2015

Adherence to and Competence in Cognitive Behavioral Therapy for Youth Anxiety: Psychometric Evaluation

Cassidy C. Arnold
Virginia Commonwealth University

ADHERENCE TO AND COMPETENCE IN COGNITIVE BEHAVIORAL THERAPY FOR YOUTH ANXIETY: PSYCHOMETRIC EVALUATION

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

By: CASSIDY C. ARNOLD
B.A., University of Colorado, Denver, 2006
M.S., Virginia Commonwealth University, 2015

Major Director: Michael A. Southam-Gerow, Ph.D.
Professor
Department of Psychology and Pediatrics

Virginia Commonwealth University
Richmond, Virginia
August, 2015

Acknowledgment

This dissertation has been possible thanks to my wonderful mentors, colleagues, and, of course, family. I would first like to thank my fellow graduate students, Alexis Quinoy, Alison Eonta, Carrie Tully, Emily Wheat, Nadia Islam, Kimberly Nicolas, and Adriana Rodríguez, who have logged countless hours coding the psychotherapy sessions that serve as the basis for this study. Adriana's camaraderie has been especially wonderful and supportive.

I would like to thank Drs. Michael Southam-Gerow and Bryce McLeod for their guidance and generosity throughout my time at VCU and this dissertation process. Dr. Southam-Gerow, as my advisor and dissertation chair, provided a crucial combination of encouragement, feedback, and freedom to help me keep this project moving forward. In addition to his guidance with this project, Dr. Southam-Gerow's kindness, enthusiasm, wisdom, and humor has been invaluable in navigating the academic and life challenges posed by graduate school.

I am so grateful for my family for their support and for fostering in me the values of compassionate service for others, education and science, and hard work. Your guidance and support were what led me to pursue my interests in clinical science and a doctorate in psychology. In my eyes, Lloyd and Joan Arnold, your work with the YMCA and primary school education, and your actions in daily life, have taught me how to live in concordance with these values. Rick Arnold, you are the kindest man I know and as fine a role model as there is. Your dedication to your patients, students, colleagues, community near and far, and family are characteristics that I aspire to emulate. Teri Clifford, you have modeled for me a dedication to underserved youth and families, the power of education, and a commitment to service improvement through your work in childhood education. Dan O'Connell, you have nurtured my love of science and modeled the practices and virtues of being a lifelong learner. Gina Clifford, I only wish that I had more memories of your wisdom, compassion, and love. I thank you all for your love!

And finally, Lily Christon Arnold, you have been my guide, confidant, cheerleader, right hand, editor, and crashpad since the first days of graduate school. You have endured more conversations about reliability and validity than anybody should be subjected to and remained as kind and loving as ever. I am forever grateful to you!

Table of Contents

List of Tables

List of Figures

Page

Abstract

ADHERENCE TO AND COMPETENCE IN COGNITIVE BEHAVIORAL THERAPY FOR YOUTH ANXIETY: PSYCHOMETRIC EVALUATION

By Cassidy C. Arnold, M.S.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2015

Major Director: Michael A. Southam-Gerow, Ph.D.,
Professor
Department of Psychology and Pediatrics

Treatment integrity—the extent to which a treatment is delivered as it was intended—has long been recognized as critically important in treatment evaluation research, but has garnered increased attention in recent years within the context of dissemination and implementation science. However, the field's development has been hindered by inadequate measurement tools. This project is focused on developing and evaluating the psychometric strength of two measures of treatment integrity. To evaluate the psychometric strength of the Cognitive-Behavioral Therapy for Youth Anxiety Therapist Adherence Scale (CBAY-A) and the Cognitive-Behavioral Therapy for Youth Anxiety Therapist Competence Scale (CBAY-C), 954 psychotherapy sessions from two treatment evaluation studies were coded. Analysis of the evidence for reliability and

validity of the item scores for each measure provide substantial support for each measure, while also highlighting areas in need of further evaluation. The discussion focuses on interpreting the psychometric strength of the CBAY-A and CBAY-C compared to other measures of treatment integrity, next steps for evaluating the psychometric strength of the two measures, and potential applications of the CBAY-A and CBAY-C.

Adherence To and Competence In Cognitive Behavioral Therapy For Youth Anxiety: Psychometric Evaluation.

The purpose of this project is to further the science of treatment integrity by developing psychometrically strong measures of therapist adherence to and competence in cognitive-behavioral therapy (CBT) for youth anxiety. This project is part of a larger Treatment Integrity Measurement Study (TIMS) funded by the National Institute of Mental Health, (RO1 MH086529). Treatment integrity—the extent to which a treatment is delivered as it was intended (Perepletchikova & Kazdin, 2005; Waltz, Addis, Koerner, & Jacobson, 1993)—is critical to: (a) treatment evaluation studies; (b) training, dissemination, and implementation projects; and (c) basic science in clinical psychology.

The aims of this study are to evaluate the reliability and validity evidence of two new measures of therapist adherence and therapist competence, the Cognitive-Behavioral Therapy for Youth Anxiety Therapist Adherence Scale (CBAY-A) and the Cognitive-Behavioral Therapy for Youth Anxiety Therapist Competence Scale (CBAY-C). Data for this project was collected via observational coding of audio and video recordings from the Individual CBT condition in one efficacy trial (Kendall et al., 2008) and one effectiveness trial that included two active treatment conditions (Southam-Gerow, Weisz, Chu, McLeod, Gordis, & Connor-Smith, 2010). Both trials included therapists administering CBT guided by the Coping Cat Manual (see Kendall & Hedtke, 2006); however, each trial used slightly different versions of the Coping Cat Manual. The effectiveness trial also included a "usual care" (UC) condition. The therapists providing the therapy in the UC condition were instructed to provide the same type and quantity of therapy that they regularly provide to youth seeking services (Southam-Gerow et al., 2010). A total of 954

treatment sessions from 89 youth and 45 therapists were coded and included in analysis for this study.

Measurement of therapist adherence (i.e., the extent to which treatment includes prescribed components and does not include proscribed components) and therapist competence (i.e., the skillfulness of treatment delivery), along with treatment differentiation (i.e., determining that a treatment is distinct from other treatment types being evaluated), has been a growing area of research in recent years and this study adds to this trend (McLeod, Southam-Gerow, & Weisz, 2009). The majority of treatment integrity studies have included one or more significant limitations. For instance, measures of adherence or competence have relied on behavioral observations from a limited number of sessions drawn from the total pool of available sessions (Barber, Mercer, Krakauer, & Calvo, 1996; Marder, 2007), adherence and competence have been coded by the same coder making evaluating the relationship between the two measures difficult (see Yale Adherence and Competence Scale; Carroll et al., 2000), or checklist-based measures have been completed by the treatment developer (see Kendall, Hudson, Gosch, Flannery-Schroeder, & Suveg, 2008; Southam-Gerow et al., 2010). Each of these issues is discussed in greater detail later in this document. This study is unique in that all available sessions were coded and included in analyses[1] and therapists' adherence and competence were independently coded by two separate teams of coders.[2]

[1] Some sessions were not coded due to technical problems at the time of the session, prohibitively poor quality of the recording itself, or because the sessions or portions of the sessions were held outside the therapy room where the audio or video recording device was placed. Further discussion of coded and not coded sessions is provided in the Method section.

[2] Due to staff turnover, some of the adherence and competence coding was completed by the same coding team though the actually coding was done at separate times.

The development of measures of adherence to and competence in delivering CBT for youth anxiety with strong reliable and valid evidence can make a meaningful contribution to multiple veins of research within clinical psychology.

Literature Review

This chapter will: (a) define treatment integrity, its components, and other relevant terms; (b) articulate the value of treatment integrity within various areas of clinical psychology; (c) discuss the methods for measuring adherence and competence and present representative efforts that use each method; and (d) introduce the two measures being evaluated in this paper. This chapter will close by presenting the specific hypotheses to be tested within this study.

Definitions. This section will focus on defining treatment integrity and the components that comprise treatment integrity as conceptualized in this project: therapist adherence, therapist competence, and treatment differentiation. Some have included treatment receipt and treatment enactment within treatment integrity. These terms will also be defined and the rationale for not including them in the conceptualization of treatment integrity used for this project will be presented. As all of these terms can be conceptually challenging, this section will include an analogy to help clarify adherence, competence, and differentiation. The section concludes with a discussion of the relationship between the different aspects of treatment integrity.

Treatment Integrity. Treatment integrity refers to the degree to which a therapist delivers a treatment consistent with a specific treatment manual or a general treatment model and in a skillful manner (Waltz et al., 1993; Southam-Gerow & McLeod, 2013). Stated more simply, was the treatment delivered in the manner in which it was intended (Perepletchikova & Kazdin, 2005)? Treatment integrity consists of therapist *adherence* to the principles and practices of the treatment, therapist *competence* in delivering the treatment, and treatment *differing* (i.e.,

treatment differentiation) from other forms of treatment being evaluated. Adherence and competence are particularly important in treatment efficacy research where the primary question is whether or not a treatment, when used with a specific population, administered as intended, and with sufficient skill, results in desirable clinical outcomes (Perepletchikova & Kazdin, 2005).

Measures of treatment differentiation are designed to assess the uniqueness of treatments within a treatment outcome study (Southam-Gerow & McLeod, 2013; Marder, 2007). Treatment differentiation is especially important when ensuring that different conditions of an intervention study (e.g., efficacy or effectiveness trials) are truly different from one another (McLeod & Weisz, 2010; Chambless & Ollendick, 2001).

Adherence. *Adherence* is defined as the extent to which a treatment includes components and approaches that are prescribed by the treatment (Waltz et al., 1993)[3]. Measures of adherence can vary in the specificity with which they measure adherence to a treatment manual. Some adherence measures (e.g., the Adherence/Competence Scale for Individual Drug Counseling for Cocaine Dependence; Barber et al., 1996) have items that are designed to measure specific practices prescribed by a treatment manual. For example, if a manual includes a focus on relaxation in a particular session, then the therapist is only considered adherent if s/he focuses the session on relaxation by using the specific practices called for in the manual. More flexible approaches to adherence (e.g., Yale Adherence and Competence Scale; Carroll et al., 2000) include items that measure the use of components and approaches that are common to the

[3] Unless otherwise noted, for the remainder of this document, the term *adherence* will always refer to therapist adherence and *competence* will always refer to therapist competence as the terms are defined in this section. Adherence, as it is used in this document, should not be confused with the concept of *treatment adherence, treatment compliance, or regimen adherence* which refers to the degree to which the patient's behaviors (e.g., taking medications, following diets, practicing behaviors or skills, and making lifestyle changes) are consistent with those prescribed to them by their mental health or medical professional (Lemanek, Kamps, & Chung, 2001).

treatment orientation for the treatment target. For example, the therapist would be given credit for being adherent whenever treatment orientation-consistent components and approaches are used to address the treatment target regardless of specific language used or sequence of delivery. The first type of measure described is more appropriate when evaluating the use of a particular treatment manual whereas the second type of measure is more appropriate when the research question addresses active and essential components of treatment or treatment manuals. The scales used to reflect therapist performance in measures of treatment adherence will be discussed later (in the Scale Type section).

Competence. Professional competence, as defined by Epstein and Hundert (2002) for medical professionals, is "the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflections in daily practice for the benefit of the individual and community served" (Epstein & Hundert, 2002, p. 226). This is consistent with other definitions of competence provided for mental health professionals (see Waltz et al., 1993; Barber, Sharpless, Klostermann, & McCarthy, 2007). One of the distinguishing characteristics of competence versus adherence is the role of context (Barber et al., 2007). That is, both concepts involve the delivery of specific treatment components but competence is context-dependent whereas adherence is context-independent. For example, a therapist attempting to lead a child through the steps of a coping plan would be considered adherent in the delivery of CBT for youth anxiety but would only be considered competent if the child's attentiveness to the conversation was promoted and appropriate examples and language were used to facilitate the child's learning.

The literature on therapist competence has focused on global competence and limited-domain competence (Barber et al., 2007; Kaslow, 2004). Global competence refers to the

skillfulness and judgment of the therapist that is independent of the particular treatment orientation and treatment objectives. Indicators of high global competence are therapist behaviors that promote a strong alliance between therapist and patient and promote high client involvement (Southam-Gerow & McLeod, 2013). Limited-domain competence refers to the therapist's skillfulness, judgment, timing, and appropriateness of intervention as she or he delivers a specific manual-guided treatment or treatment from a particular therapeutic orientation and for a particular problem (Barber et al., 2007). Aspects of limited-domain competence include timeliness of delivering treatment components, adapting the treatment to make it more relevant to the patient's life and reasons for seeking treatment, and modifying the delivery of session content to match the patient's cognitive and developmental level (Barber et al., 2007). Modifying the content based on the patient's cognitive and developmental level is especially relevant when the treatment is being delivered to youth. For this project, competence is defined as the skillfulness and responsiveness with which the therapist delivers treatment components consistent with CBT for youth anxiety (Southam-Gerow & McLeod, 2013). This definition is closely aligned with Barber's (2007) definition of limited-domain competence.

One of the challenges in measuring competence is the lack of data linking moment-by-moment therapist behaviors to treatment outcome (Barber et al., 2007). That is, part of what is assumed when discussing competent delivery of an intervention is that more competent delivery will be associated with better outcomes. Unfortunately, the empirical link between competence and outcome is limited.

As depicted by Marder (2007; see figure 1), therapist adherence and competence are inherently related to one another in that adherence to a treatment is a prerequisite of competence, while competence is not guaranteed with adherence. Consequently, measures of adherence and

competence are frequently highly correlated when both have been measured in the same study (e.g., Adherence/Competence Scale for Individual Drug Counseling for Cocaine Dependence, Barber et al., 1996; MATCH Tape Rating Scale, Carroll et al., 1998; Yale Adherence and Competence Scale, Carroll et al., 2000; Therapist Behavior Rating Scale – Competence, Hogue et al., 2008).

It is important to discuss the nature of adherence and competence and measures of each to clearly outline what is being measured and how it should be conceptualized. Depending on the specific focus of the adherence and competence measure, the item(s) evaluate either a specific behavior that corresponds to a practice described in a treatment manual, a practice element that is common to many different treatment manuals, and/or behaviors that are consistent with a theoretical orientation. However, regardless of this focus, adherence and competence measures are intended to measure behaviors exhibited by the therapist, and conclusions drawn from these measures should focus on those behaviors and not an underlying characteristic of the therapist. In this way, measures of adherence and competence are not measures of an underlying construct in the way that Classical Test Theory (CTT) defines a construct (i.e., characteristics, attributes, or traits that cannot be directly measured but are believed to exist; Cronbach & Meehl, 1955), and the measures are not consistent with the assumptions of CTT (e.g., the items of the measure are all influenced by variance due to the latent variable and error variance; Bollen & Lennox, 1991; Borsboom, Mellenbergh, & Van Heerden, 2003). For measures consistent with CTT, the items on a measure covary with one another *because* they are all mutually influenced by the latent variable or construct (Bollen & Lennox, 1991; Borsboom et al., 2003; DeVellis, 2003).

The CBAY-A and CBAY-C measure various therapist behaviors and it is the quantity and quality of those behaviors that determine the adherence and competence of the therapist in the observed session. Items *may* covary but it is not an assumption that they *will* covary. In fact, for an individual session, it is not possible for a therapist to receive high adherence ratings for every item, though it is possible for them to receive low adherence ratings for every item. In this sense, there is an opportunity cost for delivering a high dose of one treatment element (e.g., relaxation) in that there is less of an opportunity to deliver a high dose of other treatment elements, as there is only so much time in the session.

Treatment differentiation. Treatment differentiation is the process of characterizing a treatment and ensuring that treatments that are intended to be different are in fact distinct from one another in important ways (Perepletchikova & Kazdin, 2005; Marder, 2007). When investigating a treatment from a specific therapeutic orientation, treatment differentiation measures must be sensitive to the inclusion of proscribed treatment components (Waltz et al., 1993).

Two general methodologies can be used in treatment differentiation. First, in a comparative treatment study evaluating the outcome of treatment A versus treatment B, treatment differentiation can be accomplished by using an adherence measure designed for treatment A with sessions from treatment B and vice versa. If the ratings for treatment A and treatment B, when evaluated by an adherence scale developed for opposite treatment, are sufficiently low, then it can be said that the treatments are distinct. This approach was used in the Treatment of Depression Collaborative Research Program (TDCRP; Hill et al., 1992).

The second method involves using a measure that simultaneously assesses for the presence of treatment components and approaches from a variety of different treatment

orientations. The Treatment Process Observational Coding System-Strategies (TPOCS-RS, McLeod & Weisz, 2010) is an example of a measure that assesses broad treatment strategies from commonly used treatment orientations (i.e., behavioral therapy, cognitive therapy, psychodynamic therapy, family therapy, and client-centered therapy). Although the relative strength of the two methods has not been evaluated empirically, the first method has the potential to be more sensitive to issues of treatment diffusion while the second method may be more appropriate for characterizing treatments of an unknown type such as the treatment provided in the UC condition in a treatment efficacy trial.

Other aspects of treatment integrity. Other literature has included treatment receipt, treatment enactment, child involvement, and therapeutic alliance as additional components of treatment integrity (Jones, Clarke, & Power, 2008; McLeod, Southam-Gerow, Tully, Rodríguez, & Smith, 2013). Treatment receipt refers to the extent to which the patient understands the factual knowledge presented in treatment and is capable of implementing the skills taught in the treatment. Treatment enactment refers to how much the patient actually utilizes the treatment in his or her life (Marder, 2007; Lichstein, Riedel, & Grieve, 1994). Child involvement has been defined to include aspects of attitude (e.g., willingness to participate), positive behavior (e.g., asking questions, initiating responses), and the absence of negative behavior (e.g., avoidance; Chu & Kendall, 2004; Jones et al., 2008). The alliance between child and therapist has been hypothesized to include aspects related to their bond (referring to emotional aspects of the child-therapist relationship) and the task of therapy (referring to the child's level of participation in therapy activities; McLeod & Weisz, 2005; Liber et al., 2010). While they may be potentially valuable aspects of treatment integrity to measure in future work, these components were not measured or included in the current study.

An illustrative analogy. When discussing treatment evaluation research and defining treatment integrity and its components (adherence, competence, and differentiation) it can be helpful to use the analogy of baking cookies and taste-testing the cookies. Imagine a taste-test (comparing the effectiveness of two treatments) that compares a chocolate chip cookie recipe versus an oatmeal raisin cookie recipe. The first test of treatment integrity, or in this case baking integrity, is to determine if the two recipes are distinct. Likely they both include flour and sugar (components that Waltz and colleagues called *essential but not unique*; 1993), but do the oatmeal raisin cookie bakers also include chocolate chips or do the chocolate chip cookie bakers include oatmeal or raisins (*unique and essential* ingredients; Waltz et al., 1993)? If they do, then the distinction between the cookies, regardless of what the recipe calls for, is diminished.

Measurement of adherence evaluates the degree to which the recipe is followed. Do the chocolate chip cookie bakers include the prescribed quantity of chocolate chips and do the oatmeal raisin cookie bakers include the called-for quantities of raisins and oatmeal? If, for example, the chocolate chip cookie baker did not include chocolate chips, then the taste test could not reasonable be said to compare chocolate chip cookies to oatmeal raisin cookies. Finally, the skillfulness or competence of the bakers is evaluated. Do the bakers mix the ingredients skillfully to ensure a thorough mixture without spilling important ingredients? Do the bakers adjust the cooking time called for in their recipe to account for differences in oven temperature? If the cookies are not sufficiently distinct, if the bakers do not bake the cookies as called for by the recipe, and if the bakers are not sufficiently and comparably competent, then the results of the taste test should be interpreted with caution.

Marder (2007) represented the theoretical relationships between the components of treatment integrity by placing each at a different level of a pyramid (see Figure 1) with

differentiation at the base, adherence in the middle, and competence at the top, implying that the aspects of treatment integrity closer to the base measure broader and prerequisite components as those closer to the top of the pyramid. The following sections address the relationships between the different components within treatment integrity.

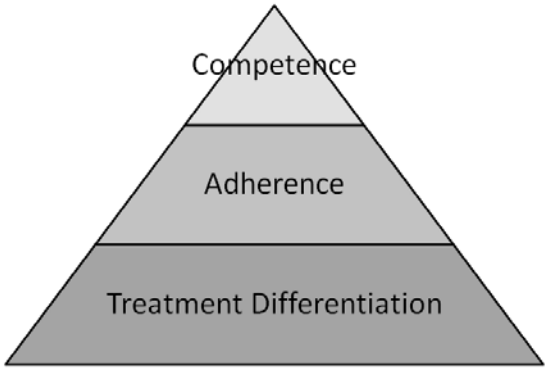

Figure 1. Treatment Integrity Components and Relationships

Adherence and differentiation. Adherence is similar to differentiation in that both evaluate the presence or absence of various interventions and treatment approaches. However, measures of differentiation must measure interventions and approaches used by *other* treatments or treatment orientations in order to be sensitive to the presence of proscribed interventions and approaches. Measures of adherence, on the other hand, are solely focused on reflecting the extent to which the treatment provided is consistent with the treatment approach or includes the treatment elements that were intended to be delivered.

Treatment differentiation is at the bottom of the pyramid because it measures treatment integrity at its broadest level and is not dependent on adherence or competence. Adherence is above differentiation in the pyramid because it assesses a particular treatment orientation in detail whereas differentiation assesses for the presence of a variety of treatment strategies from a variety of therapeutic orientations. It is expected that scores on an adherence measure for a particular treatment would be positively correlated with the subscale on a measure of treatment

differentiation that corresponds to the treatment orientation on which the treatment is based. Conversely, it is expected that scores on a measure of adherence would be minimally correlated with the other subscales on a measure of differentiation.

Adherence and Competence. As discussed earlier, one of the critical distinctions between measures of adherence and measures of competence is the role of context. Specifically, adherence is context-independent behavior whereas competence is context-dependent. For example, a therapist can receive high scores on adherence by mechanically reading a script from a therapy manual regardless of the patient's behaviors, readiness to participate in treatment, and understanding of the material. Such rote behavior would unlikely be coded as highly competent unless the patient is engaged in and responding positively to the treatment. This example illustrates the potential independence of adherence and competence. Their independence is also supported by empirical evaluations that have found the relationship between the adherence and competence to be moderate in magnitude (Miller & Binder, 2002). Some work has even described an inverse relationship between adherence and competence (Miller & Binder, 2002; Barber, Gallop, Crits-Christoph, Frank, Thase, Weiss, & Connolly Gibbons, 2006). An inverse relationship between adherence and competence is possible at very high levels of adherence if the therapist is delivering the treatment in a rote manner and not being responsive to the patient.

The conceptual model of treatment integrity illustrated by Figure 1, with adherence as a prerequisite of competence, suggests that measures of adherence and competence should be positively related. Evaluations of adherence and competence of the same psychotherapy sessions (as conducted in this study) will contribute to our understanding of the relationship between these aspects of treatment integrity.

Importance of treatment integrity research. This section will briefly describe the increasing importance of, and focus on, treatment integrity in clinical research. Following this discussion of historical context, the value of treatment integrity research will be discussed with respect to (a) treatment evaluation studies; (b) training, dissemination, and implementation projects; and (c) basic science in clinical psychology.

Developments in clinical psychology and the rising importance of treatment integrity. A number of trends over the past 30 years have led to a growing interest in treatment integrity research and the need for psychometrically strong measures of treatment integrity (Schoenwald, Henggeler, Brondino, & Rowland, 2000). These trends include (a) efforts to characterize the evidence-base supporting a particular treatment approach relative to other treatment approaches, (b) the development of manual-based treatment protocols, and (c) the scientific study of training, dissemination, and implementation.

One of the fundamental challenges of treatment research, whether the treatment is psychosocial, pharmacological, or of another modality, is demonstrating that the research has sufficient internal validity (Kazdin, 2003). Treatment evaluation researchers need to show that the treatment provided in their research to be of a known type and quality in order to make meaningful comments about the treatment's efficacy or effectiveness. These efforts are frequently described as manipulation checks. To say that a particular treatment leads to symptom reduction, one must first know that the treatment being evaluated is a single type of treatment, such as cognitive behavioral treatment or interpersonal psychotherapy, and that other treatment types or components are not included in the treatment. In fact, the American Psychological Association's (APA) Division 12 Task Force for the Development of Evidence Based-Treatments has specified that a treatment outcome study may only contribute to a

treatment's evidence base if it includes a *well-defined treatment* (Chambless & Ollendick, 2001). Furthermore, to guard against threats to internal validity in comparative treatment outcome studies, researchers must confirm that the two treatments are distinct (Barber et al., 1996; Carroll et al., 1998).

One common step taken to bolster the evidence supporting the internal validity of treatment evaluation studies is the development and use of treatment manuals that articulate the procedures and components that make up a given treatment. Though developed to help address internal validity concerns, and make treatment replicable (American Psychological Association, 1993), treatment manuals have also facilitated the development of treatment integrity checks. Treatment manuals have been particularly helpful in the study of treatment differentiation, by specifying the practices that should be included in a given treatment. This allows researchers to identify treatment practices that should not be in a given treatment (Hill et al., 1992). Manual-guided treatments heralded more sophisticated efforts to quantify the effect of a given type of psychotherapy (e.g., reduced variance in the treatment provided could lead to more confident reports on the effect size of the treatment) and comparative treatment research (Mowbray, Holter, Teague, & Bybee, 2003). Although it is generally assumed that manual-guided treatment protocols improve adherence, by no means does their use guarantee adherence without adequate training (Miller & Binder, 2002).

