form. *Gainesville, FL: University of Florida*.

Tolin, D. F., Abramowitz, J. S., & Diefenbach, G. J. (2005). Defining response in clinical trials for obsessive-compulsive disorder: a signal detection analysis of the Yale-Brown obsessive compulsive scale. *The Journal of Clinical Psychiatry*, *66*(12), 1549–1557.

Weisz, J. R., Chorpita, B. F., Frye, A., Ng, M. Y., Lau, N., Bearman, S. K., … The Research Network on Youth Mental Health. (2011). Youth top problems: Using idiographic, consumer-guided assessment to identify treatment needs and to track change during psychotherapy. *Journal of Consulting and Clinical Psychology*, *79*(3), 369–380. https://doi.org/10.1037/a0023307

Williams, J. R., Hirsch, E. S., Anderson, K., Bush, A. L., Goldstein, S. R., Grill, S., … Palanci, J. (2012). A comparison of nine scales to detect depression in Parkinson disease Which scale to use? *Neurology*, *78*(13), 998–1006.

Williams, S. (2008). *Mental health screening and assessment tools for children: Literature review* (Literature review) (p. 116). UC Davis Extension.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35.

Table 1. Differences between intake and post-treatment on each measure as a predictor of each outcome.

Outcome: Remission	AUC	Different from chance?	Cut-point	Effect size	PPV	NPV	Different from other significant classifiers?
Child STAIC	.69	*p*<.001	7.07	.38	.79	.58	No difference
Child TP	.62	*p*<.05	2.91	.25	.77	.46	No difference
Parent STAIC	.52	N.S.	-	-			-
Parent TP	.58	N.S.	-	-			-
CBCL-I	.47	N.S.	-	-			-
CBCL-E	.56	N.S.	-	-			-
Outcome: Child Response	AUC	Different from chance?		Effect Size			Different from other sig classifiers?
Child STAIC	.89	*p*<.001	11.04	.62	.90	.72	>Child TP***, >Parent STAIC***
Child TP	.67	*p*<.001	2.82	.30	.72	.57	<Child STAIC***
Parent STAIC	.64	*p*<.01	12.95	.24	.71	.53	<Child STAIC***
Parent TP	.46	N.S.	-	-			-
CBCL-I	.54	N.S.	-	-			-
CBCL-E	.50	N.S.	-	-			-
Outcome: Parent Response	AUC	Different from chance?		Effect Size			Different from other sig classifiers?
Child STAIC	.61	*p*<.05	8.44	.25	.80	.61	<Parent STAIC***, < Parent TP*, <CBCL-I**
Child TP	.61	*p*<.05	2.72	.21	.64	.57	<Parent STAIC***, <Parent TP**, <CBCL-I***
Parent STAIC	.84	*p*<.001	11.15	.56	.77	.81	>Child STAIC***, >Child TP***, >CBCL-E***
Parent TP	.75	*p*<.001	3.62	.38	.65	.60	>Child STAIC, >Child TP
CBCL-I	.79	*p*<.001	6.97	.48	.74	.75	>Child STAIC**, >Child TP***, >CBCL-E**
CBCL-E	.68	*p*<.001	3.71	.28	.67	.61	<Parent STAIC***, <CBCL-I**

*<.05 **<.01 ***<.001

Note: > means the change measure was significantly better at predicting the outcome (e.g., youth-reported STAIC was significantly better at predicting child response than either youth-reported target problems or parent-reported STAIC). < means the change measure was significantly worse at predicting the outcome.

Table 2. Comparisons among measures at different treatment intervals for remission.

Outcome: Remission	Treatment Interval	AUC	Better than chance?	Cut-point	Effect size	PPV	NPV	Measures that were worse
Child STAIC	Weeks 0-4	.61	*p*=.06	-	-	.87	.48	-
Child TP	Weeks 0-4	.59	*p*=.12	-	-	.77	.43	-
Child STAIC	Weeks 0-8	.62	*p*<.05	8.5	.26	.77	.51	-
Child TP	Weeks 0-8	.67	*p*<.01	3.33	.16	.8	.45	-
Child STAIC	Weeks 0-12	.62	*p*<.05	7.5	.26	.71	.55	-
Child TP	Weeks 0-12	.61	*p*=.07	-	-	.73	.52	-

Table 3. Comparisons among measures at different treatment intervals for child response.