One of the primary criticisms of manual-guided treatments is that they curtail the use of therapeutic skill and flexibility (Kendall, Chu, Gifford, Hayes, & Nauta, 1997; Addis & Krasnow, 2000). To address this criticism, some treatment developers have encouraged therapists to flexibly and creatively use the manuals as a guide to treatment delivery, rather than being rigidly constrained in treatment delivery (Kendall et al., 1997). Essentially, the call for

flexible delivery of manual-guided treatments is recognition that the context within which a treatment is delivered and the therapist's skillfulness in delivering the treatment is important. On its surface, such a recommendation would seem to increase the variability with which a treatment is administered. However, that is an empirical question that can be evaluated with treatment integrity research conducted with psychometrically strong measures of treatment adherence and competence.

Interpreting treatment outcome studies and evaluating treatment components. The cornerstone of treatment evaluation research in clinical psychology and other fields—the randomized clinical trial—is based on the principle that an intervention with known components, dosage, purity, and quality is delivered to equivalent groups of participants and resulting differences in the groups following the intervention are due to the intervention itself (Hogue et al., 1996). When treatment studies mitigate the threats to internal validity (e.g., maturation, regression to the mean, history, etc.) and control for alternative explanations for changes in the dependent variable (e.g., symptom severity, presence or absence of diagnosis, and functional impairment), then the independent variable can reasonably be assumed to have caused this change (Kazdin, 2003). In this context, measurement of treatment integrity refers to verification that the independent variable (i.e., the treatment) was manipulated as intended (Mowbray et al., 2003). This model of treatment evaluation research is similar to that used in medical and pharmaceutical research.

A major challenge faced in psychotherapy research is defining the specific characteristics of the independent variable (i.e., the treatment). In light of this challenge, few psychotherapy treatment outcome studies for children and adolescents report any form of an adherence check. In a review of treatment evaluation studies targeting youth psychopathology, only 32.2% of trials

used supervision or an adherence check to ensure that the treatment was being delivered to a sufficient degree (Weisz, Jensen-Doss, & Hawley, 2005). Though treatment integrity is still relevant in medical and pharmaceutical research (see Pribluda et al., 2012), medication, which is consistently manufactured and produced, is more consistent across patients in terms of dosage and quality. Psychotherapy treatments, on the other hand, may include multiple treatment elements and the dosage delivered to the patients is more difficult to measure.

These issues—unknown dosage and multiple treatment elements delivered in variable ways—threaten the internal validity and construct validity of psychotherapy treatment outcome research. Internal validity refers to the extent to which changes in the dependent variable can be attributed to the independent variable or intervention rather than other variables and influences (Kazdin, 2003). Construct validity refers to confidence with which the meaningful aspects of the intervention are what they are claimed to be in intervention description (Kazdin, 2003). For example, is the cause of participant behavior change the elements of the treatment or did a non-treatment element of the study (e.g., expectations of the experimenter) lead to the participants' behavior change (Kazdin, 2003)? When a treatment study detects or fails to detect an effect for a given treatment, statements about the efficacy of the treatment rest on the degree to which the treatment actually provided to the participants is the same as the treatment which was intended to be evaluated.

If there is no symptom improvement within a randomized controlled trial, one cannot reasonable interpret that this is the result of an inefficacious treatment *if* the treatment was not administered as it was intended to be administered or if it was administered incompetently. These additional variables may reasonably explain the lack of symptom improvement. Alternatively, a treatment cannot be considered efficacious if the intervention that was actually

administered included treatment components from treatments other than the one being evaluated. These issues further highlight the need for treatment integrity measurement to be incorporated in to psychotherapy randomized controlled trials.

There have been a number of puzzling results that have come out of the psychotherapy treatment literature. For example, the use of CBT for youth anxiety is considered a *probably efficacious*[4] treatment for a variety of anxiety disorders (e.g., social phobia; Silverman, Pina, & Viswesvaran, 2008), though multiple studies have failed to show that CBT for youth anxiety is more effective than usual care (e.g., Barrington, Prior, Richardson, & Allen, 2005; Southam-Gerow et al., 2010). One alternative explanation for these finding may be that the independent variables (the treatments) in these studies were not adequately controlled. Without verification that the treatment delivered included the prescribed components (adherence), was delivered skillfully (competence), and did not include proscribed components (differentiation) it is impossible to say what null or negative findings mean (Mowbray, et al., 2003; Southam-Gerow & McLeod, 2013). Further, it is not known to what extent the control treatments overlapped with the experimental treatment (differentiation), making the interpretation of null/negative findings even more complicated. Developing psychometrically strong measures of treatment integrity will help researchers account for variations in the quantity (measured by adherence scales) and quality (measured by competence scales) of the treatment provided in these treatment evaluation studies. This may help to clarify the reasons for these mixed findings (Mowbray, et al., 2003).

As it relates to evaluating the potency of treatment components, therapist adherence and competence measures that are sensitive to general components consistent with a treatment

[4] See Chambless & Ollendick (2001) for a description of the criteria for categorizing treatments as Well-established treatments, Probably efficacious treatments, and Experimental treatments.

approach (rather than tightly defined behaviors such as a specific relaxation script) have greater potential to benefit treatment development research than do narrower measures (e.g., tied to a specific treatment manual). When measures are strictly tied to a particular treatment manual, the potential implications for the larger field of treatment development and treatment evaluation are limited. Findings will only apply to the specific treatment manual. There is also evidence that manual-guided treatments provided in community settings (a) include treatment components from a variety of theoretical orientations and (b) may be administered with low adherence (McLeod & Weisz, 2010). Therefore, adherence measures that are too closely tied to a particular treatment manual (e.g., Coping Cat for Youth Anxiety; Kendall & Hedtke, 2006a, 2006b) may not detect treatment practices that differ from prescribed instructions in the manual, but do adequately reflect the approach of a particular orientation (e.g., cognitive behavioral therapy). When treatment components consistent with a particular theoretical orientation are evaluated, the results can be more broadly applied.

In summary, measures of treatment integrity are critical to treatment evaluation research in two major ways. First, treatment outcome research is based on the premise that the independent variable is well defined. That is, the treatment being tested includes specified characteristics of known quantity and quality. Treatment manuals are an important step but are not enough—the quantity and quality of the treatment that is actually provided should be measured. Only by evaluating the *actual* treatment provided in research studies through the use of psychometrically sound treatment integrity measures can researchers say with a high degree of confidence that the treatment provided was what it claimed to be. Second, it is important to measure treatment integrity with greater sophistication than provided by a dichotomous scale in order to make more meaningful interpretations of the results of the treatment outcome studies.

This is particularly important given the mixed support for the effectiveness of CBT for youth anxiety.

Treatment integrity in training, dissemination, and implementation work. Treatment integrity research has important implications for training, dissemination, and implementation work. This section will address the ways in which psychometrically strong measures of adherence to and competence in CBT for youth anxiety will be beneficial with respect to training clinicians and disseminating and implementing CBT for youth anxiety in community service settings.

The tasks of improving psychotherapy training and evaluating dissemination and implementation efforts each call for teaching specific psychotherapy practices and supporting therapists as they learn and deliver these skills. However, despite the proliferation and widespread adoption of evidence-based treatments (EBTs) in research settings, research on evidence-based training and supervision practices is still a relatively new endeavor (Miller & Binder, 2002). One method for evaluating the effectiveness of training and supervision efforts is to measure the degree of treatment integrity with which the therapists are practicing. The premise of psychotherapy training (e.g., degree program, continuing education, etc.) is that the quality of the therapy delivered by the therapists will improve as a result of the training. This is an empirical question, but without psychometrically strong measures of treatment adherence and competence, this outcome cannot be adequately evaluated. With data from treatment integrity measures, the effects of various training programs can be directly compared. Additionally, the dose and quality of therapy provided in community mental health clinics can be quantified prior to and following efforts to disseminate a new treatment intervention. The field of implementation science would benefit from measures of treatment integrity that could indicate

when therapists are delivering psychotherapy services below a minimum standard of adherence and competence. Such information would allow for additional training on specific skills as needed.

Adherence and competence in practice. There are real-world implications for measures of therapist adherence and competence. For instance, psychotherapy training centers, third-party payers, and community mental health center administrators all have an interest in knowing what type of treatment is being provided by therapists, and therapists' level of adherence and competence to CBT (Hayes, Barlow, & Nelson, 1999; Hogue et al., 2008).

Psychometrically strong measures of adherence to and competence in psychotherapy modalities with demonstrated efficacy could be used by a variety of stakeholders interested in ensuring that therapists are delivering high quality psychotherapy services. Specifically, graduate school training programs likely want to know that graduating therapists meet a certain standard. Similarly, third-party payers for psychological services, such as insurance companies and state and local governments, want to know that the psychotherapy they are paying for is of sufficient quality to promote symptom reduction (Mowbray et al., 2003). Researchers interested in the science of dissemination and implementation science—moving evidence-based practices from university-based research settings to community clinics—would also benefit from these measures as they measure the effects (therapists' behaviors) of their dissemination and implementation interventions (McLeod et al., 2013).

Benchmarking refers to a method of comparing a practice with questionable quality to a "gold-standard" practice along some metric (Weersing & Weisz, 2002). This metric can be an outcome, such as level of symptom reduction, or a practice, such as adherence to or competence in a manual-guided treatment (McLeod et al., 2013). Using psychometrically strong measures of

adherence and competence, benchmarks could be used by: (a) graduate schools and psychotherapy training programs (e.g., those providing continuing education credits) to characterize therapists' level of adherence and competence, their need for additional training, and readiness for practice; (b) third-party payers to identify which therapists provide psychotherapy of a sufficient quality that merits payment; (c) dissemination and implementation researchers to describe the level of intervention required to obtain a particular quality of service; (d) administrators in community mental health centers to identify which therapists need increased supervision or remedial training in order to meet a particular quality of service and which therapists can practice with greater independence. Intervention drift—changes in practice over time—is a concern not only for researchers but also for those interested in ensuring treatment quality in community settings (Mowbray et al., 2003).

Basic research. At the level of basic research, the development of gold-standard measures of therapist adherence to CBT for youth anxiety and therapist competence in delivering CBT for youth anxiety will prove a valuable tool for basic science in clinical psychology. This section will briefly describe two areas of basic science that will be furthered by the development of high quality measures of adherence to and competence in CBT for youth anxiety: measure development and comparative treatment research.

These gold-standard measures of adherence and competence will provide a yardstick against which other methods for measuring adherence and competence can be compared. By developing observational coding methods for evaluating adherence and competence, researchers will be able to compare behavioral observation measures with therapist self-reported adherence and competence (e.g., Schoenwald, Carter, Chapman, & Sheidow, 2008) to see whether self-reported measures can produce meaningful data. A critique of therapist self-report measures is

that the therapists delivering interventions are too biased to provide this information (Mowbray et al., 2003). Using therapist self-report data represents a significantly more cost effective method of evaluating treatment integrity than observational coding, though these savings are only worthwhile if the data collected is sufficiently accurate (Carroll et al., 1998).

Measures of treatment integrity have a great potential for furthering comparative treatment research. In this discussion, it is worth briefly noting what has come to be known as the Dodo Bird Verdict and the basic arguments on each side (Luborsky et al., 2002; Chambless, 2002; Budd & Hughes, 2009). The Dodo Bird Verdict claims that all treatments are equivalent and, therefore, all are appropriate treatment approaches for any given disorder. To refute this conclusion, one must simply show that, for a given problem, one or more treatments are superior to other treatments (Chambless, 2002). On the one hand, the comparative treatment literature (particularly meta-analyses of head-to-head treatment trials) may be interpreted to indicate that the differential effects of different treatments are relatively small and mostly insignificant (e.g., Luborsky et al., 2002; Budd & Hughes, 2009). On the other hand, there is concern that reliance on meta-analyses misses real differences between some treatments and that the Dodo Bird Verdict is inappropriate for several reasons. Meta-analysis is insufficient in detecting patient characteristic (e.g., diagnosis) by treatment interactions that form the cornerstone of the evidence-based treatment literature (see Chambless & Ollendick, 2001). Further, and most germane to this study, meta-analytic approaches fail to account for the *ingredients in the therapies* (i.e., the treatment integrity) being evaluated (Chambless, 2002; Budd & Hughes, 2009). Some have speculated that low adherence and competence and high treatment overlap are major causes of null and negative findings in comparative treatment research (Mowbray, et al.,

2003). Measures of adherence and competence can facilitate the evaluation of the treatment integrity in randomized clinical trials.

Treatment integrity research also can play a critical role in the treatment utility of assessment research. Treatment utility of assessment refers to the degree to which psychological assessment instruments and practices contribute to beneficial treatment outcomes (Hayes, Nelson, & Jarrett, 1987). The challenge of research on the treatment utility of assessment is that it requires the accurate and detailed measure of treatment components that were actually delivered to individual patients rather than broad characterizations of the treatment. Researchers need to demonstrate that there is an interaction between treatment components and patient characteristics influencing outcome (Carroll et al., 1998). Measurement of adherence on a continuous scale, rather than simple presence or absence, allows for greater sensitivity to interactions between patient characteristics and treatment components (Carroll et al., 1998). Therefore measures should include continuous, rather than dichotomous, items of treatment adherence.

Methods for measuring therapist adherence and competence. Over the past three decades, efforts have been made to measure adherence and competence using a variety of different methods. These methods, along with their strengths and weaknesses, are discussed in the following section. Particular attention is given to *type of data* (e.g., observational coding, therapist-report, or informant-report), *focus of measurement* (i.e., adherence to principles underlying the therapy versus adherence to specific practices called for by the treatment manual), *type of scale* used (i.e., dichotomous scale, frequency of intervention, multi-point Likert-type scale), and *sampling plan* (portion of session and proportion of sessions).

Type and source of data. Potential types of data include observational coding, self-report from the therapist, informant report, record review, and checklists. Potential sources include coders of different levels of expertise (e.g., undergraduate students, graduate students, or doctoral-level) and different informants (e.g., youth or parent).

Observational coding and self-report measures represent the two most common methods for measuring treatment integrity. Observational coding, like observational assessment in clinical contexts, has some distinct advantages over measures that rely on therapist-report, patient-report, supervisor-report, or chart review. Observational coding is not subject to the same biases that threaten the validity of self- or other-reported measures, though coder bias cannot be ruled out (McLeod, Islam, & Wheat, 2013). Specifically, there is concern that the informant reports what they believe the researcher wants to hear (social desirability; Mowbray et al., 2003). With therapist-report in particular, there is concern that the data collected on the therapist's behaviors reflects what the therapist intended to do or thinks that he or she should have done rather than what he or she actually did. In fact, Carroll and colleagues (1998) found that therapists were more likely to endorse administering a treatment component than was observed when the same session was coded using an observational coding system. These same concerns hold for supervisor-report. Supervisor-report also is suspect because the supervisor may not have seen or directly observed the therapist's actual behavior and may be reporting what the therapist told them happened during supervision sessions or based on clinical notes. Finally, patient-report is limited by the patient's knowledge of what was intended to take place in the therapy session and their understanding of psychological terms and practices (McLeod et al., in press). Data from patients may be inappropriately skewed by their overall perception of the relationship with their therapist (McLeod et al., in press). Despite these drawbacks, data

collected via therapist-report, patient-report, and supervisor-report has the distinct advantage of being relatively inexpensive to collect, making it more feasible to collect data from every therapy session and every dyad within a treatment outcome study (Carroll et al., 1998). A major benefit of developing observational coding measures of treatment integrity may be to evaluate self-reported measure of adherence and competence (the less costly and more time efficient methods of assessing treatment integrity), for future use.

One of the challenges of measuring treatment integrity is the complexity of the behaviors being measured (Waltz et al., 1993). In order to accurately measure adherence or competence, the individual making the ratings must not only have an understanding of the treatment manual being used, but they also must have strong clinical training so that they can recognize when the therapist is providing a particular intervention in a novel way. For example, therapist flexibility (i.e., altering a treatment to make the treatment more applicable to the patient) is a component of competence. To an insufficiently-trained coder, appropriate alterations may go unrecognized and result in lower ratings even if the essential treatment component was present in a novel form. Additionally, definitions for some of the adherence and competence items may be too complex to be fully understood and captured with informant-report methods when the informant is not highly trained.

Examples of observational coding measures include the Collaborative Study Psychotherapy Rating Scale (CSPRS; Hill et al., 1992; selected measures are discussed in further detail later in this chapter) and the Yale Adherence and Competence Scale (YACS; Carroll et al., 2000). The original version of the Therapist Adherence Measure (TAM), used to evaluate adherence to Multisystemic Family Therapy (MST), had versions for the therapist, youth, and caregiver to complete (Schoenwald et al., 2000). A revision of the TAM, the TAM-R, only

included a caregiver-report version, as caregiver-report was most strongly associated with clinical outcome (Schoenwald et al., 2008). Finally, therapist checklists have been used in a variety of studies (e.g., Carroll et al., 1998; Kendall et al., 2008; Southam-Gerow et al., 2010).

Observational coding systems vary in the level of expertise expected of the coders. For the observational coding systems reviewed for this project, the expertise of the coders ranged from undergraduate-level research assistants to graduate students to experts in the field (i.e., individuals with many years of experience administering the type of therapy being evaluated for the population being treated). For example, the CSPRS (Hill et al., 1992) used advanced graduate students in clinical and counseling psychology, whereas the Adherence/Competence Scale for Individual Drug Counseling for Cocaine Dependence (Barber et al, 1996) employed expert therapists with multiple years of experience administering individual drug counseling. Barber and colleagues (1996) deemed this higher level of experience necessary because of the focus on competence as well as adherence.

Focus of measurement. As mentioned earlier, the options for measures of therapist adherence are to tie the measure to either: (a) a particular manual-guided treatment or (b) general practices that target a specific treatment goal (e.g., anxiety) consistent with a treatment orientation (McLeod et al., 2013). These approaches have been described as *molar* (focus on practices from a specific treatment manual; "Was the manual followed?") and *molecular* (focus on practice components as generic ingredients common to many manual-guided treatments from a particular therapeutic orientation; "What practice elements were used?"; McLeod et al., 2013). Similarly, the options for therapist competence are either limited-domain competence or global-competence.

With regard to measuring global competence versus limited-domain competence, the majority of published competence measures have focused on limited-domain competence (Barber et al., 2007). When discussing treatment integrity, limited-domain competence is more applicable because it specifically addresses the issue of treatment integrity: does the therapist administer the specific therapy that they are intended to administer in a skillful manner? This is not to suggest that global competence, or the development of measures of global competence, is not important. There has been a wealth of research on common factors of therapy (e.g., empathy and warmth, therapeutic alliance, and positive regard; see Asay & Lambert, 1999; Brown, 2011). Some claim that these factors are the necessary and sufficient ingredients of therapy and are what account for therapeutic change (Asay & Lambert, 1999), while others claim that they are important but are better viewed as prerequisites to the active components of psychotherapy found in manual-guided treatments (Chambless, 2002). Since limited-domain competence is the more applicable form of competence within this discussion of measures of treatment integrity, for the remainder of this document, the term competence will refer to limited-domain competence unless otherwise specified.

Scale type. Different measures of adherence use different scale types with different levels of complexity. The most common examples include dichotomous scales, frequency scales, and extensiveness scales with each of these representing progressively more detailed ways of representing the therapists' adherence. Adherence measures with dichotomous scales essentially ask whether or not the therapist administered the treatment. These scales are insensitive to varying degrees of adherence. Dichotomous scales have been used primarily in checklist-type adherence scales such as the Cognitive-Behavioral Checklist (Carroll et al., 1998) and unpublished versions such as those used in treatment trials (Kendall et al., 2008; Southam-

Gerow et al., 2010). These measures provide a rudimentary level of detail that may support broad claims that therapists were adherent to the treatment protocol but they do not allow for analysis of different levels of adherence and the influence on treatment outcome.

Another form of adherence data is frequency counts or Likert- type scale ratings. Measures that assess the frequency with which adherence treatment practices or strategies take place provide a greater level of detail than the dichotomous scales. In theory, frequency counts could reflect the number of verbal exchanges that take place addressing a treatment component, the amount of time spent discussing a component, or a time-sampling technique where in a rater records the number of time segments (e.g., 1, 5, or 10 minute segments) in which a component is addressed. This type of data would reflect the dose of the therapy provided to the patient, though it would not capture the intensity of the dose or the extent to which each aspect within a component is covered. Frequency measures of adherence commonly use a 7-point, Likert-type scale with higher numbers corresponding to greater frequency rather than a count of adherence behaviors. No existing measures of adherence included frequency counts, though the Adherence/Competence Scale for Individual Drug Counseling for Cocaine Dependence (Barber et al., 1996) reports frequency ratings on a 7-point Likert-type scale.

The final and most common type of scale used to report therapist adherence and the only type of scale used to report competence is a Likert-type scale (typically with a 7-point Likert-type scale though others have also been used). With regard to adherence, the Likert-type scale usually reflects *extensiveness* (e.g., Hogue et al., 1998; Carroll et al., 2000; Marder, 2007). Extensiveness has been defined as the inexact combination of frequency of the intervention plus the thoroughness with which the interventions were covered (Hogue et al., 1998; Carroll et al., 2000; Marder, 2007). Adherence scales that use extensiveness ratings also provide a greater

level of detail than dichotomous scales. Rather than just reflecting frequency, extensiveness ratings are a combination of frequency and thoroughness of the adherence treatment practices and strategies used by the therapist (e.g., Therapist Behavior Rating Scale, Hogue Liddle, Rowe, Turner, Dakof, & LaPann, 1998). The CBAY-A coding manual states:

> … a therapist might be scored highly for extensiveness based on a thorough CBT intervention for child anxiety that occurs during a brief segment of a session. Conversely, a high extensiveness score may be given when a therapist uses a CBT intervention for child anxiety frequently, but not thoroughly. But, the highest marks ("7's") are reserved for CBT interventions for child anxiety that are both thoroughly executed and frequently employed within a session. (Southam-Gerow, McLeod, Arnold & Rodríguez, Unpublished Manual, p. 5)

Extensiveness ratings are also typically recorded on a Likert-type rating scale.

Therapists who spend more time on an intervention will receive higher frequency ratings and therefore higher extensiveness ratings. Similarly, therapists who more thoroughly cover various aspects of the intervention and deliver the intervention in a variety of ways (e.g., didactic teaching and rehearsal) will receive higher extensiveness ratings.

Sampling plan. An additional decision point for measures of treatment integrity involves the sampling of therapist behaviors. Researchers must select the portion of youth-therapist dyads within a treatment trial, the portion of sessions per course of treatment, and the portion of the session itself from which to draw data. To date, there is not empirical data suggesting that any one method is more appropriate than the others. Despite the lack of empirical support of particular sampling plans, and because of the high costs of observational coding, the

overwhelming majority of therapy process research involving observational coding procedures has involved coding only a limited sample (Carroll et al., 1998).

In creating a sampling plan, issues to consider include cost of data collection, burden of data collection, and the theory underlying the assessment. Clearly, gathering more data is more costly and more burdensome. This is particularly true for observational coding measures, as it requires: (a) the therapist to record each session and, (b) time dedicated to coding the session. A more conceptually challenging issue relates to how different sampling plans would influence the data obtained.

The majority of treatment integrity studies have collected data from a random sample of sessions or specific sessions from the early, middle, and end of treatment. Such sampling plans are more easily justified, despite the lack of empirical support of the decision, if the assessment is believed to measure a behavior or construct that is consistent across the course of treatment. Sessions within the course of treatment can be thought of as different settings for which different behaviors are called for. Consequently, by sampling specific sessions from a treatment, researchers are running the risk of over-detecting specific behaviors while under-detecting other behaviors. Alternatively, ratings of global competence may be less sensitive to time in treatment than limited-domain competence and, thus, may be less sensitive to sampling error.

The adherence and competence measures developed for and evaluated in this project include items addressing specific CBT for youth anxiety that are not believed to be equally and randomly distributed across sessions within a dyad. For example, within Coping Cat, Relaxation is only prescribed to be delivered in two sessions. Sampling plans that do not include every session would underestimate the rate at which Relaxation and other treatment elements took place.

Since the aim of this study is to develop and evaluate gold-standard measures of adherence to and competence in CBT for youth anxiety, methods were selected with a focus on collecting high-quality data (with less focus on reducing costs and burden). A full description of the coding methodology and decisions is presented at the end of this chapter and in the Method section.

Examples of adherence and competence measures. The CBAY-A and CBAY-C are not the first measures of adherence or competence. The following section, along with Table 1, will identify some of the more important and influential measures of adherence or competence. Strengths and weaknesses of these measures are discussed, including how CBAY-A and CBAY-C addresses the weaknesses found in previous measures. Additionally, other efforts to ensure treatment integrity in randomized clinical trials will be discussed.

Collaborative Study Psychotherapy Rating Scale. The first widely used measure of treatment integrity was the Collaborative Study Psychotherapy Rating Scale (CSPRS) that was initially used as part of the Treatment of Depression Collaborative Research Program (TDCRP; Hill et al., 1992). The CSPRS is a 96-item, observational coding measure of therapist behaviors that was used to differentiate between three forms of psychotherapy (Cognitive Behavioral Therapy, Interpersonal Therapy, and Clinical Management) used in the TDCRP. Pairs of trained raters independently coded recordings of psychotherapy sessions and provided ratings on a 7-point Likert-type scale with one indicating that the practice did not take place and ratings two through seven indicating greater levels of the behavioral presence with four representing an average level of the behavior (Hill et al., 1992). As a result, scores on the CSPRS represent the frequency of various therapist activities with higher scores indicating greater frequency. CSPRS items were organized into seven scales and many subscales: (1) *Cognitive-Behavioral Therapy*

Scale (six subscales and 28 items); (2) *Tangential Cognitive-Behavioral Scale* (two subscales and four items); (3) *Interpersonal Therapy* (seven subscales and 28 items); (4) *Tangential Interpersonal Therapy Scale* (four items); (5) *Clinical Management Scale* (five subscales and 20 items); (6) *Facilitative Conditions Scale* (eight items); and, (7) *Explicit Directiveness Scale* (four items).

Many of the practices used with the CSPRS were replicated in other measures of adherence and competence including the CBAY-A and the CBAY-C. Specifically, the approaches of using a 7-point rating scale and two coders with expertise in the treatment modality have been used with many measures. On the other hand, procedures used with the CSPRS can be viewed as practical shortcuts without an empirical justification. For example, raters only coded four sessions (sessions 1, 4, 7 or 8, and 14 or 15) from each therapist-patient dyad. While this practice significantly reduced the coding burden and, therefore, cost of the project, there is no empirical support for such a practice. It is also noteworthy that the CSPRS did not measure therapists' competence, nor did the TDCRP studies include another measure of therapists' competence. Readers interested in a more comprehensive review of measures of adherence to treatments for depression are encouraged to review Marder (2007).

Yale Adherence and Competence Scale. The Yale Adherence and Competence Scale (YACS; Carroll et al., 2000) is a 55-item measure of therapist adherence to and competence in behavioral interventions for substance use disorders. Like the CSPRS, the YACS is an observational coding measure designed for use by raters with expertise in the treatment being evaluated (in this case substance use disorders) and focuses on therapist behaviors. Unlike the CSPRS, the YACS used a 5-point Likert-type scale and includes ratings of both adherence and competence. Efforts were made to code every therapy session instead of selected sessions. The

scale ranged from 1 (not at all) to 5 (extensively), and each item was scored for both adherence and competence. The YACS includes six scales: (1) *Assessment* (five items), (2) *General Support* (five items), (3) *Goals of Treatment* (five items), (4) *Clinical Management* (10 items), (5) *Twelve-Step Facilitation* (nine items), and (6) *Cognitive-Behavioral Treatment* (six items; Carroll et al., 2000). It is important to note that the YACS include items that address global competence (e.g., those in the *General Support* scale) and limited-domain competence (e.g., those in the *Twelve-Step Facilitation* and *Cognitive-Behavioral Treatment* scales). As mentioned earlier, measures of competence have primarily focused on limited-domain competence (Barber et al., 2007).

Multisystemic Family Therapy Therapist Adherence Measure. The Multisystemic Family Therapy (MST) Therapist Adherence Measure (TAM) takes a very different approach to the measurement of therapist adherence than the other measures reviewed (Schoenwald et al., 2000). The TAM and the revised version (TAM-R; Schoenwald et al., 2008) grew out of a need for a more cost-effective method for evaluating therapist adherence, and is therefore based on caregiver-report, youth-report, and therapist-report of therapist behaviors rather than observational coding. As a result, it may be more appropriate to consider these measures of treatment receipt and treatment enactment (see Marder, 2007) rather than therapist adherence. Part of the rationale for developing the measure in this way was that MST is designed to be highly responsive to a family's needs, rather than a sequential set of intervention techniques (Schoenwald et al., 2000). Nevertheless, the use of an informant-report instrument versus observational coding represents a variation in methodology for assessing treatment integrity.

The MST TAM consists of 26 items, rated on a five-point Likert-type scale, and was designed to evaluate the therapist's coverage of the nine principles of MST (Schoenwald et al.,

2000). The TAM-R retained 19 of the original items and added an additional nine items resulting in a 28-item measure (Schoenwald et al., 2008). The factor structure of the TAM varied for caregiver-, youth-, and therapist-report and included six, four, and five factors respectively with each measure including a factor believed to represent adherence to MST principles (Schoenwald et al., 2000). The TAM-R is only a caregiver-report measure and has a single-factor structure and a two-point rating scale (Schoenwald et al., 2000). To score the TAM, the mean rating is used. Results from a MST trial indicated that caregiver-report was the best predictor of outcome (Schoenwald et al., 2000). This procedure works in part because the measure is designed to assess adherence to general principles rather than specific techniques.

Adherence measures used in CBT for youth anxiety treatment outcome studies. Currently, there are no measures of therapist adherence to CBT for youth anxiety with sufficient published psychometric data. However, since 1994 there have been over 112 treatment evaluation studies of CBT for youth anxiety (Southam-Gerow, McLeod, Arnold, Cox, Rodriguez, Reise, Bonifay, Weisz, & Kendall, in press). Of those 112 studies, 34 include some form of treatment adherence evaluation, but they were mostly of low or unknown quality, with most using a binary (present/absent) scoring approach and without psychometric data supporting the reliability or validity of the measurement. Methods for evaluating therapist adherence to CBT for youth anxiety include (a) comparing the delivered treatment to checklists corresponding to what would be expected based on the session number and treatment manual (e.g., Kendall et al., 1997; Southam-Gerow et al., 2010) and (b) supervisors reviewing recordings of the treatment sessions (e.g., Cohen et al., 2004).

Binary evaluation of therapist adherence ignores substantial variability in therapist adherence to treatment. For example, the adherence check by Kendall and colleagues (1997)

reported that 100 percent of therapists were adherent to the treatment, and Southam-Gerow and colleagues (2010) and Cohen and colleagues (2004) both reported that therapists were adherent to the treatment greater than 95 percent of the time. Additionally, since this type of coding is binary (adherent or not), ratings of dosage or extensiveness of the treatment provided to the youth are not possible. For example, if a CBT for youth anxiety session is supposed to address relaxation, a binary system may indicate that a therapist that has a simple conversation about the value of relaxation is equally as adherent, and is providing the same dose of CBT, as a second therapist that provides detailed psychoeducation about the role of tension and relaxation in anxiety, describes multiple relaxation strategies, and practices relaxation in session. An adherence measure that includes a 5- or 7-point, Likert-type scale of extensiveness will better reflect the quantity of CBT for youth anxiety therapy that is taking place in the session and lead to greater variance in ratings.

Table 1 summarizes the purpose, type and source of data, focus, scale type, and sampling plan used with various adherence and competence measures.

Features of the Present Study

The present study sought to develop measures of adherence to (CBAY-A) and therapist competence in (CBAY-C) CBT for youth anxiety by building upon the successes of past measures and learning from their shortcomings and evaluating the psychometric strength of the measures. Both measures used observational coding to collect therapy process data. Since there is no empirical support for the practice of sampling from random sessions or some combination of early, middle, and late sessions, data were collected from every codable therapy session from two treatment evaluation trials. For the same reason, the entire session was coded rather than time-sampling sessions. These features are thoroughly discussed in the Method section.

Table 1.

Summary of Select Adherence and Competence Measures

Measure	Aspect of Treatment Integrity Measured	Treatment Measured	Type of data	Source	Focus	Scale type	Sampling Plan
Collaborative Study Psychotherapy Rating Scale (Version 6; Hill, O'Grady, Elkin, 1992)	Adherence	Cognitive Therapy, Interpersonal Psychotherapy, and Clinical Management for Adult Depression	Observational coding	Advanced doctoral students in clinical and counseling psychology	General treatment elements found in CBT, IPT, and CM treatment manuals	7-point Likert-type scale	25% of treatment, 100% of session
Sheffield Project Rating Scale (Shapiro & Startup, 1992)	Adherence	Adherence to Exploratory Psychotherapy	Observational coding	Ranging from graduate students in psychology to clinical psychologists	General practices consistent with Exploratory Psychotherapy	Extensiveness of adherence	25% or 50% of treatment, 100% of session
Penn Adherence/ Competence Scale for Supportive-Expressive (SE) Dynamic Psychotherapy (Barber & Crits-Christoph, 1996)	Adherence and Limited-Domain Competence	Supportive-Expressive Therapy for Cocaine Dependence	Observational coding	Ph.D. level clinical psychologist	How much and "how well" specific practices were used	7-point Likert-type scale	~25% of treatment, 100% of session
Adherence/Competence Scale for IDC for Cocaine Dependence (Barber, Mercer, Krakauer, & Calvo, 1996)	Adherence and Limited-Domain Competence	Individual Drug Counseling for Cocaine Dependence	Observational coding	Therapist experience in treatment modality	Frequency and quality of components delivered from treatment manual	7-point Likert-type scale	~4% of treatment, ~25% of session

Table 1 continued

Measure	Aspect of Treatment Integrity Measured	Treatment Measured	Type of data	Source	Focus	Scale type	Sampling Plan
MATCH Tape Rating Scale (MTRS; Carroll, Connors, Cooney, DiClemente, Donovan, Kadden, Longabaugh, Rounsaville, Wirtz, & Zweben, (1998)	Adherence and Global Competence	CBT, Motivational Enhancement Training, and Twelve-Step Facilitation	Observational coding	Masters and PhD level therapists, most with alcohol treatment experience	Unspecified and therapist skillfulness	Unspecified Likert-type scale for adherence and unspecified for competence	100% of treatment, "a portion" of session.
Therapist Behavior Rating Scale (Hogue, Liddle, Rowe, Turner, Dakof. & LaPann, 1998)	Adherence	Dynamic CBT and Multidimensional Family Therapy for Substance Use	Observational coding	Graduate and undergraduate students	Extensiveness (thoroughness and frequency) of interventions	7-point Likert-type scale	~20% of treatment, 100% of session
Yale Adherence and Competence Scale (Carroll, Nich, Sifry, Nuro, Frankforter, Ball, … & Rounsaville, 2000)	Adherence and Global and Limited Domain Competence	Behavioral treatment for substance use disorders	Observational coding	Masters level therapists experienced in substance use treatment	General CBT treatment for substance use components	5-point Likert-type scale	~85% of treatment, 100% of session
Therapy Procedures Checklist (Weersing, Weisz, & Donenberg, 2002)	Adherence	Psychodynamic, Cognitive, and Behavioral treatments	Therapist Report	Therapist	Frequency of specific practice use	3-point Likert-type scale	Unspecified, N/A

Table 1 continued

Measure	Aspect of Treatment Integrity Measured	Treatment Measured	Type of data	Source	Focus	Scale type	Sampling Plan
Therapist Behavior Rating Scale - Competence (Hogue et al., 2008)	Adherence and Limited-Domain and Global Competence	CBT and Multidimensional Family Therapy	Observational Coding	Professional mental health workers in the community	Extensiveness of adherence and competence to general practices	7-point Likert-type scale	Variable (~33% of treatment), 100% of session
MST Therapist Adherence Measure - Revised (Schoenwald et al., 2008)	Adherence	Adherence to MST	Caregiver-Report	Caregiver	Extensiveness	5-point Likert-type scale	Unspecified, N/A

Both the CBAY-A and the CBAY-C were developed with a focus on therapist behaviors specifically as they relate to CBT for youth anxiety. The CBAY-A was developed to be sensitive to CBT for youth anxiety across different manual-guided treatments or generic CBT targeting youth anxiety. This aspect of the CBAY-A is important to its potential use with a variety of different manual-guided CBTs for youth anxiety within community mental health centers as a service evaluation tool and dissemination and implementation science.

The CBAY-A and the CBAY-C were coded independently. As discussed earlier, therapist adherence and therapist competence are theoretically distinct but related aspects of treatment integrity. Thus far, when they have both been coded within the same study and for the same sessions, they have been included within the same measure. As a result, it is possible that their relationship was artificially inflated due to the shared variance of having a common coder (Foster & Cone, 1995). By having adherence and competence coded by separate coders, the relationship between the two can be more accurately characterized.

Hypotheses. This study will focus on evaluating the CBAY-A and the CBAY-C with regard to their reliability, validity, and uniqueness.

Evaluating reliability. Reliability refers to "the proportion of variance attributed to the true score of a latent variable" (DeVellis, 2003, p. 27). The form of reliability that is most important for a measure is dependent on characteristics of the measure (e.g., the source of the data, the type of scale used, and the theoretical relationship between the items within the measure). For observational coding measures, it is critical that the variance in scores on the measure is due to variance associated with the behaviors being measured rather than some characteristic of the raters. As a result, reliability for the CBAY-A and the CBAY-C should be evaluated using a measure of inter-rater reliability. Since both measures use a 7-point Likert-

type scale (as opposed to categorical or dichotomous scales), intra-class correlations (ICC) are the most appropriate way of measuring inter-rater reliability (Arnold, 2011; Shrout & Fleiss, 1979). It is hypothesized that both the CBAY-A and the CBAY-C will have adequate inter-rater reliability.

Evaluating validity. Validity refers to the extent to which the scores derived by a measure are meaningful and interpretable (Foster & Cone, 1995). Validity cannot be conclusively established for a measure. Rather, evidence supporting the validity of a measure for a particular purpose can be gathered. Both the CBAY-A and the CBAY-C should be supported by content validity, construct validity (convergent and divergent), and discriminative validity. Content validity refers to the extent to which a measure includes items that adequately cover all aspects of the construct being measured and is typically achieved during measure development (Arnold, 2011; DeVellis, 2003). In theory, this is accomplished by identifying all possible items assessing the construct and randomly selecting from these items. This procedure is not realistically feasible so, in practice, content validity is achieved through consultation with experts in the field to review a measure and verify that all possibly important items included. Since content validity is not statistically demonstrated, the discussion of CBAY-A and CBAY-C measure development (presented in the Method section) will address the issue of content validity.

Construct validity refers to the degree to which the items within a measure assess the construct of interest and can be meaningfully interpreted (DeVellis, 2003; Foster & Cone, 1995; Arnold, 2011). There are a variety of different methods for demonstrating construct validity. For the purposes of this study, convergent validity, the degree to which the measure being evaluated correlates with other measures of the same or related constructs, and divergent

validity, the degree to which the measure being evaluated is independent of measures of unrelated constructs, will both be evaluated (Foster & Cone, 1995; Arnold, 2011).

Criterion-related validity refers to a measure's ability to predict meaningful events (Foster & Cone, 1995). Meaningful events can include performance on some real-world task or a test that is considered the gold standard. As there is not established gold standard measure of adherence to and competence in CBT for youth anxiety, this study will employ exploratory analyses to evaluate whether known groups differ in their scores on the CBAY-A and CBAY-C (discriminative validity).

Uniqueness of measure. The final hypothesis being evaluated in this study addresses the relationship between the CBAY-A and the CBAY-C. As noted throughout this paper, adherence and competence are theoretically distinct. However, in the past literature, never before have the same treatment sessions been coded for adherence and competence by independent raters. This study presents a unique opportunity to empirically test the uniqueness of these measures. It is hypothesized that the CBAY-A and the CBAY-C will, in fact, measure unique behaviors.

Method

Data for this project were collected by: (1) coding audio and video recordings of therapy sessions for youth being treated for anxiety disorders; (2) obtaining information from demographic and other questionnaires completed by the youth and families at the time the youth were in therapy; and (3) obtaining information from questionnaires and other data recorded by the therapists and therapists' supervisors at the time of therapy. The therapy sessions used for this study were drawn from one efficacy and one effectiveness study, with three study conditions overall. From the efficacy trial, the individual CBT for youth anxiety condition was used (Coping Cat [CC]; Kendall et al., 2008). From the effectiveness trial, the individual CBT for

youth anxiety condition (Youth Anxiety Study-Coping Cat [YAS-CC]) and the usual care condition (Youth Anxiety Study-Usual Care [YAS-UC]) were used (Southam-Gerow et al., 2010). Hereafter these three conditions will be referred to with their abbreviated labels: CC, YAS-CC, and YAS-UC.

Participants

Participants for this study consisted of youth in treatment and their therapists. Participation on the part of the youth and therapists was based solely on their participation in the treatment outcome study in which they originally participated. That is, no active participation was required of them for the current study. Characteristics of the coders who completed the observational coding for the current study are also presented.

Youth. Across the three treatment conditions from which therapy sessions were drawn, there were a total of 89 youth participants. Some youth were excluded from the present study's analyses because, among other reasons, too few session recordings were available or because they dropped out of treatment before completing the treatment. CC consisted of 51 youth ranging in age from 7 to 14 years (M = 10.36; SD = 1.90; four youth included in the trial analyses were excluded from this study). YAS-CC consisted of 17 youth ranging in age from 8 to 15 years (M = 11.32; SD = 2.32; seven youth from this study condition were excluded). YAS-UC consisted of 21 youth ranging in age from 8 to 14 years (M = 10.44; SD = 1.91; three youth from this study condition were excluded). Thus, the overall age range was from 7 to 15 years (M = 10.56; SD = 2.0). Overall, the participants were 47.2% percent female (39.2% CC, 70.6% YAS-CC and 47.6% YAS-UC in the three conditions respectively) and most (65.2% for the total sample) were European American (86.3%, 41.2%, and 33.3% respectively) followed by

African American (9.8%, 0%, and 9.5% respectively), Latino/Hispanic (2.0%, 17.6%, and 42.9% respectively), and other ethnicities (2.0%, 5.9%, and 9.5% respectively)[5].

Youth in the CC trial were recruited following being referred to a university-based clinical psychology training clinic for treatment of an anxiety disorder between 2000 and 2006. To be eligible for the trial, youth had to be between seven and 14 years of age, meet diagnostic criteria for one of three primary anxiety disorders (Generalized Anxiety Disorder, Separation Anxiety Disorder, or Social Phobia), and be willing to be randomized to either Individual CBT, Family CBT, or Family Education/Support/Attention (FESA). Only those participants from the Individual CBT treatment condition were included in the current study. Exclusion criteria were psychotic symptoms, intellectual disability, a disabling medical condition, or concurrent extra-trial psychological or pharmacological treatments. A total of 161 youth participated in the trial; 55 were randomized to the Individual CBT condition. Four of these youth dropped out of treatment prior to completion of treatment and were therefore not included in the present analyses. The primary Diagnostic and Statistical Manual of Mental Disorders (4th edition; DSM-IV; American Psychiatric Association, 1994) diagnosis of the youth in the CBT efficacy trial was either Generalized Anxiety Disorder (GAD; 37.3%), Separation Anxiety Disorder (SAD; 29.4%), or Social Phobia (SOP; 33.3%), assessed via structured interviews (Anxiety Disorder Interview Schedule for Children [ADIS-C/P]; Silverman & Albano, 1996) with both youth and parent. Diagnoses were assigned to the youth if either the youth or the parent interview indicated that youth met diagnostic criteria for an anxiety disorder. Further details about the youth participants in the CBT efficacy trial and associated procedures can be found in Kendall et al. (2008).

[5] Ethnicity is unknown for one youth in the YAS-UC condition and six youth in the YAS-CC condition.

Youth in the effectiveness trial (YAS-CC and YAS-UC) were recruited for participation in the study during the routine intake process at six community mental health clinics in a large urban setting. In order to be included in the study, youth had to be between 8 and 15 years of age, have a primary DSM-IV diagnosis of GAD, SAD, SOP, or specific phobia (SP), and have anxiety be identified by the family as the treatment priority. Exclusion criteria included a diagnosis of a pervasive developmental disability, psychotic disorder, or intellectual disability. Youth diagnoses were assessed using the Diagnostic Interview Schedule for Children Version 4.0 (DISC 4.0; Schaffer, Fisher, & Lucas, 1996). As with the efficacy trial, diagnoses were assigned if either the youth or parent responses indicated that the diagnosis was present. Of the 268 youth who were assessed for eligibility into the study, 48 were enrolled and randomized to one of the two treatment conditions (24 in each condition). Of these 48 youth, 17 from the YAS-CC and 21 from the YAS-UC were included in the present study. Participants had a primary diagnosis of GAD (5.9% and 14.3% respectively), SAD (35.3% and 38.1), SOP (23.5% and 28.6%), or SP (35.3% and 19.0%). As expected with this sample, assessed youth had multiple diagnoses at the time of intake (YAS-CC: M = 2.6 diagnoses, SD = 1.18 and YAS-UC: M = 3.0 diagnoses, SD = 1.16). These diagnoses included other anxiety disorders (SP, 64.7% in YAS-CC and 71.4% in YAS-UC; SAD, 41.2% and 57.1%; SOP, 47.1% and 38.1%, GAD, 17.6% and 23.8%; panic disorder, 11.8% and 9.5%; and post-traumatic stress disorder, 5.9% and 0%) and non-anxiety disorders (attention deficit/hyperactivity disorder, 17.6% and 42.9%; oppositional defiant disorder, 23.5% and 33.3%; conduct disorder, 0% and 14.3%; major depressive disorder, 11.8% and 0%; and dysthymic disorder, 5.9% and 0%). Further details about the youth participants in the two conditions in this effectiveness trial and associated procedures can be found in Southam-Gerow et al. (2010).

Therapists and treatment. The therapists (N = 16) administering CBT in the CC condition were either master's-level therapists with between two and three years of experience at the clinic, or doctoral-level therapists. All therapists participated in two 3-hour workshops on the CBT approach used in the study (Coping Cat), reviewed treatment manuals, and participated in weekly two-hour group supervision sessions with doctoral-level supervisors with six or seven years of experience. The weekly supervision took place for the duration of the trial. Since therapy cases were randomly assigned to therapists, and therapists in the CC condition were also trained in Family CBT and FESA, it is likely that these therapists were also administering Family CBT or FESA in addition to Coping Cat.

The therapists in the YAS-CC condition (N = 13) and YAC-UC condition (N = 16) within the effectiveness trial had a variety of different professional backgrounds (i.e., social workers, 30.8% and 25.0% respectively; doctoral-level psychologists, 15.4% and 6.3%; master's-level psychologists, 7.7% and 0%; and other backgrounds, 46.2% and 43.8%; 25% of the therapists from the YAS-UC condition did not report their highest degree). The average age of the YAS-CC therapists was 36.0 years (SD = 12.52; *range* 26 – 65) and they were predominately European American (53.8%) followed by Latino/Hispanic, Asian American/Pacific, and mixed/other ethnicity (15.4% each). Their average age of the YAS-UC therapists was 29.3 years (SD = 3.85; *range* 25 – 40) and they were predominately Latino/Hispanic (37.5%), followed by European American (31.3%) and mixed/other ethnicity (6.3%). Four YAS-UC therapists did not provide age and ethnicity data. Of the therapists that participated in the original effectiveness trial, the average number of years of clinical training was 4.40 (SD = 2.20) and the average number of years of professional experience was 4.90 (SD = 8.20). (Note: These data are not available by study condition and include therapists that

participated in the original trial but whose data were not included in the present study). Therapists were randomized to either the CBT condition or the UC condition. No significant demographic differences were found between therapists in the CBT and UC condition within the effectiveness trial (Southam-Gerow et al., 2010).

Therapists in the CBT condition received training and supervision in CBT for youth anxiety that consisted of a one-day, six-hour training and weekly supervision with one of two doctoral-level therapists. UC therapists received the form and quantity of supervision that they received for their non-study patients. The CBT for youth anxiety treatment in which the therapists in the two CBT conditions were trained and that they administered was based on the Coping Cat Manual (CC; Kendall & Hedtke, 2006a, 2006b). The CC is a 16- to 20-session treatment program that was developed specifically to treat youth anxiety. It includes a skills development phase (sessions 1 through 8) and an exposure or practice phase (sessions 9 through termination). The skills development phase includes sessions focusing on relaxation, emotion education (e.g., identifying different emotions), cognitive coping (e.g., identifying and challenging maladaptive anxious thoughts), problem solving, and the use of rewards for approach behavior. The exposure or practice phase includes multiple sessions focusing on imaginal or in-vivo exposure to the youth's feared stimuli. While concurrent pharmacotherapy was prohibited in the efficacy trial, decisions about concurrent pharmacotherapy in the effectiveness trial were made consistent with what was typically done at the clinic.

Coders

A total of four graduate student coders (two teams of two coders) reviewed recordings (audio or video) of each therapy session. Coding occurred from 2011-2014. When coding began, the coders ranged in age from 24 to 31 years of age and were between their first and

fourth year in a clinical psychology doctoral program. Of the four coders, three were female and one was male. Two coders self-identified as being White, one as Mexican American, and one as White/Hispanic. Three of the four coders were specializing in child clinical psychology, the fourth was specializing in adult clinical psychology, and two of the four coders had their master's degree at the time that they started coding. The two coders coding therapist CBAY-A coded all of the sessions (N = 954), while the two coders that began coding CBAY-C coded approximately half of the sessions. The remaining CBAY-C coding was completed by the coding team that began coding therapist adherence (see Table 10 for complete numbers). Each coding team received supervision from the measure developers. These meetings occurred weekly during pilot coding and early stages of coding and tapered to monthly as the coders became more familiar with the coding manuals. Throughout the study, efforts were made to keep the coders blind to the treatment condition of the session that they were coding, though as will be further covered in the Discussion, there were limitations to how blind coders could truly be after seeing so many sessions (e.g., cues from clinic rooms, etc.).