Outcome: Youth Response	Treatment Interval	AUC	Better than chance?	Cut-point	Effect size	PPV	NPV	Measures that were worse
Child STAIC	Weeks 0-4	.70	*p*<.001	5.5	.38	.82	.57	Child TP*; STAIC-P**
Child TP	Weeks 0-4	.54	*p*=.47	-	-	.73	.47	-
Parent STAIC	Weeks 0-4	.55	*p*=.43	-	-	.68	.45	-
Child STAIC	Weeks 0-8	.82	*p*<.001	6.5	.50	.85	.64	Child TP***; STAIC-P***
Child TP	Weeks 0-8	.60	*p*=.08	-	-	.70	.47	-
Parent STAIC	Weeks 0-8	.56	*p*=.06	-	-	.66	.44	-
Child STAIC	Weeks 0-12	.79	*p*<.001	6.5	.51	.81	.70	Child TP*; STAIC-P**
Child TP	Weeks 0-12	.65	*p*<.05	2.17	.24	.73	.50	-
Parent STAIC	Weeks 0-12	.60	*p*=.10	-	-	.78	.60	-

Table 4. Comparisons among measures at different treatment intervals for parent response.

Outcome: Parent Response	Treatment Interval	AUC	Better than chance?	Cut-point	Effect size	PPV	NPV	Measures that were worse
Child STAIC	Weeks 0-4	.64	$p<.05$	10.5	.22	.77	.52	-
Child TP	Weeks 0-4	.53	$p=.55$	-	-	.66	.5	-
Parent STAIC	Weeks 0-4	.76	$p<.001$	6.5	.37	.77	.61	Child TP*; Parent TP**; CBCL-E*
Parent TP	Weeks 0-4	.49	$p=.86$	-	-	.57	.54	-
CBCL-I	Weeks 0-4	.68	$p<.05$	3.5	.41	.83	.59	Parent TP**
CBCL-E	Weeks 0-4	.56	$p=.42$	-	-	.92	.47	-
Child STAIC	Weeks 0-8	.62	$p<.05$	4.5	.27	.66	.62	-
Child TP	Weeks 0-8	.56	$p=.31$	-	-	.66	.49	-
Parent STAIC	Weeks 0-8	.76	$p<.001$	8.5	.38	.77	.62	Child TP*; Parent TP**; CBCL-E*
Parent TP	Weeks 0-8	.63	$p<.05$	1.80	.26	.74	.55	-
CBCL-I	Weeks 0-8	.74	$p=.001$	5.5	.43	.88	.57	CBCL-E**
CBCL-E	Weeks 0-8	.57	$p=.29$	-	-	.69	.55	-
Child STAIC	Weeks 0-12	.62	$p<.05$	5.5	.30	.69	.61	-
Child TP	Weeks 0-12	.62	$p=.06$	-	-	.73	.54	-
Parent STAIC	Weeks 0-12	.87	$p<.001$	9.5	.61	.86	.75	Child STAIC***; Child TP***; Parent TP***; CBCL-E**
Parent TP	Weeks 0-12	.70	$p<.01$	2.5	.33	.79	.56	-
CBCL-I	Weeks 0-12	.80	$p<.001$	6.5	.48	.90	.59	Child STAIC*; Child TP*
CBCL-E	Weeks 0-12	.69	$p<.01$	3.5	.38	.81	.56	-

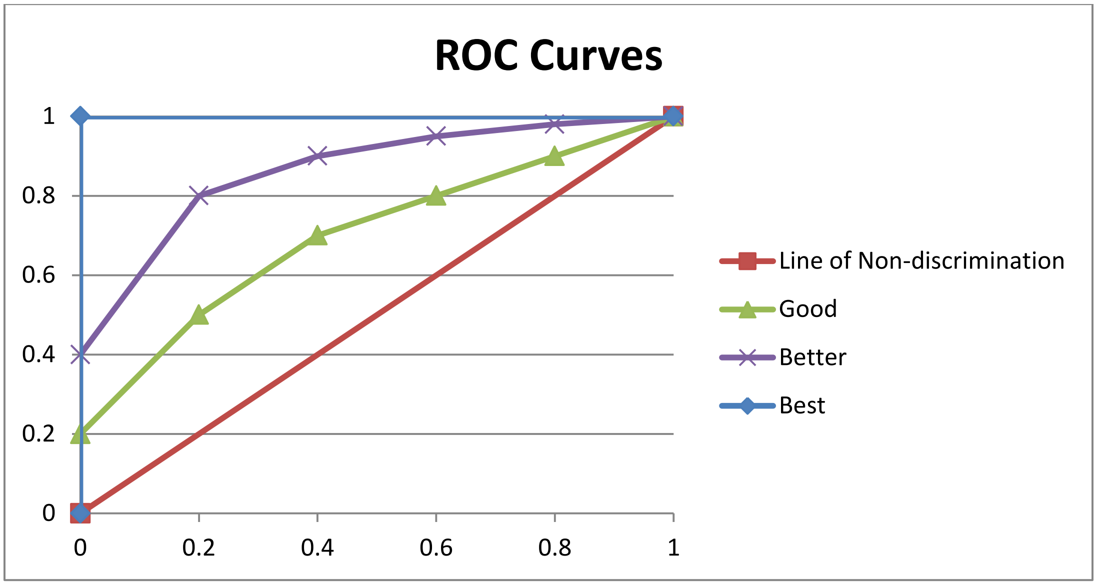

Figure 1. Examples of different ROC curves.