Session Recordings

During the two trials, efforts were made to record every session using audio or video recording devices. These recordings were then transferred to audio CDs or video DVDs. Coders viewed and coded the therapy sessions from the three treatment conditions using these recordings. Out of a potential 1428 sessions (*n* = 812, 286, and 330 for condition CC, YAS-CC, and YAS-UC respectively), 954 sessions (66.80%) were codable and were used in analyses (*n* =532, 212, and 210 from the respective conditions). Reasons for sessions not being included in analyses included: (a) sessions were not provided by the original trials to this project for unknown reasons; (b) youth were dropped from analysis due to having too few codable sessions;

(c) significant audio or visual problems resulting in the session content being inaudible; (d) session recordings were incomplete (i.e., less than 15 minutes in length); and (e) majority of a session occurring in a non-English language.

Measures

CBT for Youth Anxiety Therapist Adherence Scale (CBAY-A). The CBT for Youth Anxiety Therapist Adherence Scale (CBAY-A) is a 22-item[6], observational coding measure of therapist adherence to cognitive behavioral therapy for youth anxiety. Items are organized into three categories: (a) *Standard Items* that reflect basic CBT elements (e.g. agenda setting and homework review); (b) *Model Items* that reflect core interventions in CBT for youth anxiety (e.g. relaxation, cognitive coping, and exposure); and (c) *Delivery Method Items* that reflect how model items are delivered (e.g. collaborative teaching, modeling, and rehearsal). Each item is coded on a 7-point Likert-type *Extensiveness* scale (1= 'Not at All', 3= 'Somewhat', 5= 'Considerably', 7= 'Extensively'). Therefore, sessions for which a behavior is observed, ratings range from two through seven. The Extensiveness ratings are similar to those used in previous measures (e.g., Hogue et al., 1998; Carroll et al., 2000; Marder, 2007) that consider the frequency and thoroughness of interventions.

The CBAY-A is intended to assess the therapist's adherence to treatment components commonly found in evidence-based CBT for youth anxiety rather than the treatment components and practices specific to a particular manual-guided CBT for youth anxiety treatment (e.g. Coping Cat). As such, the sequencing of treatment components and specific delivery methods and terminology is not evaluated. For further details on the CBAY-A, interested readers should

[6] The CBAY-A originally had 23 items but one, Weekly Ratings, was removed because it was so infrequently observed in the recordings for this study and was, therefore, not possible to evaluate.

refer to the unpublished CBAY-A coding manual (available upon request from McLeod and Southam-Gerow, 2010).

CBAY-A measure development. The procedures used for developing the CBAY-A included the following steps. First, subscales were developed that mirrored the PASCET Manual Adherence Scale (PMAS; Marder, 2007), an adherence to CBT for youth depression based on the Primary and Secondary Control Enhancement Training (PASCET; Weisz, Moore, Southam-Gerow, Weersing, Valeri, & McCarty, 1997). These subscales were (a) *Standard Items*, (b) *Model Items*, and (c) *Delivery Method* Items. *Standard Items* are meant to measure therapist behaviors associated with features of CBT that occur in most sessions (e.g., Agenda Setting, Homework Review, and Homework Assigned). *Model Items* are designed to measure therapist behaviors associated with specific treatment components common to CBT for youth anxiety (e.g., Psychoeducation-Anxiety, Relaxation, Problem Solving, and Exposure). *Delivery Method Items* are designed to measure the therapist behaviors associated with different methods of teaching, practicing, or otherwise conveying the content associated with different *Model Items*.

The second step in the measure development process was item development. Items were developed within each of the three subscales in two ways. First, CBT for youth anxiety treatment manuals (e.g., Coping Cat, Kendall & Hedtke, 2006a, and Modular Approach to Therapy for Children with Anxiety, Depression, Trauma, or Conduct Problems, Chorpita, 2007) were reviewed and items were developed associated with prescribed content. Second, experts in the field of CBT of youth anxiety reviewed the list of items generated from the first step and were asked to add additional items as they felt was appropriate. From this list of items, items with similar content were collapsed. The CBT for youth anxiety treatment manual was then

reviewed to ensure that the content prescribed in the select treatment manuals was covered by the final list of items.

The third step involved selecting a scoring strategy. A 7-point Likert-type scale reflecting the extensiveness to which the therapist was adherent to CBT for youth anxiety was selected in order to allow coders to indicate the degree to which sessions were implemented with adherence. Extensiveness, as stated in the CBAY-A manual, "refers to two dimensions: the *thoroughness* of a CBT intervention for child anxiety and the *frequency* of the CBT intervention for child anxiety" (Southam-Gerow, McLeod, Arnold, & Rodríguez, 2010, p. 4). As noted elsewhere in this paper, this type of extensiveness scale had been used in other treatment integrity measures.

The fourth and final step in the development of the CBAY-A coding manual involved pilot coding, meetings between coders and measure developers, and revisions to the coding manual. During this step, the two CBAY-A coders reviewed randomly selected psychotherapy sessions from the three conditions, coded therapist adherence, identified questions regarding variable definitions, identified exemplars for variables, and identified points of potential item overlap. A mixture of independent and joint coding was done during this step. The questions, exemplars, and instances of item overlap were reviewed during the meetings with measure developers. Modifications to the coding manual were made when appropriate. The measure development phase ended when each coder and the measure developer agreed that the coding manual was clear and each variable assessed the appropriate therapist behaviors.

CBT for Youth Anxiety Therapist Competence Scale (CBAY-C). Similar to the CBAY-A, the Cognitive-Behavioral Therapy for Youth Anxiety Therapist Competence Scale (CBAY-C) includes 23 items that fall under three categories: (a*) Standard Items* for basic CBT elements that

occur in each session (e.g. homework review); (b) *Model Items* that represent the core interventions in individual CBT for youth anxiety (e.g. exposure); and (c) *Delivery Items* that refer to how model items are delivered by the therapist (e.g. role playing). Each observed item was rated on a 7-point Likert-type *Competence* scale (0 = 'Not Present', 1 = 'Very Poor', 3 = 'Acceptable', 5 = 'Good', 7 = 'Excellent'). For this measure, therapist competence was defined as including a combination of the therapist's *skillfulness* (i.e., the technical quality of the intervention performed by the therapist) and *responsiveness* (i.e., the timing and appropriateness with which the therapist administers the treatment). For further details on the CBAY-C, interested readers should see the unpublished CBAY-C coding manual (available upon request from McLeod and Southam-Gerow, 2010).

CBAY-C measure development. The aim in developing the CBAY-C was to create a measure that assessed therapist competence in delivering CBT for youth anxiety. The steps in the measure development of the scoring manual for the CBAY-C (Southam-Gerow, McLeod, Quinoy, & Eonta, 2010) were similar to the measure development of the CBAY-A. Because this measure was designed to assess competence in delivering CBT for youth anxiety rather than competence in global therapeutic activities, the measure was designed to assess the therapist's competence in delivering the content measured by the CBAY-A and the delivery methods assessed by the CBAY-A. Therefore, with regard to step one and two of measure development, the same three subscales were used for the CBAY-C as was used in the CBAY-A and the same items were used for the *Model Items* subscale and the *Delivery Methods Items* subscale. Some items in the *Standard Items* subscale were different. As with the CBAY-A, CBAY-C items were developed following a review of established measures of therapist competence and through consultation with experts in CBT for youth anxiety.

At step three (selecting a scoring method for the CBAY-C), the goal was to select a method that would be sensitive to variance between therapists and variance between therapeutic tasks. Similar to the CBAY-A, a 7-point Likert-type scale was selected. Scores for each item reflect a combination of technical quality with which the intervention was delivered (*Skillfulness*) and the timing and appropriateness of the intervention for the particular youth (*Responsiveness*). Specifically, "raters are asked to consider the extent to which a therapist demonstrated the following dimensions that comprise Competence (see Carroll et al. 2000): (a) expertise, commitment, motivation; (b) clarity of language and communication; (c) appropriate timing of interventions/actions (responsiveness); and (d) ability to read and respond to where the client appears to be (responsiveness)" (Southam-Gerow, McLeod, Quinoy, & Eonta; Unpublished, p. 4). Unlike the CBAY-A, which is designed to assess only the therapist's behaviors, scores on the CBAY-C also take the youth's (or parent's) responses into consideration. Lastly, the pilot coding and coding manual revision process for the CBAY-C was identical to that used with the CBAY-A.

Therapy Process Observational Coding System-Strategies (TPOCS-S). The Therapy Process Observational Coding System-Revised Strategies (TPOCS-S) is a 31-item observational coding measure (McLeod & Weisz, 2010). The TPOCS-S was designed to characterize the therapeutic practices of therapists and as a measure of treatment differentiation. The 31 items on the TPOCS-S are organized into five scales corresponding to different therapeutic orientation (i.e., *Cognitive Therapy*, *Behavioral Therapy*, *Family Therapy*, *Psychodynamic Therapy*, and *Client-Centered Therapy*) and one scale (*General*) assessing common treatment practices. Unlike the CBAY-A and the CBAY-C, the TPOCS-S is not specific to the target of therapy (e.g., anxiety disorder treatment).

An evaluation of the TPOCS-S indicated that it has strong psychometric properties (McLeod & Weisz, 2010). Specifically, inter-rater reliability ranged from 'Good' to 'Excellent' at both the item level (ICCs ranged from .66-.95, *M* = .84, *SD* = .08) and the subscale level (ICC's ranged from .79-.97, *M* = .89, *SD* = .07; McLeod & Weisz, 2010). Internal consistency of the items within the scales of the TPOCS-S had alpha coefficients ranging from .74 to .86. To evaluate the validity of the TPOCS-S, the subscale ICCs were compared to the correlation between subscales. In each case, the ICC values were greater than the correlations between subscales, thus supporting the validity of the TPOCS-S (McLeod & Weisz, 2010). For further details on the TPOCS-S, interested readers should see McLeod and Weisz (2010) and the unpublished TPOCS-S coding manual (available upon request from McLeod).

Of the five TPOC-S scales, the *Cognitive* and *Behavioral* scales include items that are similar to items on the CBAY-A. Conversely, the items on the *Psychodynamic*, *Family*, and *Client-Centered* scales should be sensitive to different behaviors than the items on the CBAY-A. Based on data provided in the psychometric evaluation of the TPOCS-S, the mean correlation between the *Cognitive*/*Behavioral* scales and the *Psychodynamic*/*Family*/*Client-Centered* scales equals .193 (i.e., 6 cells; McLeod & Weisz, 2010)[7]. This value will be used as the criterion value against which comparisons used to evaluate the divergent validity of the two measures will be compared.

For this project, the Therapy Process Observational Coding System-Revised Strategies (TPOCS-RS) was used. The TPOCS-RS is very similar to the TPOCS-S but has 12 additional

[7] When the same set of correlations, the *Cognitive* and *Behavioral* scales and the *Psychodynamic*, *Family*, and *Client-Centered* scales, is run using the Therapy Process Observational Coding System-Revised Strategies (TPOCS-RS) and the sample used for this study, the mean magnitude of the correlation coefficients was were substantially greater (r=.331 for Item-Level scores; r=.516 for Item-Mean scores). Additionally, for the present study, the *Cognitive* and *Behavioral* Scales were negatively correlated with the *Psychodynamic* and *Family* Scales and positively correlated with the *Client-Centered* Scale, whereas all six correlations were positive in the original TPOCS-S publication.

items (McLeod, Smith, Southam-Gerow, Weisz, & Kendall, 2014). Similar coder training and coding procedures were used for the TPOCS-RS as was described for the CBAY-A and the CBAY-C with less attention focused on item development and manual modifications. Additionally, characteristics of the TPOCS-RS coders for the present study were very similar to the coders of the CBAY-A and CBAY-C (i.e., graduate students in a clinical psychology training program).

Common Factor Therapist Competence Scale for Youth Psychotherapy. The Common Factor Therapist Competence Scale for Youth Psychotherapy (COMP-CF) scale is an observational coding measure designed to measure elements of therapist competence that are common across youth psychotherapies regardless of therapeutic orientation (Brown, 2011). The COMP-CF consists of 14 microanalytic items and five domain-level items that correspond to the five domains measured: Alliance-Building, Increasing Positive Expectancies, Instigating Change, Focusing Treatment, and Responsiveness. The microanalytic items measure therapist behaviors specific to the five domains listed above. The focusing treatment and instigating change domain-level items are of greatest interest within this study. Preliminary data indicates that these items can be coded reliably (ICC(2,2) = .77 and .79 respectively).

Procedures

The procedures for this study can be roughly organized into five phases: (a) measure development (described earlier), (b) pilot coding, (c) certification, (d) coding, and (e) data analysis. This section will specifically address the pilot coding, certification, and coding phases as the measure development phase was described earlier and the data analysis will be discussed in a subsequent section that also includes the study hypotheses. Unless noted otherwise, the

procedures for the development, pilot coding, and coding phases for the CBAY-A and CBAY-C are the same.

Pilot coding. The pilot coding phase of this study served two purposes. First, during this phase, the coders reviewed the CBAY-A or CBAY-C coding manual and the Coping Cat manual, familiarized themselves with the operational definitions for each code, and practiced coding. Second, the coders and the measure developers edited the coding manuals to improve the definitions of items and to generate *item distinction* sections in the coding manual (descriptions of how items are distinct from one another). During this phase, coders coded sessions from all three conditions, sometimes coding sessions together and sometimes coding independently. Sessions were selectively assigned to the coders to ensure that they had exposure to each item on their respective scale. Weekly supervision meetings were held with the coding teams and the measure developers. In these supervision meetings, questions about items were discussed and inter-class correlations (ICCs) between coders coding the same sessions were monitored. Additional training and focus was given toward items with low or falling ICCs.

Certification. During the certification phase, each coder coded 32 sessions independently at a rate of approximately 17 sessions per week. Certification sessions were carefully selected to sample sessions from early, middle, and late portions of therapy and sessions that included each item on the respective scale. Unlike during the pilot coding phase and later coding phase, no supervision meetings were held and all sessions were coded independently.

After coding the 32 certification sessions, each coder's ratings (i.e., extensiveness for the CBAY-A and competence for the CBAY-C) were compared to the other coder for the measure and to the ratings made by the measure developers. The ratings by the measure developers were

obtained by taking the mean of their ratings after they independently coded the same certification sessions. Each coder was independently deemed "certified" in the coding system if the ICCs between his or her ratings and the mean ratings of the measure developers were above .60 (mean ICC). For individual items with ICC values below .60 for the certification sample, case-by-case decisions were made regarding how to proceed. In some instances (e.g., Maintenance), low ICC values were seen as a consequence of infrequent observations of the practice and not interpreted as problematic. In other cases (e.g., Psychoeducation Anxiety and Modeling), low ICC values were addressed with discrepancy analysis and ongoing monitoring during weekly meetings.

Discrepancy analysis consisted of identifying sessions for which the codes provided by the two coders were highly different, both coders reviewing the session together, identifying the cause of the discrepancy, and, in consultation with the research team, making modifications to the coding manual to address the cause of the discrepancy (e.g., modifying the definition of the item, describing therapist behavior that should receive high and low codes, and describing differences between items). The coders then made an effort to code future sessions consistent with these modifications to the coding manual. This standard has been used in evaluating other behavioral observation measures of treatment integrity (e.g., Brown, 2011). Discrepancy analysis, coder team meetings, and rereading the coding manual were the primary steps taken to prevent and address coder drift.

Coding. During the coding phase of the study, each coder independently coded sessions from all three conditions at a rate of approximately 17 sessions per week. As noted previously, supervision meetings during this phase of the study were initially held weekly but tapered to monthly as familiarity with the coding systems increased. Throughout the coding phase, when large discrepancies in ratings between two coders were identified for a particular item and for a

particular session, the coders re-watched the session together and discussed the item. Codes were not changed following these reviews unless the discrepancy was attributed to clerical or data entry errors.

As noted earlier, the CBAY-A coding team coded every codable recording. The CBAY-C coding team, on the other hand, coded approximately two thirds of the codable recordings prior to leaving the .As a result, the coding team that coded the sessions using the CBAY-A then coded the remaining CBAY-C sessions. Prior to doing this, they went back through the pilot coding phase and certification phase for the CBAY-C measure.

Data analysis. This section describes the process of data entry and management, as well as scoring and analysis.

Data entry and management. After code sheets were completed by the coders, they were entered into two independent but identical electronic databases (SPSS, Version 20) by trained undergraduate research assistants. To confirm the accuracy of the data, the two datasets were then compared using SPSS Syntax that subtracted corresponding data points from one another (e.g., VAR1_de1 – VAR1_de2). Any resulting non-zero numbers, indicating non-agreement between the first and second entry of the data, were checked to determine the cause of the inconsistency. Any of these data entry errors were resolved by returning to the paper copy of the scoring sheet for verification. Additionally, in some instances, checks for logical inconsistencies in the data were made (e.g., for the CBAY-A scale, if a practice is seen by the coder [frequency greater than zero], it must receive an extensiveness score between two and seven). Instances of logical inconsistencies were investigated and errors resolved. When the error was made at the level of the coder, the coder was asked to fix the error. In some instances, this required the coder to re-watch all or part of a session. Finally, session characteristics were

compared between coders and discrepancies were identified. Examples of discrepancies included substantial differences in the length of the session, and number of people involved in the session (which were captured on the coding sheets). Though rare (e.g., three instances were found for the adherence coding team), when they were found, the coding teams were asked to recode the session to ensure accuracy.

The integration of the dataset created from the code sheets and the datasets from the original trials was accomplished by matching the data by youth identification numbers and merging the data using SPSS Syntax. All syntax was double-checked by at least one member of the research team with extensive experience using SPSS Syntax.

Scoring. Prior to discussing the scoring options and procedures for the CBAY-A and the CBAY-C, it is important to review each rating scale and how the score for the two raters were combined, as they are different in important ways. As noted earlier, the CBAY-A is scored on a 7-point Likert-type interval scale with scores ranging from 1 'Not Present' to 7 'Extensive'. As such, there is a direct relationship between the rating for an item and the extensiveness of the therapist's delivery of the element associated with the corresponding item.

In contrast, the CBAY-C measure has the following scoring options: 0 = 'Not Present' and 1 = 'Very Poor' to 7 = 'Excellent'. The CBAY-C can be considered a combination of an ordinal scale (dichotomous; 0 = 'Not Present' and 1-7 = 'Present in Some Fashion') and an interval scale. Specifically, it is a dichotomous scale in that coders first must determine if a particular therapist behavior took place or not. If not, the coders provide a rating of 0 = 'Not Present'. If the coders determine that the behavior did take place, then the measure uses the 7-point Likert-type interval scale and provide a rating of the therapist's competence using rating ranging from 1 = 'Very Poor' to 7 = 'Excellent'. The term "positively scored" refers to instances

where both coders determined that the therapist behavior was present and provided a rating ranging from 1 to 7, the interval portion of the rating scale.

For the purposes of these analyses, scores for an individual item (e.g., Relaxation) for both measures reflect the mean of the two raters' ratings. This procedure is straightforward with regard to the CBAY-A since the ratings have a direct relationship with the extensiveness of delivery of that item by the therapist. For example, if one rater does not see the treatment element while the other rater sees the element at a low level, their ratings may be 1 and 2 respectively. The mean of these two values (1.5) is easily interpretable.

Given the same circumstance, the coder of the CBAY-C scale who does not see the treatment element would provide a rating of 0 while the coder who sees a small amount of the element must rate the therapist on her or his delivery of the element. In this situation, the scale should be considered ordinal and the mean of the two scores is not meaningfully interpretable. Therefore, CBAY-C ratings where both raters agreed that the treatment element did not occur will correspond to a combined rating of 0. Instances where both raters detect the presence of the element will result in a rating that is equal to the mean of the two coders' ratings. However, instances where one coder detects the treatment element (scoring the item 1-7) while the other coder does not detect the treatment element (scoring the item 0 = 'Not Present') will be coded as 'Unscored Due to Disagreement'. These procedures mean that there will be sessions for which some treatment elements have ratings indicating that the element was delivered on the CBAY-A while the corresponding CBAY-C rating is 'Unscored Due to Disagreement'.

There are many different ways in which the item scores can be used. These include, but are not limited to, simply using the mean score from the two raters on each item, combining all of the scores for a single session, or combining all of the scores for a single item from each

session delivered to a youth. For this study, scores were derived by taking the mean score of the two raters. This scoring approach is very simple, has broad utility, and acts as the basis for many other potential scoring approaches. For the CBAY-A, scores will indicate the degree of extensiveness that the treatment element was delivered. For the CBAY-C, scores will indicate the degree to which the therapist competently delivered the item or that the element was not detected by the raters.

Missing data. Prior to evaluating the psychometrics of the CBAY-A and CBAY-C, the patterns of non-present data were evaluated. Though one of the goals for this study was to code every delivered session, some sessions were not coded and both CBAY-A and CBAY-C data include missing data (see the Session Recordings section for reasons why sessions were not coded).

Missing data was evaluated with two approaches. First, the rate of *sessions held per youth*, the rate of *sessions coded per youth*, and the *percent coded per youth* were compared across the three study conditions using analysis of variance (ANOVA). Because the same set of sessions was coded for the CBAY-A and the CBAY-C, these comparisons applied to both measures. Next, steps were taken to evaluate whether there was evidence for a pattern within the missing data. Evaluating the patterns of non-present data is important step when considering the confidence that can be placed on analyses of the data that is *present*. It is possible that some sort of bias existed in the pattern of the values that were missing within the current dataset; systematic bias in missing data would reduce the confidence that can be placed on the conclusions drawn from the analyses conducted on the dataset.

To evaluate the pattern of values for the missing data, procedures outlined by Schlomer, Bauman, and Card (2010) will be used. The first step to assessing the pattern of missing values

involved evaluating whether the data were "*missing not at random"* (MNAR; Schlomer et al., 2010). If data are found to be MNAR, this would be problematic, as the data values that are missing would depend on unobserved data (i.e., data is not present for a specific reason), and bias would be introduced into the existing data. Evaluating whether or not data are MNAR is done conceptually, rather than statistically. While it is possible that the uncoded sessions were disproportionally high or low in adherence and competence, this seems unlikely and was not raised as a concern in the original articles describing the trials. Therefore, while it is impossible to rule out MNAR (Schlomer et al., 2010), the potential for these data to be MNAR seems acceptably small.

The second step in evaluating the pattern of missing data was to determine if the data were *missing at random* (MAR) or *missing completely at random* (MCAR; Schlomer et al., 2010). If the uncoded sessions were more likely to be from one study condition versus the others or particular therapists, then the data would be MAR (i.e., the reason for the pattern is identified and can be attributed to observed variables in the dataset). If data are determined to be MCAR, then missingness is not related to other observed variables (e.g., session number, study condition, therapist, sessions held, sessions coded).

Schlomer et al. (2010) recommend then conducting Little's (1988) omnibus MCAR test on all variables to be used for analyses to assess whether the data that are missing completely at random. Little's test involves calculating a chi-square statistic that tests whether data points that are missing within a dataset either (a) exhibit an identifiable pattern (i.e., the data points that are non-present are related to another variable, such as a characteristic that would make all sessions missing for a particular condition), or (b) do not exhibit an identifiable pattern. When data points that are missing within a dataset are found to have no identifiable pattern, they are classified as

MCAR. When this statistical test is applied more generally, having no identifiable pattern for data points that are missing means that they may be presumed to occur completely at random (i.e., the values are independent of both observed variables and of unobserved parameters). Therefore, when values that are missing are found to have no identifiable pattern via Little's test, the analyses performed on the data may be presumed to be unbiased by other variables that may otherwise have an influence on these missing values.

To evaluate the pattern of uncoded sessions, the following variables were included in Little's test: (a) Agenda Setting from each session (1-53); (b) Study Condition; (c) Therapist; (d) Sessions Held; (e) Sessions Coded; and (f) Percentage of Sessions Coded. Since each coder provided a rating for every item for every coded session on both the CBAY-A and CBAY-C, the selection of the CBAY-A Agenda Setting item versus another item is inconsequential. These findings are outlined in the Results section.

Item performance. The first phase of evaluating the psychometrics of the CBAY-A and the CBAY-C was to describe the item performance. This consisted of describing each item's mean, standard deviation, range, and normality (see Tables 4 and 9). While there were expectations for how the measures would perform (i.e., mean near the midpoint of the scale; the full range of scores used), there were no specific hypotheses. Items were not removed if they did not perform as expected. The purpose of evaluating item performance was twofold. First, poor item performance may be due to a poorly-worded item definition, insufficient coder training, or use of the item with a sample (in this case therapy sessions) that is ill-suited for the item. Identifying these issues can make it possible to address any of these problems. Second, understanding item performance can influence how the item is used in future analyses. For

example, statistical analyses that depend on normally distributed variables should be avoided if items are found to be non-normal.

One method of evaluating normality is to look at the skewness and kurtosis statistics for each item. However, this approach is problematic with the data in this study because of the large sample size. Field (2005) warns against using skewness and kurtosis statistics for samples that are greater than N=200; instead, Field recommends visually inspecting histograms for each variable. While visual inspection of histograms may help identify severely non-normal distributions, it is relatively unhelpful when distributions are approaching normality. To address this issue, the Shaprio-Wilk test of normality compares the distribution of interest to computer generated, normal distributions with the same mean and standard deviation (Field, 2005). If the Shapiro-Wilk test is significant (p<.05), then it is likely that the distribution in question came from a non-normal sample.

***Reliability*.** Reliability (the degree to which scores on a measure are the result of the behavior itself rather than characteristics of the rater, form, or time) is a critical prerequisite to measure is put into widespread practice. Different types of measures (e.g., self-report versus observational data) call for different tests of reliability. When scores are provided by a coder or rater, it is important to establish that the coder is a small source of the variance in scores relative to the phenomenon being measured. This is done by comparing different coder's scores on the measure when rating the same stimuli (recordings of psychotherapy session in this case). Depending on the type of scores being provided by the coder (e.g., categorical variables versus continuous variables), the comparison of scores focuses on agreement (categorical) or correlation of scores between coders (continuous).

The CBAY-A uses an interval scale while the CBAY-C uses an ordinal scale and an interval scale. For the interval scale portion of each measure, intra-class correlation (ICC) is the most appropriate method of evaluating the inter-rater reliability (Cicchetti, 1994; Shrout & Fleiss, 1979). The ICC(2,2) analyses for the CBAY-C only include sessions for which both coders rated the item as present. This method emphasizes the importance of consistency of each rater (i.e., that a given observed set of behaviors will consistently result in a specific score and that a specific rater's tendency to rate therapist behaviors is consistently rated favorably or unfavorably relative to other raters) as opposed to agreement. For the ordinal scale portion of the CBAY-C, agreement is important. Therefore, inter-rater reliability of that portion of the measure is evaluated with Cohen's Kappa (Cicchetti, 1994; Ludbrook, 2002; Sims & Wright, 2005), which takes into account the influence of agreement by chance.

The model of ICC (ICC[2,2]) used to evaluate the reliability of the CBAY-A and CBAY-C reflects that the same coders coded every session, and that the analysis is based on the average between the two coders. Interpretation of the ICC(2,2) and Cohen's Kappa are based on guidelines put forth by Cicchetti (1994). Specifically, ICC(2,2) values less than .40 are considered '*Poor*', values between .40 and .59 are considered '*Fair*', values between .60 and .74 are considered '*Good*', and values above .75 are considered '*Excellent*'.[8]

Validity. The validity of a measure can be demonstrated in a variety of ways, but is generally demonstrated by showing that the scores on a measure are related to independent observations of the behavior or construct of interest, or independent observations of behaviors or constructs that are thought to be closely related or co-occurring (Groth-Marnat, 2003). Two

[8] Sims and Wright propose different cutoffs for interpreting Cohen's Kappa. Specifically, they suggest that Kappa coefficient values can be interpreted as follows: "[less than] 0=poor, .01–.20=slight, .21–.40=fair, .41–.60=moderate, .61–.80=substantial, and .81–1=almost perfect (Sims & Wright, 2005, p. 264). For these analyses, the more conservative interpretation guidelines presented by Cicchetti will be used.

types of validity that will be discussed and evaluated: construct validity (convergent and divergent validity) and discriminative validity.

To evaluate the construct validity of the CBAY-A and CBAY-C, the convergent validity and the divergent validity of each measure were investigated. First, to evaluate convergent validity, CBAY-A data were compared to items and scales from the TPOCS-RS that, based on the definitions of the items in the two coding manuals, are similar to one another. The expectation was that these data would be positively correlated. Specifically, it was hypothesized that correlation coefficients would be moderate in magnitude (i.e., equal to or greater than .30; Arnold, 2011; Cohen, 1992; Hemphill, 2003) when correlated with other measures theorized to be related to therapist adherence and competence, respectively. The specific item comparisons are listed in Table 2 along with a brief rationale for why the specific TPOCS-RS item and scale was selected.

Conversely, divergent validity of the measures is supported if each measure is only minimally correlated with measures of theoretically unrelated constructs or behaviors. For the purposes of this evaluation, the cut-off for divergent validity will be equal to the mean of six correlations derived from the (a) *Cognitive* Scale and *Behavioral* Scale of the TPOCS-S (scales that are theoretically similar to the CBAY-A and CBAY-C) and (b) *Psychodynamic* Scale, *Family* Scale, and *Client-Centered* Scale of the TPOCS-S (scales which the CBAY-A and CBAY-C are theoretically unrelated to; i.e., $r < .193$; McLeod & Weisz, 2010). This value is roughly equal to the midpoint of what is considered a "small" correlation (i.e., .10 to .30; Cohen, 1992) and the 33rd percentile mark of correlation coefficients found in published assessment, treatment, and meta-analytic literature (Hemphill, 2003). The rationale for using the mean of the correlations between these two classes of scales is that, while it is possible that the

Table 2.

Comparisons for Evaluating Construct Validity of the CBAY-A

CBAY-A Item	*TPOCS-RS Item*	*Rationale*
Agenda Setting	Session Goals item	Both CBAY-A Agenda Setting and TPOCS-RS Session Goals are designed to capture the establishment and review of session goals. They are different in that TPOCS-S Session Goals is coded for non-CBT related goals whereas CBAY-A Agenda Setting is only coded when the goal is CBT for youth anxiety related.
HW Review	Homework item	The CBAY-A HW Review item focuses on the therapist's efforts to review homework with the youth while the TPOCS-RS Homework item includes homework review and homework assigned.
HW Assigned	Homework item	The CBAY-A HW Assigned item focuses on the therapist's efforts to assign homework with the youth and encourage completion of homework while the TPOCS-RS Homework item includes homework review and homework assigned.
Rapport Bldg	N/A	There are no corresponding items on the TPOCS-RS that can be used to evaluate the CBAY-A Report Building item.
Psychoed-Anx	Cognitive Education item	The CBAY-A Psychoed-Anx item measures the delivery of information about anxiety and treatment (e.g., Normalization of anxiety, components of anxiety). The TPOCS-RS Cognitive Education item addresses the cognitive model of anxiety.
Emotion Ed	Cognitive Education item	The CBAY-A Emotion Education item focuses on the identification of emotions, particularly anxiety, while the TPOCS-RS Cognitive Education item includes a focus on the relationship between sensations and anxiety.
Fear Ladder	Respondent item	The CBAY-A Fear Ladder item focuses on the development of a hierarchy of feared stimuli, this is a component of the TPOCS-RS Respondent item.
Relaxation	Relaxation Item	Both the CBAY-A Relaxation item and the TPOCS-RS Relaxation Strategies item measure teaching and encouraging relaxation practices.
Cognitive-Anx	Cognitive Distortion item (TPOCS-RS) Cognitive Scale (TPOCS-RS)	The CBAY-A Cognitive-Anx item measures teaching about the role of cognition in anxiety and corresponds to the TPOCS-RS Cognitive Distortion item. The Cognitive Scale include items that address the role of thoughts in anxiety, the connection between thoughts and feelings, identifying thoughts and thinking errors, and how to alter thoughts to manage anxiety. All of these items relate to the CBAY-A Cognitive-Anx item.
Problem Solving	Coping Skills (TPOCS-RS)	The CBAY-A Problem Solving item focuses on the teaching and practice of a multi-step problem solving strategy, and this is also a component of the TPOCS-RS Coping Skills item.

Table 2 continued

CBAY-A	*TPOCS-RS Item*	*Rationale*
Self-Reward	Operant Strategies (TPOCS-RS)	The CBAY-A Self-Reward and TPOCS-RS Operant Strategies items measure when the therapist's focus is on using rewards to reinforce approach and coping behaviors.
Coping Plan	Cognitive Scale (TPOCS-RS) Behavioral Scale (TPOCS-RS)	The CBAY-A Coping Plan item measures focuses on teaching or practice of a multi-step coping approach and is similar to many of the items on the TPOCS-RS Cognitive Scale and Behavioral Scale.
Exposure: Prep	Respondent item	The CBAY-A Exposure: Prep item focuses on preparation for conducting an exposure. The TPOCS-RS Respondent Strategies item similarly focuses on preparing for and conducting exposures.
Exposure	Behavioral Scale (TPOCS-RS)	The CBAY-A Exposure item focuses on the practice of conducting exposures. The Behavior Scale of the TPOCS-RS includes multiple items (e.g., Behavioral Focus and Respondent Strategies) that are measured by the CBAY-A Exposure item.
Exposure: Debrief	Operant Strategies item Respondent item	The CBAY-A Exposure: Debrief item focuses on learning occurring during exposure activities while the Operant Strategies item focuses includes instituting the principles of operant conditioning, a component of the exposure debrief. The TPOCS-S Respondent Strategies item also focuses on reviewing exposures.
Exposure: Debrief	Behavioral Scale (TPOCS-RS)	The Behavior Scale of the TPOCS-S includes multiple items (e.g., Behavioral Focus, Operant Strategies, and Respondent Strategies) that are measured by the CBAY-A Exposure: Debrief item.
Maintenance	N/A	There are no items on the TPOCS-S that correspond to the CBAY-A Maintenance item.
Didactic Teaching	N/A	There are no items on the TPOCS-S that correspond to the CBAY-A Didactic Teaching item.
Collaborative Teaching	N/A	There are no items on the TPOCS-S that correspond to the CBAY-A Collaborative Teaching item.
Modeling	Modeling Item	Both the CBAY-A and TPOCS-S Modeling items focus on the therapist demonstrating skills.
Rehearsal	Rehearsal Item	Both the CBAY-A and TPOCS-S Modeling items focus on the therapist encouraging the youth to practice skills.
Coaching	Coaching Item	Both the CBAY-A and TPOCS-S Modeling items focus on the therapist providing guidance to the youth as they practice a skill.
Self-disclosure	Self-Disclosure Item	Both the CBAY-A and TPOCS-S Modeling items focus on the therapist disclosing information about him/herself to facilitate the youth learning a skill.

Psychodynamic Scale, *Family* Scale, and *Client-Centered* Scale may pick up on certain common psychotherapy elements also present in the *Cognitive* Scale and *Behavioral* Scale, it is most likely that these three scales will represent a combination of non-CBT psychotherapy elements.

Evaluating the convergent validity of the CBAY-C is a more complex process than for the CBAY-A because it includes both a categorical component, evaluated in step one, and an interval component, evaluated in step two. First, to evaluate the extent to which the CBAY-C items are sensitive to the appropriate therapist behaviors, CBAY-C and CBAY-A scores will be recoded to reflect whether each measured therapist behavior was present (or 1) or absent (as 0). By recoding in this way, the measures are equivalent: scores will reflect whether the therapist engaged in a particular CBT for youth anxiety behavior or not.[9] Then, these scores will be compared using Cohen's Kappa (as noted previously, Kappa is appropriate for evaluating the rate of agreement between two ratings on an ordinal scale; Cohen, 1968).

Cohen's Kappa is most commonly used to evaluate inter-rater or alternate-form reliability when measures use a nominal scale (Brennen & Prediger, 1981; Cohen, 1968; Ludbrook, 2002). It is an improvement over percent agreement because Cohen's Kappa accounts for chance agreement. Therefore, to get a high Kappa value, the measure must result in a higher level of agreement than would be achieved by chance. Kappa can also be used to assess the validity of one test or measure when one test or measure is a criterion (i.e., a standard against which other measures should be compared; Brennen & Prediger, 1981). A concern of using Cohen's Kappa to evaluate validity is that a measure should not simply improve on chance agreement but should improve on the base rate, when that base rate is stable (Brennen & Prediger, 1981). For instance, if you knew that exposure occurred in 2/3 sessions, it would be important to ensure that whatever

[9] Three items on the CBAY-C (Within Session Focus, Across Session Focus, and Structure/Phase) do not have corresponding items of the CBAY-A and therefore will not be compared to a CBAY-A item.

level of prediction that you have is better than 2/3, not just better than chance agreement. However, the concern raised by Brennen and Prediger (1981) does not apply for two reasons. First, Brennen and Prediger's (1981) warning about using Cohen's Kappa when the base rate of the measured event (e.g., the delivery of a particular CBT for youth anxiety treatment element) is known does not apply since the rates are not known. Second, the rates of the measured events are variable. Therefore, Cohen's Kappa is a viable statistic for evaluating the construct validity of the dichotomous portion of the CBAY-C items.[10]

Second, the CBAY-C scores will be compared to the Focusing Treatment and Instigating Change domain-level items on the COMP-CF. Together, these two comparisons will evaluate the extent to which the CBAY-C is sensitive to the appropriate therapist behaviors and measures therapist competence.

The divergent validity, an aspect of construct validity, of the CBAY-A and CBAY-Care evaluated by comparing the relationship between scores on the two measures with scores that measure treatment elements unrelated to CBT for youth anxiety. To do this, each score will be compared to the *Client-Centered* Scale, *Family* Scale, and *Psychodynamic* Scale from the TPOCS-RS. The three scales from the TPOCS-RS each measure treatment elements consistent with treatment models other than CBT. It is hypothesized that the magnitude of each of these correlations will be less than or equal to .193, as discussed previously.

It is common for psychometric evaluation studies to evaluate the criterion-related validity of the measure in question. However, currently there are no gold-standard measures of adherence to or competence in CBT for youth anxiety, so evaluating the CBAY-A and CBAY-C

[10]While appropriate for evaluating the validity of the dichotomous portion of the CBAY-C, Kappa is not appropriate for evaluating the continuous portion of the CBAY-C.

in that way is not possible. Therefore, this study will evaluate a related form of validity, discriminative validity.

Discriminative validity refers to the degree to which a measure performs in predicted ways when used with different groups (Cicchetti, 1994; Foster & Cone, 1995; Hattie & Cooksey, 1984). This method for evaluating validity has been suggested when treatments have different characteristics (e.g., high versus low complexity of the treatment), therapists were trained differently (e.g., high versus low supervision to case load ratios), and therapists receive different supervision (e.g., extensive versus less training; Perepletchikova & Kazdin, 2005). If the scores are found to be significantly different from one another in the expected direction, then the discriminative validity of the measure is supported. To evaluate the discriminative validity of the CBAY-A and CBAY-C, scores from the CC, YAS-CC, and YAS-UC will be compared with the expectation that the CC scores will be significantly higher than the YAS-CC scores and the YAS-CC scores will be significantly higher than the YAS-UC.

There is a potential for analyses from this study to be significantly affected by the nesting structure of the data, especially when the analyses include between-group comparisons, as is the case with these sets of analyses. In this case, of interest is whether there are different levels of adherence between the three different study conditions. Since therapists and clients are nested within study condition (i.e., each therapist and client participate in only one of the three study condition), the variance associated with each is not evenly distributed to the three study conditions. Furthermore, it is possible that the therapists and clients in the three conditions are not from the same populations (i.e., they differ in important ways) and will exert systematic differences on the CBAY-A and CBAY-C scores.

To account for the possible effects of nested data, a mixed modeling approach can be used if it is determined that that the nested variables systematically influence the adherence or competence scores. However, this approach is only necessary if the nesting structure is determined to have a significant influence on the CBAY-A and CBAY-C scores (Taylor, 2010). To test influence of the nesting structure on the CBAY-A scores, ICC(1,1) were run between the scores and levels of the nesting structure (session number, youth, therapist, and condition; Field, 2005, 2013). These analyses were computed using Mplus statistical package. This procedure was then repeated for the CBAY-C data. These analyses essentially compare the within-group variance relative to the between-group variance (Field, 2013). Higher ICC(1,1) coefficients would indicate that the between-group variance is high relative to the within-group variance and would indicate that the level of the nesting structure being evaluated has a higher influence on the scores and needs to be accounted for in further analyses (Field, 2013; Taylor, 2010). If these ICC(1,1) coefficients were shown to be high (greater than 0.1; Taylor, 2013) then a mixed modeling approach would have been used to evaluate the discriminative validity of the CBAY-A and CBAY-C.

Tables 3 and 4 show the ICC(1,1) for the item scores on the CBAY-A and CBAY-C respectively. Since many of the ICC(1,1) for the item scores on the CBAY-A are greater than 0.1, a mixed method approach is used. Specifically, variance component analyses were used to compute the estimated mean and standard error for the item scores on the CBAY-A. These values were then used to compute Student's *t*-test. On the other hand, since the ICC(1,1) for the item scores on the CBAY-C are almost all less than 0.1 (the Maintenance scores at the Session Number level and the Homework Assigned at the Youth level were the only two ICC(1,1) greater than 0.1) Student's *t*-test were run using the raw data.

Relationship between CBAY-A and CBAY-C. The final hypothesis was that the CBAY-A and CBAY-C are unique measures. To test this hypothesis, the correlation between the CBAY-A and CBAY-C was evaluated. The hypothesis that the CBAY-A and CBAY-C each measure unique variance is supported if the magnitude of the correlation is less than .70 (see Tabachnick & Fidell, 2007).

Table 3.

Evaluating CBAY-A Nesting Structure

CBAY-A Item	Session number ICC(1,1)	Youth ICC(1,1)	Therapist ICC(1,1)	Condition ICC(1,1)
Agenda Setting	0.009	0.313	0.313	0.384
HW Review	0.062	0.378	0.371	0.486
HW Assigned	0.067	0.328	0.323	0.442
Rapport Bldg.	0.092	0.131	0.080	0.042
Psyched-Anxiety	0.112	0.081	0.075	0.114
Emotion Ed	0.471	0.023	0.041	0.080
Fear Ladder	0. 069	0.136	0.126	0.170
Relaxation	0. 233	-0.015	0.010	0.053
Cognitive- Anxiety	0.250	0.041	0.041	0.071
Problem Solving	0.223	0.017	0.006	0.038
Self-Reward	0.242	-0.035	-0.008	0.037
Coping Plan	0.141	0.302	0.294	0.403
Exposure: Prep	0.359	0.157	0.158	0.278
Exposure	0.233	0.100	0.092	0.180
Exposure: Debrief	0.242	0.121	0.116	0.204
Maintenance	0.162	0.021	0.016	0.008
Didactic teaching	0.182	0.287	0.273	0.382
Collaborative teaching	0.107	0.448	0.451	0.652
Modeling	0.144	0.231	0.239	0.204
Rehearsal	0.123	0.443	0.450	0.679
Coaching	0.013	0.080	0.038	0.040
Self-disclosure	0.100	0.153	0.141	0.084

Table 4.

Evaluating CBAY-C Nesting Structure

CBAY-C Item	Session number	Youth	Therapist	Condition
Within Session Focus	-0.023	0.070	0.068	0.061
Across Session Focus	-0.023	0.070	0.068	0.061
Structure/Phase	-0.024	0.066	0.065	0.057
Homework Review	-0.015	0.106	0.069	0.064
Homework Assigned	-0.015	0.089	0.055	0.040
Psychoed-Anx	0.014	0.003	0.014	0.020
Emotion Ed	0.050	0.027	0.023	0.017
Fear Ladder	-0.015	0.075	0.056	0.019
Relaxation	-0.016	0.026	0.021	0.003
Cognitive-Anx	-0.019	0.055	0.052	0.031
Problem Solving	-0.021	0.054	0.002	0.007
Self-Reward	-0.023	-0.001	-0.011	0.006
Coping Plan	0.000	0.033	0.039	0.076
Exposure: Prep	-0.015	0.030	0.006	0.006
Exposure	-0.021	0.019	-0.003	0.006
Exposure: Debrief	-0.011	0.012	-0.006	0.005
Maintenance	0.111	-0.007	-0.010	0.002
Didactic teaching	0.046	0.008	0.002	0.031
Collaborative teaching	0.069	0.041	0.032	0.040
Modeling	-0.008	0.024	0.030	0.061
Rehearsal	0.021	0.053	0.049	0.086
Coaching	0.011	0.005	0.004	0.036
Self-disclosure	-0.011	0.045	0.046	0.054

Results

Results of the psychometric evaluation of the CBAY-A and the CBAY-C are presented in the following sections: (a) data preparation, (b) missing data, (c) item performance, (d) reliability statistics, (e) construct and criterion-related validity statistics, and (f) correlations between the two measures (uniqueness analysis). Since the data preparation steps and missing data analyses were the same for both the CBAY-A and CBAY-C, that section will address both measures. The final section will cover the uniqueness of the CBAY-A and CBAY-C measures.

Data Preparation

After clean data sets were created (see Data Entry and Data Management), the lead project investigators, with support from graduate students familiar with the original studies, conducted further data preparation. This included: (a) matching session identifiers to youth identifiers and therapist identifiers, (b) calculating session numbers, (c) removing sessions that did not meet inclusion criteria, and (d) creating datasets on youth and therapist characteristics (e.g., total time in treatment, total number of sessions attended, and diagnosis) that were linked by youth identifiers and therapist identifiers. Data sets were merged as appropriate, and summary scores were computed using SPSS syntax.

Missing Data

Analysis of variance (ANOVA) indicated that the rate of the sessions held per youth and sessions coded per youth were not significantly different for the three study conditions (see Table 5). Overall, 67.40% of the sessions were coded.

To evaluate the pattern of values for missing data (i.e., uncoded sessions), two versions of Little's test were run. The first one included all session numbers (1-53) and the second included only sessions 1-20. The second variation of Little's test was run because the majority (94.4%) of youth received 20 or fewer sessions (corresponding to 97.7% of coded sessions) and this analysis would exclude sessions that were only held by a minority of therapist-youth dyads. In both cases, Little's test was non-significant: χ^2 (1125, N= 89) = 1079.67, p= .83 and χ^2 (1103, N= 89) = 1088.80, p= .61 respectively. These results indicate that the pattern of data is consistent with being missing completely at random (MCAR), and that the uncoded sessions are presumably unrelated to any of the variables included in analysis. With this added confidence to support that

Table 5.

Rate of Coded Sessions

	CC	YAS-CC	YAS-UC	Total	*F*-statistic*	*p*-value
Sessions held	812	286	330	1428		
Sessions coded	532	212	210	954		
Mean session held per youth(SD)	15.92 (1.43)	16.82 (5.02)	15.71 (9.34)	16.04 (5.07)	.256	.775
Mean sessions coded per youth (SD)	10.43 (2.84)	12.47 (4.61)	10.00 (6.01)	10.72 (4.17)	1.974	.145
Mean % coded per youth (SD)	65.5(17.4)	74.3 (15.8)	66.4 (22.7)	67.4 (18.2)	1.491	.231

*Note(s):** Degrees of freedom for each F-statistic (2, 88).

the missing data are MCAR, confidence in the results derived from these data increases (i.e., the data missing is likely independent of observed and unobserved parameters).

Adherence to CBT for Youth Anxiety

CBAY-A: Item Performance. The next step in evaluating the CBAY-A was to describe the performance of the individual items (Table 6). When examining all 954 sessions, the majority of items showed values using nearly the full range of the scale, with a range of 5.5 (five instances with a scores ranging from 1-6.5) or six (14 instances with scores ranging from 1-7). While the majority of items used nearly the full scale, three items did not: Psychoeducation Anxiety had scores ranging from 1-5.5; Maintenance had scores ranging from 1-4.5; and Coaching had scores ranging from 1-4.5.It is noteworthy that Coping Cat does not specifically instruct therapists to provide extensive psychoeducation about anxiety, discuss maintaining treatment gains following the termination of treatment, or provide feedback in a manner that would be captured by the Coaching item.

While the range of scores on the majority of the CBAY-A is very promising, the normality or distribution of those scores is also important to assess. The Shapiro-Wilk tests were significant ($p<.05$) for every item, indicating that the distribution of data for every CBAY-A

Table 6.

CBAY-A Item Performance

CBAY-A Item	Mean of two coders: N=954		
	Range	*Mean* (*SD*)	*Mean* (*SD*)*
Agenda Setting	1 – 6.5	2.07 (1.18)	3.78 (1.00)
Homework Review	1 – 7	2.67 (1.70)	4.13 (1.10)
Homework Assigned	1 – 7	2.44 (1.61)	4.06 (1.04)
Rapport Building	1 – 6.5	1.33 (0.94)	4.09 (1.03)
Psychoeducation Anxiety	1 – 5.5	1.35 (0.67)	2.96 (0.76)
Emotion Education	1 – 7	1.65 (1.42)	4.73 (1.57)
Fear Ladder	1 – 7	1.55 (1.07)	4.01 (1.10)
Relaxation	1 – 7	1.43 (1.17)	4.39 (1.46)
Cognitive Anxiety	1 – 7	1.52 (1.24)	4.58 (1.54)
Problem Solving	1 – 7	1.17 (0.79)	5.06 (1.26)
Self-Reward	1 – 6.5	1.26 (0.92)	4.43 (1.35)
Coping Plan	1 – 6.5	2.25 (1.48)	4.04 (1.14)
Exposure: Prep	1 – 7	1.84 (1.60)	4.58 (1.21)
Exposure	1 – 7	1.58 (1.34)	4.46 (1.17)
Exposure: Debrief	1 – 7	1.48 (1.06)	3.57 (1.09)
Maintenance	1 – 4.5	1.05 (0.29)	3.22 (0.75)
Didactic Teaching	1 – 6.5	2.13 (1.21)	3.44 (0.91)
Collaborative Teaching	1 – 7	2.96 (1.67)	4.51 (1.14)
Modeling	1 – 7	1.78 (1.21)	3.91 (1.06)
Rehearsal	1 – 7	3.23 (2.06)	4.91 (1.17)
Coaching	1 – 4.5	1.08 (0.33)	2.80 (0.75)
Self-Disclosure	1 – 7	1.32 (0.74)	3.25 (1.10)

Note(s): *Mean and standard deviation based on data that does not include Not-Present data.

item was non-normal. The Shapiro-Wilk test of normality was then run separately for each of the three study conditions, to assess normality of the scores within each condition. Again, the 22 CBAY-A items were each found to be non-normal. These results indicate that for future analyses using the CBAY-A measure, analyses that do not assume item normality should be used whenever possible. Results based on analyses that do assume item normality should be interpreted with caution. Consequently, the correlations used in the construct validity,

discriminative validity, and uniqueness analyses of this dissertation use a Spearman Rho correlation since normality is not an assumption for Spearman Rho.

In summary, the range of scores for the CBAY-A scores is encouraging, as it demonstrates that the full range of scores is possible. Furthermore, it is encouraging that the few items that have a more restricted range (e.g., Psychoeducation Anxiety, Maintenance, and Coaching) are items that are conceptually important for a measure of adherence to CBT for youth anxiety to include but are not specifically prescribed in the Coping Cat manual, which may have contributed to their restricted range. The lack of normality for all items means that statistical analyses conducted that assume that scores are normally distributed should be interpreted with caution.

CBAY-A: Reliability. The next step in the psychometric evaluation of the CBAY-A was to examine the inter-rater reliability of the items. These data are presented in Table 7. However, ICC is, at its core, a correlation, and based on the assumption that the data are normally distributed. Since the data used for the ICC are scores and since these scores were found to violate assumptions of analyses, the validity of these analyses is questionable. Despite this concern, the decision to use ICC will be maintained due to the lack of viable alternatives (i.e., a review of the literature did not identify any nonparametric tests of intra-class correlation).

Overall, the interrater reliability of the CBAY-A appears strong, with ICC(2,2) coefficients ranging from .43 to .93 (*M* = .77; *SD* = 0.15, *Median*= .81). Three items' ICC coefficients (Psychoeducation Anxiety, Maintenance, and Coaching) are considered 'Fair', six are considered 'Good' (Agenda Setting, Coping Plan, Didactic Teaching, Collaborative Teaching, Modeling, Self-Disclosure), and 13 are considered 'Excellent' (Homework Review,

Table 7.

CBAY-A Inter-rater Reliability

CBAY-A Item	*Mean (SD)* Coder 1	*Mean (SD)* Coder 2	ICC	95% CI
Adherence: Agenda Setting	2.13 (1.32)	2.00 (1.46)	0.62	.57 - .67
Adherence: Homework Review	2.53 (1.62)	2.81 (2.01)	0.85	.83 - .87
Adherence: Homework Assigned	2.26 (1.47)	2.61 (2.00)	0.81	.79 - .83
Adherence: Rapport Building	1.27 (0.79)	1.39 (1.22)	0.80	.78 - .83
Adherence: Psychoeducation Anxiety	1.46 (0.96)	1.25 (0.65)	0.49	.42 - .55
Adherence: Emotion Education	1.59 (1.51)	1.71 (1.44)	0.92	.91 - .93
Adherence: Fear Ladder	1.75 (1.35)	1.35 (0.99)	0.78	.75 - .81
Adherence: Relaxation	1.48 (1.31)	1.38 (1.11)	0.92	.91 - .93
Adherence: Cognitive Anxiety	1.55 (1.37)	1.50 (1.28)	0.86	.84 - .88
Adherence: Problem Solving	1.19 (0.87)	1.14 (0.76)	0.93	.92 - .94
Adherence: Self Reward	1.30 (1.01)	1.23 (0.90)	0.91	.90 - .92
Adherence: Coping Plan	2.12 (1.63)	2.38 (1.73)	0.71	.67 - .75
Adherence: Exposure: Prep	1.91 (1.74)	1.77 (1.56)	0.93	.92 - .94
Adherence: Exposure	1.65 (1.50)	1.50 (1.29)	0.91	.90 - .92
Adherence: Exposure: Debrief	1.53 (1.16)	1.42 (1.09)	0.86	.84 - .88
Adherence: Maintenance	1.06 (0.39)	1.04 (0.30)	0.52	.46 - .58
Adherence: Didactic Teaching	2.44 (1.57)	1.82 (1.14)	0.73	.69 - .76
Adherence: Collaborative Teaching	3.77 (2.15)	2.15 (1.63)	0.69	.65 - .73
Adherence: Modeling	1.66 (1.34)	1.89 (1.38)	0.74	.71 - .77
Adherence: Rehearsal	3.25 (2.18)	3.21 (2.15)	0.89	.88 - .90
Adherence: Coaching	1.07 (0.40)	1.10 (0.43)	0.43	.36 - .50
Adherence: Self-Disclosure	2.13 (1.32)	1.41 (0.92)	0.71	.67 - .74

Homework Assigned, Rapport Building, Emotion Education, Fear Ladder, Relaxation, Cognitive Anxiety, Problem Solving, Self-Reward, Exposure: Prep, Exposure, Exposure: Debrief, and Rehearsal).

CBAY-A: Construct Validity. The results of the convergent validity analyses were supportive of the validity of the CBAY-A as a measure of therapist adherence to CBT for youth anxiety (see Table 8). All 22 correlation coefficients were greater than .30 (M=0.56, SD=0.16). All of the correlation

Table 8.

CBAY-A: Construct Validity

CBAY-A Item	TPOCS-RS Item	Convergent validity: Correlation with corresponding TPOCS-RS Variable	Divergent validity: Correlation with Non-CBT*		
			Psychodynamic Scale	Family Therapy Scale	Client-Centered Scale
Agenda Setting	Session Goals Item	.424	-0.347	-0.339	0.142
HW Review	Homework Item	.789	-0.322	-0.519	0.270
HW Assigned	Homework Item	.755	-0.304	-0.505	0.239
Rapport Building	**	**	-0.086	-0.199	0.104
Psychoeducation-Anx	Cognitive Education Item	.379	-0.194	-0.068	-0.003
Emotion Ed	Cognitive Education item	.512	-0.146	-0.197	0.042
Fear Ladder	Respondent item	.421	-0.242	-0.104	0.061
Relaxation	Relaxation Item	.669	-0.123	-0.120	0.038
Cognitive-Anxiety	Cognitive Distortion item	.525	-0.141	-0.220	0.104
Cognitive-Anxiety	Cognitive Scale	.481			
Problem Solving	Coping Skills	.321	-0.083	-0.109	0.107
Self-Reward	Operant Strategies item	.383	-0.083	-0.034	0.018
Coping Plan	Cognitive Scale	.703	-0.291	-0.261	0.183
Coping Plan	Behavioral Scale	.585			
Exposure: Prep	Respondent item	.773	-0.194	-0.317	0.185
Exposure	Behavioral Scale	.552	-0.143	-0.247	0.139
Exposure: Debrief	Operant Strategies item	.440	-0.175	-0.278	0.200
Exposure: Debrief	Respondent item	.733			
Exposure: Debrief	Behavioral Scale	.622			
Maintenance	**	**	0.001	0.057	0.050
Didactic teaching	**	**	-0.308	-0.187	0.035
Collaborative teaching	**	**	-0.420	-0.543	0.236
Modeling	Modeling Item	.620	-0.282	-0.357	0.207
Rehearsal	Rehearsal Item	.861	-0.426	-0.576	0.283
Coaching	Coaching Item	.309	-0.064	-0.137	0.083
Self-disclosure	Self-Disclosure Item	.503	-0.175	-0.269	0.174

*Note(s): **: Only one analysis per CBAY-A item was run to evaluate divergent validity. Therefore, some cells are left blank

**: Not applicable because not corresponding item from the TPOCS is available.

coefficients were positive, with Rehearsal having the greatest magnitude (r=.861; corresponding with the TPOCS-RS Rehearsal item) and Coaching having the least magnitude (r=.309; corresponding with the TPOCS-RS Coaching item). Among the model items, the mean correlation coefficients were 0.54 (SD=0.14).

CBAY-A: Divergent Validity. The divergent validity analyses consisted of 66 correlation coefficients (22 CBAY-A items and the Psychodynamic, Family Therapy, and Client Centered subscale scores on the TPOCS-RS). When looking at the 30 correlation coefficients associated with the four Standard Items (Agenda Setting, Homework Reviewed, Homework Assigned, and Rapport Building) and the six Delivery Method Items (Didactic Teaching, Collaborative Teaching, Modeling, Rehearsal, Coaching, and Self-Disclosure), 20 of Spearman correlation coefficients were greater than the criterion of .193. When looking at the 36 correlation coefficients associated with the 12 Model Items, 11 are greater than the criterion of .193. Three Model Items (Coping Plan, Exposure: Prep, and Exposure: Debrief) had two of three correlations greater than .193.These results indicate that the CBAY-A is generally inversely related to the Psychodynamic Scale (mean r = -0.21; SD = 0.12) and Family Therapy Scale (mean r = -0.25; SD = 0.17) and positively associated with the Client-Centered Scale (mean r = 0.13; SD = 0.09).

CBAY-A: Discriminative Validity. Many of the ICC(1,1) coefficients between the CBAY-A scores and the levels of the nesting structure (Session Number, Youth, Therapist, and Study Condition; presented in Table 7) were greater than 0.1, thus requiring a statistical procedure that accounts for the nesting structure. Estimated means and standard errors, computed using variance component analyses, are presented in Table 9 along with Student's t-tests based on those data.

Table 9.

Discriminative Validity of CBAY-A Scores

	CC	CC vs. YAS-CC T-TESTS			YAS-CC	YAS-CC vs. YAS-UC *t*-tests			YAS-UC
	Mean (SE)	*t*-statistic	*df*	*p*-value	*Mean (SE)*	*t*-statistic	*df*	*p*-value	*Mean (SE)*
CBAY-A Item	N=532				N=212				N=210
Agenda Setting	2.51 (0.04)	7.05	742.00	<.001	1.93 (0.07)	8.42	211.00	<.001	1.10 (0.07)
HW Review	3.32 (0.06)	5.47	742.00	<.001	2.69 (0.10)	12.11	211.00	<.001	1.02 (0.10)
HW Assigned	3.08 (0.06)	8.26	742.00	<.001	2.16 (0.09)	8.01	211.00	<.001	1.09 (0.09)
Rapport Building	1.42 (0.04)	0.4	742.00	0.6879	1.39 (0.06)	3.91	211.00	<.001	1.04 (0.06)
Psychoed-Anx	1.50 (0.03)	4.7	742.00	<.001	1.27 (0.04)	3.41	211.00	0.0007	1.07 (0.04)
Emotion Ed	1.83 (0.01)	0.42	742.00	0.6765	1.79 (0.08)	6.35	211.00	<.001	1.06 (0.08)
Fear Ladder	1.83 (0.04)	5.86	742.00	<.001	1.35 (0.07)	3.46	211.00	0.0006	1.01 (0.07)
Relaxation	1.59 (0.05)	2.08	742.00	0.0375	1.40 (0.08)	3.2	211.00	0.0015	1.05 (0.08)
Cognitive-Anx	1.71 (0.05)	1.84	742.00	0.0668	1.53 (0.08)	4.08	211.00	<.001	1.05 (0.08)
Problem Solving	1.28 (0.03)	3.46	742.00	0.0006	1.06 (0.05)	0.79	211.00	0.4318	1.00 (0.05)
Self-Reward	1.36 (0.04)	1.22	742.00	0.2219	1.27 (0.06)	2.95	211.00	0.0034	1.01 (0.06)
Coping Plan	2.81 (0.05)	7.37	742.00	<.001	2.06 (0.09)	8.61	211.00	<.001	1.01 (0.09)
Exposure: Prep	2.41 (0.05)	12.3	742.00	<.001	1.22 (0.08)	1.9	211.00	0.0586	1.00 (0.08)
Exposure	1.97 (0.05)	8.77	742.00	<.001	1.17 (0.08)	1.56	211.00	0.12	1.00 (0.08)
Exposure: Debrief	1.80 (0.04)	9.9	742.00	<.001	1.13 (0.06)	1.54	211.00	0.13	1.00 (0.06)
Maintenance	1.05 (0.01)	-1.78	742.00	0.0751	1.09 (0.02)	2.06	211.00	0.0096	1.02 (0.02)
Didactic teaching	2.57 (0.04)	7	742.00	<.001	2.02 (0.07)	9.55	211.00	<.001	1.12 (0.07)
Collaborative teaching	3.72 (0.06)	7.88	742.00	<.001	2.89 (0.09)	14.09	211.00	<.001	1.11 (0.09)
Modeling	2.10 (0.05)	4.56	742.00	<.001	1.70 (0.07)	6.27	211.00	<.001	1.04 (0.07)
Rehearsal	4.28 (0.07)	12.11	742.00	<.001	2.73 (0.11)	10.89	211.00	<.001	1.06 (0.11)
Coaching	1.13 (0.01)	2.7	742.00	0.0071	1.06 (0.02)	1.93	211.00	0.0539	1.00 (0.02)
Self-disclosure	1.43 (0.03)	1.06	742.00	0.2896	1.37 (0.05)	5.3	211.00	<.001	1.01 (0.05)

*Note(s):**: Indicates significant difference after using the Holm procedure to correct for family-wise error.

The results[11] in Table 9 present findings comparing the CBAY-A scores between the different study conditions. For 21 of 22 item comparisons (97.73%) the CBAY-A scores were greater for the therapists in the CC condition than the therapists in the YAS-CC condition. All CBAY-A scores were higher for therapists in the YAS-CC condition than therapists in the YAS-UC condition. Furthermore, CC therapists received significantly higher scores than YAS-CC therapists for 15 of 22 items (68.18%), and seven of 12 Model Items (58.33%). For every item, the YAS-CC therapists had scores significantly greater than the YAS-UC therapists.

In summary, these data suggest that some feature of the CC study resulted in higher CBAY-A score than found in the YAS-CC condition. Similarly, the features of the YAS-CC study condition resulted in higher CBAY-A score than found in the YAS-UC condition.

Competence in CBT for Youth Anxiety

The following sections focus on the CBAY-C. Where applicable, issues relevant to the CBAY-A will be referenced, as the methodology used to evaluate the CBAY-C is largely similar to that of the CBAY-A.

CBAY-C: Item Performance. This section will focus on the item performance of the CBAY-C. Table 10 presents the item performance data for the CBAY-C including: the number of sessions for which data was coded as present by both coders; the number of sessions for which data is non-present due to disagreement; and the range, mean, and standard deviation. Shapiro-Wilk statistics were calculated to evaluate the normality of the scores for each item.

The first noteworthy issue is the variable number of data points included in the CBAY-C analyses. Out of a total of 954 sessions coded, the number times that each behavior was coded as

[11] Since analyses for this hypothesis used significance testing rather than comparing results to a criterion value, the Holm procedure was used to control for family-wise error (see Holm, 1979; Holland & Copenhaver, 1988).

Table 10.

CBAY-C Descriptive Statistics

CBAY-C Item	*Mean* of two coders (*N* = 954)			
	N Coded as Present**	*N* Non-present due to disagreement (*SD*)	Range	*Mean (SD)*
Within Session Focus	696	59 (6.2%)	1.0 – 7.0	4.60 (1.39)
Between Session Focus	695	60 (6.3%)	1.0 – 7.0	4.53 (1.33)
Structure/Phase	694	61 (6.4%)	1.0 – 7.0	4.51 (1.41)
Homework Review	579	70 (7.3%)	1.0 – 6.5	4.33 (1.31)
Homework Assigned	531	89 (9.3%)	1.0 – 7.0	4.25 (1.40)
Psychoeducation-Anx	74	137 (14.4%)	1.5 – 7.0	3.85 (1.02)
Emotion Education	104	58 (6.1%)	1.5 – 7.0	4.81 (1.28)
Fear Ladder	82	80 (8.4%)	1.5 – 6.5	4.45 (1.05)
Relaxation	102	34 (3.6%)	1.5 – 7.0	4.87 (1.16)
Cognitive Anxiety	68	89 (9.3%)	1.5 – 7.0	4.81 (1.41)
Problem Solving	30	20 (2.1%)	3.0 – 7.0	5.40 (0.94)
Self-Reward	48	23 (2.4%)	2.5 – 6.5	5.16 (0.98)
Coping Plan	304	170 (17.8%)	1.5 – 6.5	4.75 (1.05)
Exposure: Prep	229	24 (2.5%)	1.0 – 7.0	4.90 (1.09)
Exposure	156	32 (3.4%)	1.5 – 7.0	4.92 (1.09)
Exposure: Debrief	194	33 (3.5%)	2.0 – 7.0	4.67 (1.11)
Maintenance	9	22 (2.3%)	4.0 – 5.5	4.67 (0.61)
Didactic Teaching	391	176 (18.4%)	1.0 – 7.0	4.48 (1.19)
Collaborative Teaching	420	185 (19.4%)	1.5 – 7.0	4.81 (1.28)
Modeling	220	142 (14.9%)	2.0 – 7.0	5.11 (0.85)
Rehearsal	434	129 (13.5%)	1.5 – 7.0	5.01 (1.11)
Coaching	53	78 (8.2%)	3.0 – 6.0	4.86 (0.68)
Self-Disclosure	137	122 (12.8%)	2.0 – 7.0	4.82 (1.00)

*Note(s):*** The number of items coded as non-present equals 954 (the total number of sessions coded) minus the sum of the number of items coded as present, plus the number of items non-present due to disagreement.

present ranged from 696 (Agenda Setting) to nine (Maintenance) sessions. Next, of the 23 items, 19 had score ranges greater than or equal to five (the maximum possible range is six; 1 – 7), demonstrating that, for the vast majority of items, the coders observed both high and low competence. Exceptions were: Problem Solving and Self-Reward, which had ranges of four; Coaching, which had a range of three; and Maintenance, which had the lowest range, one-and-a-

half. These items with restricted ranges also corresponded to the least-observed therapist behaviors, being coded as present relatively infrequently in the dataset (e.g., ranging from 53 times [Coaching] to as few as nine times [Maintenance]). The fact that the items with the lowest rates of observance were also the same items that did not approach using the full range of scores leaves open the possibility that these items could still be coded using the full range of scores, if coded with a set of sessions where the behaviors were more common.

The mean values of the items were generally in the middle of scale, ranging from a low of 3.85 (Psychoeducation Anxiety) to a high of 5.40 (Problem Solving) with standard deviations ranging from a low of 0.61 (Maintenance) to 1.41 (Homework Review). There also seems to be a relationship between the number of times an item was coded as present and the standard deviation of the mean. Of the five items with standard deviations less than 1.00, four were the items listed above with restricted ranges and low rates of being coded as present. The Shapiro-Wilk tests were significant for all variables, indicating that the items are not normally distributed and that non-parametric analyses should be used whenever possible.

CBAY-C: Reliability. The reliability analyses for the CBAY-C included first addressing the dichotomous aspect of the scale by computing Cohen's Kappa of the two coders, and then addressing the continuous aspect of the scale by calculating the ICC(2,2) of the two coders. To aid in the interpretation of Cohen's Kappa, Table 11 presents the rate of observation for each coder, Cohen's Kappa, and percentage agreement.

First, with regard to the between-coder agreement or whether the therapist engaged in the behavior measured by each item, and based on Cicchetti's guidelines for interpreting Cohen's Kappa (1994), four items (Psychoeducation Anxiety, Cognitive Anxiety, Maintenance, and

Table 11.

CBAY-C Reliability

CBAY-C Item	Coder A % Observed	Coder B % Observed	Cohen's Kappa	% Agreement	Mean (SD)* Coder 1	Mean (SD)* Coder 2	N**	ICC**	95% CI**
Within Session Focus	75.3	76.8	0.83	93.8	4.76 (1.61)	4.27 (1.57)	696	0.80	.77 - .83
Across Session Focus	75.2	76.8	0.83	93.7	4.63 (1.56)	4.25 (1.51)	695	0.77	.73 - .80
Structure/Phase	75.2	76.7	0.83	93.6	4.65 (1.62)	4.20 (1.56)	694	0.81	.79 - .84
Homework Review	63.2	65.5	0.84	92.7	4.37 (1.61)	4.03 (1.51)	579	0.74	.69 - .78
Homework Assigned	59.0	61.7	0.81	90.7	4.30 (1.66)	3.91 (1.61)	531	0.77	.72 - .80
Psychoeducation Anxiety	15.0	14.9	0.44	85.6	4.08 (1.27)	3.42 (1.19)	74	0.64	.42 - .77
Emotion Education	13.2	14.7	0.75	93.9	4.57 (1.52)	4.46 (1.44)	104	0.80	.70 - .86
Fear Ladder	13.2	12.4	0.62	91.6	4.44 (1.25)	3.97 (1.27)	82	0.69	.52 - .80
Relaxation	12.5	12.5	0.84	96.4	4.91 (1.26)	4.37 (1.52)	102	0.77	.66 - .84
Cognitive Anxiety	11.5	12.1	0.55	90.7	4.54 (1.62)	3.88 (1.78)	68	0.78	.64 - .86
Problem Solving	4.7	3.7	0.74	97.9	4.93 (1.47)	4.97 (1.29)	30	0.69	.35 - .85
Self-Reward	6.4	6.1	0.79	97.6	5.05 (1.04)	4.79 (1.36)	48	0.80	.65 - .89
Coping Plan	39.4	42.1	0.63	82.2	4.87 (1.30)	4.40 (1.24)	304	0.70	.63 - .76
Exposure: Prep	25.4	25.2	0.93	97.5	5.00 (1.30)	4.73 (1.24)	229	0.70	.61 - .77
Exposure	18.8	17.3	0.89	96.7	5.04 (1.25)	4.65 (1.30)	156	0.69	.57 - .77
Exposure: Debrief	21.7	22.4	0.90	96.5	4.87 (1.41)	4.29 (1.34)	194	0.56	.42 - .67
Maintenance	1.7	2.5	0.44	97.7	4.44 (1.15)	3.88 (1.19)	9	0.37	-1.79 - .86
Didactic Teaching	49.4	51.1	0.63	81.6	4.51 (1.43)	4.06 (1.39)	391	0.74	.68 – .78
Collaborative Teaching	58.7	48.7	0.61	80.6	4.84 (1.50)	4.45 (1.50)	420	0.74	.69 - .79
Modeling	29.1	31.9	0.65	85.1	5.14 (1.09)	4.69 (1.17)	220	0.55	.41 - .65
Rehearsal	53.6	50.9	0.73	86.5	4.99 (1.36)	4.76 (1.35)	434	0.70	.64 - .75
Coaching	8.7	10.6	0.53	91.8	5.05 (0.99)	4.45 (0.87)	53	0.55	.22 - .74
Self-Disclosure	19.1	22.4	0.61	87.2	4.93 (1.19)	4.28 (1.20)	137	0.67	.54 - .76
Mean Value	**32.6**	**33.0**	**0.71 (.10)**	**91.4**	**4.74 (1.37)**	**4.31 (1.37)**	**271.74**	**0.70 (.11)**	

Note(s): * Mean and Standard Deviation are calculated based on positively scored items only.
** Data based on items for which both coders provided positive ratings.

Coaching) had a Cohen's Kappa in the 'Fair' range. There were three *Model* Items and five *Delivery Method* Items (Fear Ladder, Problem Solving, Coping Plan, Didactic Teaching, Collaborative Teaching, Modeling, Rehearsal, and Self-Disclosure) in the 'Good' range. All five *Standard* Items and six *Model* Items (Within Session Focus, Across Session Focus, Structure/Phase, Homework Review, Homework Assigned, Emotion Education, Relaxation, Self-Reward, Exposure: Prep, Exposure, and Exposure: Debrief) were in the 'Excellent' range. The mean Cohen's Kappa value was in the 'Good' range and no items were in the 'Poor' range. Of the items in the 'Fair' range, Psychoeducation Anxiety is the most concerning, because it received a 'Fair' Cohen's Kappa value despite having a modest rate of observation (each observer Coded Psychoeducation in approximately 15% of sessions).

The ICC(2,2) values for the CBAY-C ranged from the 'Poor' range (ICC for Maintenance = 0.37) to the 'Excellent' range (ICC for Structure/Phase = 0.81), with three items the 'Fair' range, and 19 in the 'Good' or 'Excellent' ranges. The mean ICC(2,2) value for the CBAY-C was 'Good' (M = .70, SD = .11). Overall, these inter-rater reliability coefficients indicate that the CBAY-C can be coded reliably, with the possible exception of the Maintenance item. Definitive judgment of the reliability of the Maintenance item is premature at this point in time given that it was so rarely observed (N = 9).

CBAY-C: Construct validity. Analyses comparing the CBAY-C scores and CBAY-A scores, both recoded into a dichotomous form that reflects whether or not the therapist behaviors corresponding to each item were present, largely suggest that the CBAY-C items are sensitive to the therapist behaviors that they are designed to reflect. Of the 20 CBAY-C items for which there was a corresponding CBAY-A item, four had Cohen's Kappa coefficients in the 'Fair' range, four had Cohen's Kappa coefficients in the 'Good' range, and 11 had Cohen's Kappa

coefficients in the 'Excellent' range (see Table 12). This leaves one item, Maintenance, in the 'Poor' range. These findings contribute to the general trend of overall positive findings with a few weaker items.

The comparisons of the CBAY-C scores with the COMP-CF Focusing Treatment and Instigating Change domain-level items were also mostly supportive of the validity of the CBAY-C items (see Table 12). The Spearman Rho coefficients ranged from .240 (relationship between

Table 12.

CBAY-C Construct Validity

CBAY-C Item	CBAY-A Item	*N*	Cohen's Kappa	CBAY-C COMP-CF* *N*	COMP-CF Focusing Treatment	COMP-CF Instigating Change
Within Session Focus	N/A			696	0.526	0.578
Across Session Focus	N/A			695	0.514	0.524
Structure/Phase	N/A			694	0.522	0.555
Homework Review	HW Review	859	.896	579	0.406	0.420
Homework Assigned	HW Assigned	827	.845	531	0.379	0.388
Psychoeducation Anxiety	Psychoeducation Anxiety	728	.469	74	0.322	0.255
Emotion Education	Emotion Education	818	.723	104	0.570	0.525
Fear Ladder	Fear Ladder	809	.513	82	0.351	0.403
Relaxation	Relaxation	879	.893	102	0.355	0.393
Cognitive Anxiety	Cognitive Anxiety	800	.522	68	0.497	0.488
Problem Solving	Problem Solving	926	.775	30	0.443	0.337
Self-Reward	Self-Reward	908	.762	48	0.313	0.321
Coping Plan	Coping Plan	738	.772	304	0.466	0.432
Exposure: Prep	Exposure: Prep	919	.973	229	0.311	0.338
Exposure	Exposure	921	.931	156	0.240	0.294
Exposure: Debrief	Exposure: Debrief	908	.962	194	0.284	0.276
Maintenance	Maintenance	917	.326	9	0.678	0.847
Didactic Teaching	Didactic Teaching	703	.734	391	0.465	0.414
Collaborative Teaching	Collaborative Teaching	730	.735	420	0.439	0.455
Modeling	Modeling	713	.791	220	0.324	0.350
Rehearsal	Rehearsal	797	.824	434	0.405	0.408
Coaching	Coaching	843	.424	53	0.276	0.264
Self-Disclosure	Self-Disclosure	746	.684	137	0.315	0.327

*Note(s):*N/A: Not applicable because not corresponding item from the CBAY-A or TPOCS-RS is available.
* = COMP-CF CBT-Scale is derived from the mean of five items: (a) Structure and Pace, (b) Continuity of Treatment, (c) Focusing on Key Themes, (d) Uses Change Strategies Effectively, and (e) Facilitates Client's Participation.

Exposure and Focusing Treatment) to .847 (relationship between Maintenance and Instigating Change) and had a mean of 0.41 (*SD*=0.12). The CBAY-C Exposure, Exposure: Debrief, and Coaching items each had two of three Spearman Rho coefficients below the cutoff of 0.30 and Psychoeducation had one (reflecting the relationship with Instigating Change) below the cutoff.

CBAY-C: Divergent validity. Data reflecting the relationship between the CBAY-C items and the TPOCS-RS Psychodynamic, Family Therapy, and Client Centered Scale Scores are presented in Table 13. As was seen with the CBAY-A divergent validity analyses, the correlations between the CBAY-C Standard Item scores was highly related to the three non-CBT TPOCS-RS Scale Score (13 of 15 correlation coefficients were greater than the criterion of .193). The correlation coefficients reflecting the relationship between CBAY-C Delivery Method Items and the TPOCS-RS Scale Score were more varied (seven of 18 correlation coefficients were greater than the criterion of .193). Finally, 17 of the 36 correlation coefficients reflecting the relationship between the CBAY-C Model Items and TPOCS-RS non-CBT Scale Scores were greater than the criterion of .193. Emotion Education, Fear Ladder, Cognitive Anxiety, Problem Solving, Coping Plan, and Maintenance all had at least two of three correlation coefficients that were higher than hypothesized. These results indicate that the CBAY-C has a similar relationship with the non-CBT TPOCS-RS Scales as the CBAY-A: inversely related to the Psychodynamic Scale (mean r = -0.21; SD = 0.13) and Family Therapy Scale (mean r = -0.12; SD = 0.13) and positively associated with the Client-Centered Scale (mean r = 0.24; SD = 0.13).

Table 13.

CBAY-C Divergent Validity

CBAY-C Item	*N*	TPOCS-RS Psychodynamic Scale	TPOCS-RS Family Therapy Scale	TPOCS-RS Client-Centered Scale
Within Session Focus	696	-0.312	-0.301	0.285
Across Session Focus	695	-0.295	-0.223	0.279
Structure/Phase	694	-0.296	-0.287	0.244
Homework Review	579	-0.248	-0.160	0.284
Homework Assigned	531	-0.281	-0.078	0.258
Psychoed-Anx	74	-0.030	0.100	0.168
Emotion Ed	104	-0.454	-0.020	0.336
Fear Ladder	82	-0.201	-0.286	0.192
Relaxation	102	-0.072	0.208	0.155
Cognitive-Anx	68	-0.397	-0.181	0.270
Problem Solving	30	-0.498	-0.323	-0.181
Self-Reward	48	-0.031	-0.108	0.465
Coping Plan	304	-0.133	-0.227	0.279
Exposure: Prep	229	-0.132	-0.134	0.281
Exposure	156	-0.190	-0.016	0.174
Exposure: Debrief	194	-0.140	-0.117	0.225
Maintenance	9	-0.298	-0.269	0.502
Didactic teaching	391	-0.238	-0.146	0.175
Collaborative teaching	420	-0.226	-0.113	0.260
Modeling	220	-0.091	-0.056	0.195
Rehearsal	434	-0.218	-0.071	0.209
Coaching	53	-0.001	0.016	0.110
Self-disclosure	137	-0.130	-0.062	0.285

CBAY-C: Discriminative validity. The investigation of the discriminative validity of the CBAY-C will only focus on the two individual CBT study conditions (i.e., CC and YAS-CC) since so few YAS-UC therapists administered any CBT for youth anxiety treatment elements (nine observations were made of model items across the 12 model items and 210 coded sessions). Table 14 presents the data from a series of *t*-tests used to evaluate the discriminative validity of the CBAY-C. Twenty-three separate *t*-tests were run, one for each item, and the Holm procedure was used to control for family-wise error (see Holland & Copenhaver, 1988; Holm, 1979). In every instance, the mean significantly different in 21 of the 23 comparisons (while the score differences on Exposure:

Table 14.

CBAY-C Discriminative Validity

	CC *Mean (SD)*				YAS-CC *Mean (SD)*
CBAY-C Item	*N*=532	*t*-statistic	*df*	*p*-value	*N*=212
Within Session Focus	5.08 (1.05)	14.96*	240.20	<.001**	3.37 (1.37)
Across Session Focus	5.02 (0.95)	16.47*	239.35	<.001**	3.29 (1.25)
Structure/Phase	5.00 (1.08)	15.51*	249.08	<.001**	3.26 (1.33)
Homework Review	4.76 (0.96)	12.86*	192.62	<.001**	3.21 (1.33)
Homework Assigned	4.70 (1.09)	12.40*	187.18	<.001**	3.07 (1.36)
Psychoed-Anx	4.07 (0.92)	2.86	68.00	.006**	3.31 (0.98)
Emotion Ed	5.40 (0.86)	9.44	102.00	<.001**	3.53 (1.10)
Fear Ladder	4.71 (0.90)	5.25	80.00	<.001**	3.38 (0.97)
Relaxation	5.07 (1.05)	3.10	98.00	.002**	4.21 (1.22)
Cognitive-Anx	5.48 (0.95)	5.11	61.00	<.001**	4.03 (1.24)
Problem Solving	5.56 (0.81)	3.09	28.00	.004**	4.00 (1.00)
Self-Reward	5.44 (0.65)	2.98*	12.93	.011**	4.29 (1.29)
Coping Plan	5.04 (0.79)	9.09*	77.09	<.001**	3.62 (1.18)
Exposure: Prep	5.04 (0.94)	5.80*	17.33	<.001**	3.15 (1.32)
Exposure	5.06 (0.94)	3.54*	13.98	.003**	3.57 (1.54)
Exposure: Debrief	4.73 (1.06)	2.07*	11.77	.061	3.83 (1.47)
Maintenance	4.79 (0.64)	1.11	7.00	.305	4.25 (0.35)
Didactic teaching	4.88 (0.93)	10.97*	140.74	<.001**	3.50 (1.11)
Collaborative teaching	5.26 (0.96)	11.92*	157.15	<.001**	3.73 (1.21)
Modeling	5.27 (0.74)	6.45	216.00	<.001**	4.33 (0.86)
Rehearsal	5.33 (0.82)	11.19*	103.60	<.001**	3.82 (1.17)
Coaching	4.93 (0.64)	2.78	51.00	.008**	4.00 (0.71)
Self-disclosure	5.05 (0.82)	5.28*	29.96	<.001**	3.80 (1.12)

Note(s):: Levene's Test for the *t*-tests was significant meaning that equal variance cannot be assumed. Corresponding t-statistic and df used.
**: *t*-test was significant after correcting for multiple analyses using the Holm correction (Holm, 1979).

CBAY-C score was higher for the CC therapists than the YAS-CC therapists. The scores were Debrief and Maintenance were not significant). Overall, these data are consistent with the CBAY-A findings that therapists in the CC condition had higher scores on the CBAY-C than therapists in the YAS-CC condition.

Uniqueness: Correlations Between Adherence and Competence Measures

The final set of analyses focus on the uniqueness of the CBAY-A and the CBAY-C. As can be seen in Table 15, two of the *Standard* Items from the CBAY-A and three of the *Standard* Items from the CBAY-C did not have a corresponding item on the other measure and were

therefore not included in analyses. For the 20 correlations that were run, the mean Spearman Rho correlation coefficient was .44 (*SD* = .20), ranging from .01 (Coaching) to .74 (Psychoeducation Anxiety). These results are generally supportive of the uniqueness of the CBAY-A and the CBAY-C. However, two items (Psychoeducation Anxiety and Cognitive Anxiety) had scores above the threshold of .70 (.74 and .71 respectively). For these two items, concern remains that they are not measuring unique aspects of therapist behavior. Efforts to use these items from both the CBAY-A and the CBAY-C measures as predictor variables within future analyses could lead to spurious findings.

Table 15.

Correlations Between Corresponding CBAY-A and CBAY-C Items

CBAY-A Item	CBAY-C Item	*N*	Spearman Rho
Agenda Setting	--		
--	Within Session Focus		
--	Across Session Focus		
--	Structure/Phase		
HW Review	Homework Review	579	.485
HW Assigned	Homework Assigned	531	.552
Rapport Building	--		
Psychoeducation Anxiety	Psychoeducation Anxiety	74	.260
Emotion Education	Emotion Education	104	.741
Fear Ladder	Fear Ladder	82	.504
Relaxation	Relaxation	102	.624
Cognitive Anxiety	Cognitive Anxiety	68	.712
Problem Solving	Problem Solving	30	.624
Self-Reward	Self-Reward	48	.370
Coping Plan	Coping Plan	304	.363
Exposure: Prep	Exposure: Prep	229	.546
Exposure	Exposure	156	.455
Exposure: Debrief	Exposure: Debrief	194	.521
Maintenance	Maintenance	9	.123
Didactic Teaching	Didactic Teaching	391	.421
Collaborative Teaching	Collaborative Teaching	420	.478
Modeling	Modeling	220	.300
Rehearsal	Rehearsal	434	.549
Coaching	Coaching	53	.014
Self-Disclosure	Self-Disclosure	137	.118

Discussion

The purpose of this project was to evaluate the psychometric strength and uniqueness of two measures of treatment integrity: the Cognitive-Behavioral Therapy Adherence Scale for Youth Anxiety (CBAY-A) and the Competence with Cognitive-Behavioral Therapy for Youth Anxiety (CBAY-C).Results suggested that item scores from both measures show initial promise, as reflected in reliability and validity results. First, interrater reliability of the item scores of the CBAY-A and CBAY-C as measured by intraclass correlations were mostly in the 'Good' to 'Excellent' ranges, with the exception of a few items on each measure. These results support the reliability of the CBAY-A and CBAY-C. Similarly, the validity analyses of item scores for both measures supported the item score validity. That is, the data suggested that most item scores measured what they were intended to measure (CBAY-A items were highly correlated with corresponding items on an observational measure of treatment practices; CBAY-C items were highly correlated with an observational common factors competence measure) though their relationship to measures of theoretically unrelated behaviors were greater than expected. This may be due to the opportunity cost associated with engaging in therapy behaviors (i.e., some therapy behaviors are mutually exclusive). CBAY-A and CBAY-C scores were higher among therapists practicing in a research clinic and trained and supervised by the treatment developer than community-based therapists with trained and supervised by experts, and CBAY-A scores were higher among trained versus untrained community-based therapists. These results should be interpreted broadly and not be seen as an indictment of community therapists' ability to deliver high doses of quality CBT. And finally, the CBAY-A and CBAY-C item scores were not correlated with one another to a problematic degree, supporting the hypothesis is that they measure related but unique aspects of therapist behavior: quantity and quality of CBT for youth

anxiety delivered. Further interpretations of these findings in light of potential applications of the measures will be presented in subsequent sections.

Adherence

This psychometric evaluation of the CBAY-A focused on interrater reliability and various aspects of validity. As with most observational coding measures (e.g., Barber et al., 2006; Hogue et al, 2010), demonstrating that the item scores are associated with the observed therapist behaviors rather than the coder making the observations is critical to supporting the measure's reliability (see Arnold, 2011; & Fleiss, 1979). The ICC(2,2) for the CBAY-A items ranged from .43 (Coaching) to .93 (Problem Solving) and had a mean of .77 (*SD* = 0.15). The ICC (2, 2) coefficients for all but three items were in the 'Good' or 'Excellent' range; Psychoeducation Anxiety, Maintenance, and Coaching were in the 'Fair' range. Whereas it is not possible to know why the interrater reliability of those three items were less favorable than the other items' interrater reliability, the restricted range of adherence scores for those items is a possible reason. The less favorable reliability data for these items may be due to these items not being explicitly prescribed in the Coping Cat manual (Kendall & Hedtke, 2006a). To better evaluate these three items, their reliability and validity should be evaluated based on coding of sessions with a high likelihood that these therapist behaviors took place with varying levels of quantity and quality.

The interrater reliability of the CBAY-A is generally consistent with other measures of treatment adherence. For example, Hogue and colleagues (2008) obtained interrater reliability coefficients ranging from .56 to .83 (*M* = .71) while Barber and colleagues (2004) obtained interrater reliability coefficients ranging from .54 to .94 (*Median* = .85). In each case, the majority of item reliably coefficients were greater than .60. Overall, these data seem to indicate

that expert coders can code carefully defined adherence items with sufficient reliability when developed using the measure development and coding processes used with the CBAY-A and other well-supported measures of adherence.

The validity analyses for the CBAY-A focused on showing that the item scores were strongly correlated with other items or scales that measured similar therapist behaviors (convergent validity) and that the item scores are unrelated to conceptually independent measures (divergent validity). Exploratory analyses evaluated whether or not therapists from different study conditions delivered CBT for youth anxiety at different rates, thus receiving different scores (discriminative validity). These exploratory analyses were less conclusive. Every CBAY-A item (except Rapport Building, which did not have a corresponding item on the TPOCS-RS) was strongly correlated (i.e., $r > .30$) with at least one item or scale from the TPOCS-RS that measured a comparable therapist behavior. The divergent validity analyses of the CBAY-A item scores were moderately to minimally related to a conceptually unrelated scales from the TPOCS-RS, the Psychodynamic, Family, and Client-Centered scales. The magnitude of the correlations was less than the criterion value used for these analyses ($r = .193$) for 25 of 36 analyses. These data indicate that the CBAY-A items measure what they are supposed to be measuring but may be too sensitive to delivery of non-CBT treatment elements either through direct influence on the item scores or through an opportunity cost effect.

The comparison of item scores derived from CC and YAS-CC therapists (two conditions in which CC was the prescribed treatment) was moderately supportive of the discriminative validity of the CBAY-A whereas the comparison of item scores derived from the YAS-CC and YAS-UC was strongly supportive of the discriminative validity of the CBAY-A. The differences between treatment integrity scores for CC and YAS-CC therapists are more difficult to attribute

to a single cause, because the therapists were practicing in different settings (organizational climate, unmeasured in these studies, is known to influence treatment delivery; Glisson & Hemmelgarn, 1998), treating patients from different populations (patients receiving services in university based clinics differ from those receiving services in community clinics; Ehrenreich-May, Southam-Gerow, Hourigan, Wright, Pincus, & Weisz, 2011), had a different training background (CC therapists were trained in CBT throughout graduate school in addition to pre-trial manual-specific training compared to YAS-CC therapists who had a more heterogeneous training background and potentially had different attitudes toward evidence-based and manualized treatments; Beidas, Mychailyszyn, Edmunds, Khanna, Downey, & Kendall, 2012) provided more CBT and higher quality CBT than therapists from a more heterogeneous training background. As there are many variables that potentially influence the variable adherence to CBT for youth anxiety of the therapy delivered in different treatment conditions, these data should not be used to harshly critique therapists working in community settings.

Validity analyses of observational measures of adherence (see Barber et al., 2004; Hogue et al. 2008) has received less focus and been less rigorous than reliability analyses. For example, Barber and colleagues (2004) showed that the adherence scores on the Cognitive Therapy Adherence-Competence Scale could distinguish therapists delivering different therapies (discriminative validity) but did not evaluate the construct validity of the measure. Hogue and colleagues (2008) did evaluate the construct validity of the adherence scores on the Therapist Behavior Rating Scale—Competence but did so by interpreting inter-item correlations. This is problematic because the items measure different, though potentially co-occuring, therapist behaviors. Relative to previous studies, the validity analyses in this study is a strength.

Competence

The psychometric evaluation of the CBAY-C also addressed interrater reliability, construct validity, and discriminative validity (also in an exploratory fashion). There has been a consistent pattern within the literature (e.g., Cognitive Therapy Adherence-Competence Scale [Barber et al., 2004] and Therapist Behavior Rating Scale—Competence [Hogue et al., 2008]) where competence measures were less well supported than corresponding adherence measures. For example, Hogue and colleagues found interrater reliability (ICC) of the competence scores on the Therapist Behavior Rating Scale—Competence ranged from .15 to .55, substantially lower than the adherence scores. Whereas this pattern also held true for the CBAY-C and CBAY-A, the difference was less extreme and substantial support for the psychometric strength of the CBAY-C was found.

For the CBAY-C, raters' ability to reliably detect the presence or absence (Cohen's Kappa) of the target therapist behaviors was generally good, ranging from .44 to .90 (M = .70, SD = .10), and the interrater reliability (ICC) of the competence items scores (positively scored items only) ranged from .37 to .81 with only four of 23items falling short of the 'Good' range. Although slightly less favorable than the reliability of the CBAY-A, these scores indicate independent coders were able to code the CBAY-C item scores quite reliably.

The validity analyses of the CBAY-C are more complex than the validity analyses for the CBAY-A. One of the reasons why the development of the CBAY-C is so important to the field of treatment integrity for youth anxiety is that there are no other existing measures of therapist competence for CBT for youth anxiety with demonstrated psychometric strength. Unfortunately, this means that there are no established measures that can be used to directly evaluate the CBAY-C's validity. Therefore, the process of evaluating the validity of the CBAY-C included

four steps. First, CBAY-C item scores were compared to the CBAY-A item scores to determine the extent to which the coders are able to detect the appropriate therapist behaviors as measured by the CBAY-A (convergent construct validity). Second, CBAY-C item scores were compared to two COMP-CF scale scores to determine the extent to which the item scores on the CBAY-C are sensitive to global competence (convergent construct validity). Third, CBAY-C item scores were compared to the non-CBT scales from the TPOCS-RS do determine the extent to which the items are independent of theoretically independent measures of therapist behaviors (divergent validity). Finally, CBAY-C item scores across the two conditions that included therapists instructed to deliver CBT for youth anxiety were compared (i.e., CC and YAS-CC; discriminative validity).

The analyses comparing the CBAY-C with CBAY-A items indicated that, overall, the items are sufficiently sensitive to the appropriate therapist behaviors when making the determination of whether or not the item should be positively coded (Cohen's Kappa; $M = 0.73$; $SD = .19$). Comparison of positively-scored CBAY-C items and COMP-CF scores were also generally supportive of the construct validity of the CBAY-C ($M = 45$; $SD = 0.12$). The comparison of CBAY-C items scores and the Psychodynamic, Family Therapy, and Client Centered subscale scores on the TPOCS-RS were moderately supportive of the divergent validity of the CBAY-C. Finally, as hypothesized, the item scores on the CBAY-C were significantly greater for the CC therapists than the YAS-CC therapists for 21 of 23 items.

Uniqueness of the Adherence and Competence Measures

The analysis of the uniqueness of the items was based on the premise that corresponding items on the CBAY-A and CBAY-C are related yet measure unique aspects of therapist behaviors. Therefore, they should be correlated with one another, but *too* high of a correlation

would indicate that the two measures are not distinct from one another. Of the 20 items for which there were corresponding versions of the items on both the CBAY-A and CBAY-C, the mean correlation between items is r = .44 (SD = .20) and ranged from r = .01 (Coaching) to r = .74 (Emotion Education). Cognitive Anxiety was also above the criterion value of r = .70.

These findings are generally consistent with past investigations that looked at the relationship between adherence and competence. When looking at adherence and competence scales as a whole, the relationship strength ranged from correlation coefficients of .31 (Carroll et al., 2000) to .54 (Barber et al. 2004; Barber & Crits-Christoph, 1996; also see Butler et al., 1995). When looking at corresponding items on the Therapist Behavior Rating Scale—Competence, correlation coefficients ranged from .61 to .74 (Hogue et al., 2008).

In summary, these results generally suggest that the CBAY-A and CBAY-C are distinct from one another and that they each have sufficient preliminary psychometric support to warrant their use in applied settings. To further interpret the findings, it is important to take into consideration possible uses of the measures and the strengths and limitations of this study.

Strengths and Limitations of the Study

The following section will discuss strengths and limitations of various features of the present study. The features of the study that are addressed below are: scoring approach used by the two measures, focus on treatment elements rather than a specific treatment manual, decisions about sessions coded, and masking of coders.

Both measures employ a simple scoring approach with each item being coded and scored independently. The CBT for youth anxiety literature has identified treatment elements that are commonly included in evidence-based treatments or have been shown to be effective in single element treatments but the literature is insufficiently advanced to explicitly prescribe what CBT

for youth anxiety must include. Consistent with this state of the literature, this scoring approach takes a conservative approach by focusing on the individual treatment elements rather than creating composite scores. This represents a strength of the present study, because it allows the measures to be highly flexible, and supports efforts to identify other scoring procedures that facilitate specific aims but a weakness of the overall CBT for youth anxiety treatment literature. Furthermore, with this approach each item essentially becomes a separate "single-item" measure. In addition to the problems associated with single-item measures (DeVellis, 2003), this feature of the CBAY-A and CBAY-C means that the psychometric strength of each item is independent of the other items and interpretation of the reliability and validity data becomes complex.

The next feature to be highlighted is the focus on common elements of CBT for youth anxiety rather than a specific treatment manual. This feature also makes the CBAY-A and CBAY-C highly flexible treatment integrity measures. Unfortunately, this strength is also associated with a limitation of the present study. Since the vast majority of therapist behaviors that were coded as CBT for youth anxiety were drawn from the two study conditions in which Coping Cat was the prescribed treatment, this study is not able to determine the psychometric strength of the CBAY-A and CBAY-C with non-Coping Cat CBT for youth anxiety. A related issue is the low rate at which certain treatment elements (e.g., Weekly Ratings, Maintenance) were observed. The almost complete absence of therapist engaging in weekly ratings resulted in the Weekly Rating item being dropped from the CBAY-A and the few observations of Maintenance limits the confidence in the findings associated with that item. Future studies that use the CBAY-A and CBAY-C (with different coders and evaluating sessions using a treatment manual other than Coping Cat) will need to take care to ensure that coders are thoroughly trained, items are performing as expected, coders are maintaining an adequate level of interrater

reliability, and evidence of item validity is taken into consideration when interpreting findings. This is particularly true for items that were rarely seen in this sample.

Another unique feature of this study is that every codable session was reviewed in its entirety and was double-coded. The initial goal was to code every session but this was not possible because some original session recordings were not provided by the original trials and others were not codable. The high rate of coded sessions represents a significant strength of this study over other treatment integrity measurement studies. Though not designed as treatment integrity studies, Kendall's evaluation of individual versus group versus supportive psychotherapy from which the CC sessions were drawn reviewed 15-minute segments of 30% of study sessions to ensure adequate treatment integrity (Kendall et al., 2008). Currently, no studies have empirically evaluated the consequences of conducting treatment integrity research on a subsample of session, versus the whole sample. Such a study would be of enormous benefit given the time commitment and associated financial resources needed to code every session.

One issue that clearly limits confidence in the results is the fact that the effort to mask the coders from the study condition of each session was not effective. There were many contextual details (e.g., appearance of the room) and therapist behaviors (e.g., discussion of the study condition and discussion of local events [e.g., local sports teams] and places [e.g., entertainment and vacation destinations]) that were distinctive and made it impossible for coders to remain masked to the study condition. It is unclear what effect this had on the actual coding of the therapist behaviors, but it is possible that bias may have crept into the coders ratings based on expectations for the different study conditions and influenced the discriminative validity results. This is a difficult issue to do away with when working with naturalistic therapy settings, and future studies should consider more sophisticated approaches to maintaining coder masking (e.g.,

keeping the coders blind to the possible sources of the session recordings). It is not expected that this limitation had a significant impact on the coders perspectives, as efforts were made to reduce the impact of bias (e.g., randomizing coding order and discussing the value in fidelity with flexibility in coder training meetings), but it impossible to fully remove the possible threat of bias in this case.

Overall, the present study should be viewed as an important effort to evaluate the reliability and validity evidence for the CBAY-A and CBAY-C item scores. Data from this study generally, though not uniformly, supports the reliability of the CBAY-A and CBAY-C when coded by expert coders and indicate that the item scores reflect what the items are designed to measure. As noted earlier, future users of each measure should be aware of the areas of strength and weakness of the present study and should take steps to confirm the aspects of the CBAY-A and CBAY-C that are well supported (e.g., flexibility of scoring approach, focus on treatment elements, all available sessions were double coded) and address the aspects of the CBAY-A and CBAY-C that are less well supported (e.g., cumbersomeness of scoring procedure, unknown reliability and validity when used with non-Coping Cat CBT for youth anxiety).

Proposed Applications of the CBAY-A and CBAY-C

This section will focus on applications of the CBAY-A and CBAY-C in light of their psychometric strength. Specifically, the following uses will be reviewed: (a) as a manipulation check in treatment evaluation studies, (b) within therapist training efforts, and (c) by supervisors or third-party payers to evaluate therapists. These applications are not exhaustive but are meant to highlight some commonly mentioned uses of treatment integrity measures.

Manipulation check. As discussed earlier, ensuring that the treatment being implemented in a treatment evaluation study is actually delivered consistent with how it is

intended to be delivered and that it is distinct from a control or comparison treatment is a central task in treatment evaluation research and an important application for treatment adherence measures (Barber et al., 1996; Carroll et al., 1998; Chambless & Ollendick, 2001; Kazdin, 2003). One of the problems with previous treatment integrity measures is that ratings only addressed adherence and were dichotomous (e.g., Kendall et al., 2008; Southam-Gerow et al., 2010). The CBAY-A allows for the quantity of CBT for youth anxiety treatment elements to be measured on continuous scales. However, these data only evaluated the use of item-level scores and not composite score. Furthermore, these data do not identify a specific benchmark that therapists should achieve in order to be deemed "adherence" or "competent."

Therapist training. Another application of the CBAY-A and CBAY-C relates to the possibility of using the measures to evaluate therapist-training efforts. Recent years have seen an increased focus on dissemination and implementation (D & I) of evidence-based practices, helping therapists to deliver evidence-based treatments to youth in community settings such as community mental health centers and schools. These efforts include providing training in evidence-based treatment approaches to therapists and supervisors; addressing agency and/or setting, variables such as increasing the expectation that therapy is evidence-based; and affecting the referral stream so that therapists' caseloads become more homogeneous and receptive for a given treatment approach. While the ultimate goal of these efforts is to improve symptoms and functioning for the patients, a proximal goal is to improve the quantity of evidence-based treatment elements and quality with which the therapy being delivered to the patients. Through the use of treatment integrity measures such as the CBAY-A and CBAY-C, these D & I efforts can be evaluated and the science of improving mental health services in community settings can be advanced. Similarly, programs that specialize in training therapists such as graduate schools,

and providers of continuing education could use the CBAY-A and CBAY-C to measure the change in therapist behaviors before and after participation in their programs. These measures may be helpful to training organizations to demonstrate that therapist training efforts have, in fact, been effective in increasing the therapists' uptake of desired behaviors (i.e., using more CBT for youth anxiety and doing so more competently). Therapists with low adherence and competence item-level scores could be steered toward training interventions that target their particular area of needed growth. This is a particular advantage of having measures where each element is measured with a separate item rather than having a single item that reflects overall treatment delivery or collapsing separate items scores into a single summary score. With separate items it is possible to tease apart what areas therapists may need the most additional training in.

Evaluating therapists. The final application to be discussed is similar to the previous application and addresses the possibility of using these measures in the evaluation of therapists delivering CBT for youth anxiety. The list of individuals and entities that may be interested in evaluating therapists includes: (a) supervisors and agency administrators who may be interested in which therapists need additional training; (b) third-party payers (e.g., insurance companies, Medicaid, Medicare, child welfare, schools) who are interested in steering patients toward therapists that are most likely to produce positive treatment effects; and (c) patients and patient advocacy groups who want the best possible outcome for themselves and for those they advocate for.

Of the groups mentioned above, the logistics of evaluating therapists using these measures is most straightforward for supervisors and agency administrators and could be incorporated into a standard quality assurance/improvement practice. For instance, a supervisor

may evaluate a particular therapist at the end of the year on their performance with a particular child client with an anxiety disorder. This evaluation may point out areas of strength and weakness of the therapist, and identify areas for further training. It is possible that these measures could also be adapted for use in 1:1 supervision, with therapists and supervisors watching sessions together and then rating the therapist on their performance.

These three potential applications of the CBAY-A and CBAY-C are focused on different facets of the data. Those using these measures as a manipulation check would likely be interested in coding all or the vast majority of sessions over a limited period of time (the time period of the study being conducted). For those evaluating the effectiveness of therapist trainings, the primary interest would be in the capacity of the therapist to deliver a high dose of high quality therapy. Therefore, the coding demands may be lower. For those interested in evaluating therapists, the coding demands could be vast as it is likely that many sessions from many different therapists would need to be coded on an ongoing basis. In each case, the CBAY-A and CBAY-C could be useful assessments though the time commitment needed to use them stands as a significant barrier. However, as noted earlier, the results from this study should only be seen as a first step in an ongoing assessment of the psychometric evaluation of the CBAY-A and CBAY-C. It is entirely plausible that coders not involved in the development of the measures would have a more difficult time reliably using these measures.

Future Directions

Evaluating the psychometric strength of a measure is an ongoing process. In order to evaluate the generalizability of these findings, it will be important to evaluate the CBAY-A and CBAY-C with different coders and therapists delivering CBT for youth anxiety guided by a treatment manual or approach other than Coping Cat. To further expand the utility of these

measures, future research could adopt the approach used in this study to develop the coding manual and develop items to measure delivery of CBT elements for other treatment foci. The use of these measures in ongoing clinical practice will be seriously limited by the financial and time cost associated with paying masters-level therapists to code sessions. To address this barrier, future research could investigate whether coders with less education, training, and experience can use the CBAY-A and CBAY-C while maintaining sufficient reliability and validity.

Two areas of particular interest to the field of treatment integrity is the relationship between adherence and treatment outcome and competence and treatment outcome. Thus far, findings on this front have been mixed. The CBAY-A and CBAY-C have the potential to make a strong contribution to the field because they employ continuous scales and because each treatment element is measured separately. Another focus of treatment integrity research that the CBAY-A and CBAY-C can assist with is the development of therapist-report, youth-report, and/or caregiver-report measures of treatment integrity that can be evaluated against established observational coding measures of the same therapist behaviors. The development of psychometrically strong therapist-report measures of treatment integrity could have a significant impact of the science of treatment research. Finally, future researchers should partner with community mental health center administrators and therapists to explore the feasibility of various real-world applications of the CBAY-A and CBAY-C.

Conclusion

This study has provided a first glance at the psychometric properties of two new measures of treatment integrity, the CBAY-A assessing therapist adherence, and the CBAY-C assessing competence with CBT for youth anxiety. Findings from this study are encouraging but

not overwhelmingly so. They demonstrate that the majority of the items perform as expected, have strong reliability and validity even when used with therapists from diverse training backgrounds. While additional evaluation of these measures in needed, especially for items that assess therapist behaviors that were rarely seen in this sample of sessions, the CBAY-A and CBAY-C open the door to a variety of intriguing and important research questions that, prior to this point, have been inaccessible due to the lack of appropriate measurement tools.

List of References

List of References

Addis, M. E., & Krasnow, A. D. (2000). A national survey of practicing psychologists' attitudes toward psychotherapy treatment manuals. Journal *of Consulting and Clinical Psychology, 68*(2), 331-339.

American Psychological Association. (1993). Task force on promotion and dissemination of psychological procedures. [online]. Available: http://www.apa.org/divisions/div12/est/chamble2.pdf. (April 19, 2013)

Arnold, C. C. (2011). *Evaluation of a community mental health center's assessment clinic: development of a novel assessment evaluation tool.* (Unpublished master's thesis, Virginia Commonwealth University Richmond, Virginia).

Asay, T.P., and Lambert, M.J. (1999). The empirical case for the common factors in therapy: quantitative findings. In Hubble, Duncan, Miller (Eds), *The Heart and Soul of Change* (pp. 23–55)

Barber, J. P., Gallop, R., Crits-Christoph, P., Frank, A., Thase, M. E., Weiss, R. D., & Beth Connolly Gibbons, M. (2006). The role of therapist adherence, therapist competence, and alliance in predicting outcome of individual drug counseling: Results from the National Institute Drug Abuse Collaborative Cocaine Treatment Study. *Psychotherapy Research, 16*(02), 229-240.

Barber, J., & Crits-Christoph, P. (1996). Development of a therapist adherence/competence rating scale for supportive-expressive dynamic psychotherapy: a preliminary report. *Psychotherapy Research, 6*(2), 81-94.

Barber, J.P., Mercer, D., Krakauer, I., & Calvo, N. (1996). Development of an adherence/competence rating scale for individual drug counseling. *Drug and Alcohol Dependence, 43*(3), 125-132.

Barber, J.P., Sharpless, B.A., Klostermann, S., & McCarthy, K.S. (2007). Assessing intervention competence and its relation to therapy outcome: A selected review derived from the outcome literature. *Professional Psychology Research and Practice, 38*(5), 493-500.

Barrington, J., Prior, M., Richardson, M., & Allen, K. (2005). Effectiveness of CBT versus standard treatment for childhood anxiety disorders in a community clinic setting. *Behaviour Change*, *22*(01), 29-43.

Beidas, R. S., Mychailyszyn, M. P., Edmunds, J. M., Khanna, M. S., Downey, M. M., & Kendall, P. C. (2012). Training school mental health providers to deliver cognitive-behavioral therapy. *School mental health, 4*(4), 197-206.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological bulletin*, *110*(2), 305-314.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological review*, *110*(2), 203-218.

Brown, R. C. C. (2011). *The Development of the Common Factor Therapist Competence Scale for Youth Psychotherapy.* (Unpublished doctoral dissertation, Virginia Commonwealth University Richmond, Virginia).

Budd, R., & Hughes, I. (2009). The Dodo Bird Verdict—controversial, inevitable and important: A commentary on 30 years of meta-analyses. *Clinical Psychology & Psychotherapy*, *16*(6), 510-522.

Carroll, K. M., Connors, G. J., Cooney, N. L., DiClemente, C. C., Donovan, D. M., Kadden, R. R., ... & Zweben, A. (1998). Internal validity of Project MATCH treatments: Discriminability and integrity. *Journal of Consulting and Clinical Psychology*, *66*(2), 290-303.

Carroll, K.M., Nich, C., Sifry, R.L., Nuro, K.F., Frankforter, T.L., Ball, S.A., ... & Rounsaville, B. J. (2000). A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug and alcohol dependence*, *57*(3), 225-238.

Carroll, K., Nich, C., & Rounsaville, B. (1998). Utility of therapist session checklists to monitor delivery of coping skills treatment for cocaine abusers. *Psychotherapy Research,8*(3), 307-320.

Chambless, D.L., & Ollendick, T.H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology, 52*, 685-716.

Chen, Q., Kwok, O. M., Luo, W., & Willson, V. L. (2010). The impact of ignoring a level of nesting structure in multilevel growth mixture models: A Monte Carlo study. *Structural Equation Modeling*, *17*(4), 570-589.

Chorpita, B. F. (2007). Modular cognitive behavioral treatment for childhood anxiety disorders. Guilford Press: New York.

Chorpita, B. F., Daleiden, E. L., & Weisz, J. R. (2005). Identifying and selecting the common elements of evidence based interventions: A distillation and matching model. *Mental Health Services Research, 7*(1), 5-20.

Chu, B. C., & Kendall, P. C. (2004). Positive association of child involvement and treatment outcome within a manual-based cognitive-behavioral treatment for children with anxiety. *Journal of Consulting and Clinical Psychology, 72*(5), 821.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment, 6*(4), 284.

Cronbach, L.J., & Meehl, P.E.(1955).Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

DeVellis, R.F., (2003) Scale development: Theory and application (2nd ed.). Thousand Oaks, CA: Sage Publishing.

Ehrenreich-May, J., Southam-Gerow, M. A., Hourigan, S. E., Wright, L. R., Pincus, D. B., & Weisz, J. R. (2011). Characteristics of anxious and depressed youth seen in two different clinical contexts. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(5), 398-411.

Epstein, R.M., & Hundert, E.M. (2002). Defining and assessing professional competence. *JAMA: the Journal of the American Medical Association,287*(2), 226-235.

Foster, S.L., & Cone, J.D. (1995). Validity issues in clinical assessment. *Psychological Assessment, 7*(3), 248-260.

Glisson, C., & Hemmelgarn, A. (1998). The effects of organizational climate and interorganizational coordination on the quality and outcomes of children's service systems. *Child abuse & neglect,22*(5), 401-421.

Groth-Marnat, G., (2003). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: John Wiley & Sons Inc.

Hattie, J., & Cooksey, R. W. (1984). Procedures for assessing the validities of tests using the "known-groups" method. *Applied Psychological Measurement, 8*(3), 295-305.

Hayes, S. C., Barlow, D. H., & Nelson-Gray, R. O. (1999). *The scientist practitioner: Research and accountability in the age of managed care (2nd Ed.)*. Boston: Allyn and Bacon.

Hayes, S.C, Nelson, R.O., & Jarrett, R.B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *The American psychologist, 42*(11), 963-74.

Hill, C.E., O'Grady, K.E., & Elkin, I. (1992). Applying the Collaborative Study Psychotherapy Rating Scale to rate therapist adherence in cognitive-behavior therapy, interpersonal

therapy, and clinical management. Journal *of Consulting and Clinical Psychology*, *60*(1), 73.

Hogue, A., Liddle, H.A., & Rowe, C. (1996). Treatment adherence process research in family therapy: A rationale and some practical guidelines. *Psychotherapy*, *33*(2), 332-345.

Hogue, A., Dauber, S., Chinchilla, P., Fried, A., Henderson, C., Inclan, J., ... & Liddle, H.A. (2008). Assessing fidelity in individual and family therapy for adolescent substance abuse. *Journal of substance abuse treatment*,*35*(2), 137-147.

Hogue, A., Liddle, H.A., Rowe, C., Turner, R.M., Dakof, G.A., & LaPann, K. (1998). Treatment adherence and differentiation in individual versus family therapy for adolescent substance abuse. *Journal of Counseling Psychology*, *45*, 104-114.

Hogue, A., Henderson, C.E., Dauber, S., Barajas, P.C., Fried, A., & Liddle, H.A. (2008). Treatment adherence, competence, and outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology*,*76*(4), 544.

Jones, H. A., Clarke, A. T., & Power, T. J. (2008).Expanding the concept of intervention integrity: A multidimensional model of participant engagement. *Balance*, *23*, 4-5.

Kaslow, N.J. (2004). Competencies in professional psychology. *American Psychologist, 59*(8), 774-781.

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

Kendall, P.C., Chu, B., Gifford, A., Hayes, C., & Nauta, M. (1999).Breathing life into a manual: Flexibility and creativity with manual-based treatments. *Cognitive and Behavioral Practice*,*5*(2), 177-198.

Kendall, P. C., Hudson, J. L., Gosch, E., Flannery-Schroeder, E., & Suveg, C. (2008). Cognitive-behavioral therapy for anxiety disordered youth: A randomized clinical trial evaluating child and family modalities. *Journal of consulting and clinical psychology, 76*(2), 282-297.

Kendall, P. C., & Hedtke, K. (2006a). *Cognitive-behavioral therapy for anxious children: Therapist manual.* (3rd ed.). Ardmore, PA: Workbook Publishing.

Kendall, P. C., & Hedtke, K. (2006b). *Coping cat workbook* (2nd ed.). Ardmore, PA: Workbook Publishing.

Lemanek, K.L., Kamps, J., & Chung, N.B. (2001). Empirically supported treatments in pediatric psychology: regimen adherence. *Journal of Pediatric Psychology*,*26*(5), 253-275.

Liber, J. M., McLeod, B. D., Van Widenfelt, B. M., Goedhart, A. W., van der Leeden, A. J., Utens, E. M., & Treffers, P. D. (2010). Examining the relation between the therapeutic alliance, treatment adherence, and outcome of cognitive behavioral therapy for children with anxiety disorders. *Behavior Therapy*, *41*(2), 172-186.

Lichstein, K.L., Riedel, B.W., & Grieve, R. (1994). Fair tests of clinical trials: A treatment implementation model. *Advances in Behaviour Research and Therapy,16*(1), 1-29.

Luborsky, L., Rosenthal, R., Diguer, L., Andrusyna, T.P., Berman, J.S., Levitt, J.T., ... & Krause, E. D. (2002). The dodo bird verdict is alive and well—mostly. *Clinical Psychology: Science and Practice*, *9*(1), 2-12.

Marder, A. M. (2008). Measuring therapist adherence to a manual-based treatment tested in a community setting: The PASCET Manual Adherence Scale (P-MAS). (Unpublished doctoral dissertation). Virginia Commonwealth University, Richmond.

McLeod, B. D., Islam, N. Y., & Wheat, E. (in press). Designing, conducting, and evaluating therapy process research. In J. Comer & P. Kendall (Eds.), *The Oxford Handbook of Research Strategies for Clinical Psychology*. New York, NY: Oxford University Press.

McLeod, B. D., Smith, M. M., Southam-Gerow, M. A., Weisz, J. R., & Kendall, P. C. (2014,October 27). Measuring Treatment Differentiation for Implementation Research: The Therapy Process Observational Coding System for Child Psychotherapy Revised Strategies Scale. *Psychological Assessment*. Advance online publication. http://dx.doi.org/10.1037/pas0000037

McLeod, B.D., Southam-Gerow, M.A., Tully, C.B., Rodríguez, A., & Smith, M.M. (2013).Making a case for treatment integrity as a psychosocial treatment quality indicator for youth mental health care. *Clinical Psychology: Science and Practice*, *20*(1), 14-32.

McLeod, B.D., Southam-Gerow, M., & Weisz, J.R. (2009).Conceptual and methodological issues in treatment integrity measurement. *School Psychology Review*, *38*(4), 541.

McLeod, B. D., & Weisz, J. R. (2005). The therapy process observational coding system-alliance scale: Measure characteristics and prediction of outcome in usual clinical practice. *Journal of Consulting and Clinical Psychology*, *73*(2), 323-333.

McLeod, B.D., & Weisz, J.R. (2010).The therapy process observational coding system for Child Psychotherapy Strategies Scale. *Journal of Clinical Child & Adolescent Psychology*, *39*(3), 436-443.

Miller, S. J., & Binder, J. L. (2002). The effects of manual-based training on treatment fidelity and outcome: A review of the literature on adult individual psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, *39*(2), 184.

Mowbray, C. T., Holter, M. C., Teague, G. B., Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*(3), 315-340.

Perepletchikova, F. (2009).Treatment integrity and differential treatment effects. *Clinical Psychology: Science and Practice, 16*(3), 379-382.

Perepletchikova, F., & Kazdin, A.E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice,12*(4), 365-383.

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology, 75*(6), 829-840.

Pribluda, V.S., Barojas, A., Añez, A., López, C.G., Figueroa, R., Herrera, R., ... & Evans, L. (2012). Implementation of basic quality control tests for malaria medicines in Amazon Basin countries: results for the 2005–2010 period. *Malaria journal, 11*(1), 202.

Rosen, G. M., & Davison, G. C. (2003). Psychology should list empirically supported principles of change (ESPs) and not credential trademarked therapies or other treatment packages. *Behavior Modification, 27*(3), 300-312.

Schoenwald, S. K., Henggeler, S. W., Brondino, M. J., & Rowland, M. D. (2000). Multisystemic therapy: monitoring treatment fidelity*. *Family Process, 39*(1), 83-103.

Schoenwald, S. K., Carter, R. E., Chapman, J. E., & Sheidow, A. J. (2008). Therapist adherence and organizational effects on change in youth behavior problems one year after multisystemic therapy. *Administration and Policy in Mental Health and Mental Health Services Research, 35*(5), 379-394.

Schaffer, D., Fisher, P., & Lucas, C. (1996). The NIMH diagnostic interview schedule for children version 4.0 (DISC-4.0). *New York: Ruane Center for Early Diagnosis, Division of Child Psychiatry, Columbia University*.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull, 86*(2), 420-428.

Silverman, W. K., Pina, A. A., & Viswesvaran, C. (2008). Evidence-based psychosocial treatments for phobic and anxiety disorders in children and adolescents. Journal *of Clinical Child & Adolescent Psychology, 37*(1), 105-130.

Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment, 17*(4), 396-408.

Southam-Gerow, M.A., Weisz, J.R., Chu, B.C., McLeod, B.D., Gordis, E.B., & Connor-Smith, J.K. (2010). Does cognitive behavioral therapy for youth anxiety outperform usual care in community clinics? An initial effectiveness test.*Journal of the American Academy of Child & Adolescent Psychiatry*,*49*(10), 1043-1052.

Southam-Gerow, M. A., & McLeod, B. D. (2013). Advances in applying treatment integrity research for dissemination and implementation science: Introduction to special issue. *Clinical Psychology: Science and Practice*, *20*(1), 1-13.

Southam-Gerow, McLeod, Arnold, & Rodríguez, (Unpublished). Scoring Manual for the CBT for Youth Anxiety Therapist Adherence Scale (CBAY-A).

Southam-Gerow, McLeod, Quinoy, & Eonta (Unpublished). The Scoring Manual for the CBT for Youth Anxiety Therapist Competence Scale (CBAY-C)

Tabachnick, B. G., & Fidell, L. S. (2007).Using multivariate statistics. Boston: Pearson.

Taylor, P. J. (2010). An introduction to intraclass correlation that resolves some common confusions. Retrieved from http://www.faculty.umb.edu/peter_taylor/09b.pdf.

Waltz, J., Addis, M.E., Koerner, K., & Jacobson, N.S. (1993).Testing the integrity of a psychotherapy protocol: assessment of adherence and competence. Journal *of consulting and clinical psychology*, *61*(4), 620.

Weersing, V.R., & Weisz, J.R. (2002). Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. Journal *of Consulting and Clinical Psychology*,*70*(2), 299-310.

Weisz, J. R., Jensen-Doss, A., & Hawley, K. M. (2006).Evidence-based youth psychotherapies versus usual clinical care.*American Psychologist*, *61*(7), 671-689.

Weisz, J.R., Moore, P.S., Southam-Gerow, M.A., Weersing, V.R., Valeri, S.M., & McCarty, C.A. (1999). Therapist's manual: PASCET: Primary and Secondary Control Enhancement Training Program. *University of California, Los Angeles.*

Appendix A

Cognitive-Behavioral Therapy for Youth Anxiety Therapist Adherence Scale (CBAY-A)

Cognitive-Behavioral Therapy for Youth Anxiety Therapist Competence Scale (CBAY-C)

Coder ID: _____ Recording ID#: _____ Coding Date: _____/_____/_____

ADH-CBT Extensiveness Scale

Instructions: Using the grid provided below, please indicate PRESENCE of any item observed for each **five-minute** time segment. Use a "+" to indicate *extensive presence*, "X" to indicate *moderate presence*, and "-" to indicate *slight presence*. After watching the ENTIRE recording, use the 1-7 scale to assign an Extensiveness rating (Ext) for all items that are present in at least ONE (1) time period. Also, record the number of time periods each item appeared in under Frequency (Freq).

1 Not at all	2	3 Somewhat	4	5 Considerably	6	7 Extensively

Item																Freq	Ext
STANDARD																	
1. Agenda Setting																	
2. Weekly Rating																	
3. HW Review																	
4. HW Assigned																	
5. Rapport Bldg																	
MODEL																	
6. Psychoed-Anx																	
7. Emotion Ed																	
8. Fear Ladder																	
9. Relaxation																	
10. Cognitive-Anx																	
11. Problem Solving																	
12. Self-Reward																	
13. Coping Plan																	
14. Exposure: Prep																	
15. Exposure																	
16. Exposure: Debrief																	
17. Maintenance																	
DELIVERY																	
18. Didactic teaching																	
19. Collaborative teaching																	
20. Modeling																	
21. Rehearsal																	
22. Coaching																	
23. Self-disclosure																	

Coder ID: _____ Recording ID#: _____ Coding Date: _____

COMP-CBT Competence Scale

Instructions: Using the grid provided below, please indicate whether each specific item occurs during each **ten-minute** time segment. If an item occurs during a time segment place a "+" to indicate *above average* rating, "X" to indicate *average rating*, or a "–" to indicate *below average* rating in the space provided in the grid corresponding to the correct item. After watching the ENTIRE recording, use the 1-7 scale to assign a Competence (Comp) rating for all items that are present in at least ONE (1) time period.

0	1	2	3	4	5	6	7
Not Present	Very Poor	Poor	Acceptable	Adequate	Good	Very Good	Excellent

Item									Comp
STANDARD									
1. Within session focus									
2. Across session focus									
3. Structure/ phase									
4. Homework Review									
5. HW assigned									
MODEL									
6. Psychoed-Anxiety									
7. Emotion Ed									
8. Fear Ladder									
9. Relaxation									
10. Cognitive-Anxiety									
11. Problem Solving									
12. Self-Reward									
13. Coping Plan									
14. Exposure: Prep									
15. Exposure									
16. Exposure: Debrief									
17. Maintenance									
DELIVERY									
18. Didactic teaching									
19. Collaborative teaching									
20. Modeling									
21. Rehearsal									
22. Coaching									
23. Self-disclosure									

Vita

Cassidy C Arnold was born on April 26th, 1980 in Chicago, Illinois. He graduated from University Prep High School in Seattle Washington in 1998 and University of Colorado, Denver in 2006. Prior to the entering the clinical psychology program at Virginia Commonwealth University, Cassidy worked with Outward Bound and Seattle Children's Hospital. He completed his Master of Science in Clinical Psychology in 2011. He is currently finishing his predoctoral internship at the Medical University of South Carolina in Charleston and will be graduating with his Doctorate of Philosophy in Clinical Psychology in 2015.