2018

A PRACTICE ELEMENTS-BASED APPROACH TO ADHERENCE MEASUREMENT

Stephanie Violante
Virginia Commonwealth University

A PRACTICE ELEMENTS-BASED APPROACH TO ADHERENCE MEASUREMENT

A thesis submitted in partial fulfillment of the requirements for the degree Master of Science, at Virginia Commonwealth University

by

Stephanie Violante
Bachelor of Science, University of Washington, 2013

Director: Bryce D. McLeod, Ph.D.
Professor, Department of Psychology

Virginia Commonwealth University
Richmond, Virginia
November, 2018

Acknowledgment

I would like to thank Dr. Bryce McLeod for his mentorship, guidance, and feedback throughout this process, Drs. Michael Southam-Gerow and Kevin Sutherland for kindly agreeing to serve on my committee, Drs. Meghan Smith, Adriana Rodriguez, Carrie Tully, and Nadia Islam for their dedication to hours of coding, and my family and friends for their continued encouragement and support.

Table of Contents

List of Tables

List of Figures

Abstract

A PRACTICE ELEMENTS-BASED APPROACH TO ADHERENCE MEASUREMENT

By Stephanie Violante, B.S.

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2018.

Major Director: Bryce D. McLeod, Ph.D.
Professor, Department of Psychology

The measurement of treatment adherence (i.e., the frequency and thoroughness with which a therapist delivers a treatment as designed) is critical to the field of implementation science, as adherence is often considered an indicator of successful implementation. Most existing instruments of treatment adherence are treatment protocol specific; however, this approach is costly, inefficient, and disallows cross-treatment and cross-study comparisons. Thus, there is a need for adherence instruments that can be used across treatment protocols. It has been suggested that an instrument that captures adherence at the practice element level would meet this need, as it would have utility across protocols that share practice elements. The current study examines the extent to which an observational adherence instrument designed to assess the core common practice elements found in individual cognitive-behavioral therapy (ICBT) for youth anxiety (the Cognitive-Behavioral Therapy Adherence Scale for Youth Anxiety; CBAY-A) can produce scores that can be reliably and validly interpreted across two separate ICBT protocols: *Coping Cat* (a standard manualized treatment; SMT) and *MATCH* (a modular manualized treatment; MMT). This study provides initial psychometric data for scores on an SMT subscale (comprised of CBAY-A items found in the SMT protocol) and an MMT subscale (comprised of CBAY-A

items found in the MMT protocol) of the CBAY-A. Treatment sessions (n = 359 SMT; n = 243 MMT) from 38 youth participants (n = 22 SMT; n = 16 MMT; M age = 9.84 years, SD = 1.65; 52.6% male, 60.5% Caucasian) in an effectiveness study were independently coded by two coders using the CBAY-A. Inter-rater reliability intraclass correlation coefficients (2,2) for the item scores averaged 0.83 (SD = 0.07) for the SMT group and 0.80 (SD = 0.09) for the MMT group. CBAY-A SMT and CBAY-A MMT subscale scores demonstrated evidence of convergent and discriminant validity via associations with observational instruments of therapist adherence designed specifically for the SMT and MMT protocols, and observational instruments of competence and alliance. Results provide preliminary evidence that the CBAY-A can be effectively used in place of two separate protocol-specific adherence instruments, indicating that it may be a flexible, efficient, and useful tool for capturing adherence to ICBT protocols for youth anxiety in a way that allows for comparisons across treatment protocols and research studies.

Introduction

Implementation science, the study of methods to promote the systematic uptake of evidence-based practices into routine practice, has gained momentum over the past decade (Eccles et al., 2009). Both the National Institutes of Health (Insel, 2008) and the Institute of Medicine (2015) have identified implementation research as a priority, calling for the investigation of implementation strategies and outcomes. A critical issue in implementation research is how to evaluate implementation success. One way of determining implementation success is through the measurement of implementation outcomes — "the effects of deliberate and purposive actions to implement new treatments, practices, and services" (Proctor et al., 2011, p. 65).

As an emerging field, implementation science lacks high-quality outcome instruments which are essential for scientific advancement (Lewis, Stanick, et al., 2015). The systematic development and psychometric characteristics of existing implementation science instruments has been described as "weak at best" (Lewis, Stanick, et al., 2015, p. 13). Much of the existing measurement for implementation outcomes has been 'home-grown' with very little investigation of psychometric properties, limiting the ability to draw conclusions from data and confidently generalize findings (Martinez, Lewis, & Weiner, 2014; Proctor et al., 2011). A review of 104 instruments of implementation constructs delineated in the Consolidated Framework for Implementation Research (CFIR; Damschroder et al., 2009) and the Implementation Outcomes Framework (IOF; Proctor et al., 2009) revealed that limited psychometric information was available across instruments, with 46% of instruments lacking information about critical psychometric domains (Lewis, Stanick, et al., 2015). Thus, to advance understanding of the implementation process and enhance efficiency in implementation research, development and

identification of instruments of implementation outcomes with demonstrated score reliability and validity are needed.

One of the most frequently studied implementation outcomes is treatment integrity (sometimes referred to as treatment fidelity; Powell, Proctor, & Glass, 2014), which refers to the degree to which a treatment was delivered as intended (Allen, Linnan, & Emmons, 2012). Treatment integrity is comprised of adherence (the extent to which a therapist delivers the treatment as designed, consisting of both frequency and thoroughness of delivery), competence (how well an intervention is implemented as determined by the therapist's skill and responsiveness when delivering the treatment), and differentiation (the therapist's delivery of non-prescribed therapeutic interventions; Bellg et al., 2004; Perepletchikova, Treat, & Kazdin, 2007). Of the three components of treatment integrity, adherence has received the most empirical attention with some evidence suggesting that it is related to outcomes in youth psychosocial treatment (e.g., Hogue, Henderson, et al., 2008; Schoenwald, Carter, Chapman, & Sheidow, 2008).

Treatment adherence has most commonly been utilized as an independent variable check (i.e. ensuring a particular treatment was delivered as designed) in clinical trials (Kazdin, 1994; Perepletchikova & Kazdin, 2005), making its measurement necessary to draw valid inferences. For example, if individuals receiving a treatment achieved better outcomes than individuals in a control group but the treatment was not delivered as intended, it cannot be determined whether the outcomes resulted from the intervention or alternative factors (Perepletchikova et al., 2007). With increased attention to the transfer of evidence-based treatments (EBTs) from research to community-based service settings, treatment adherence has emerged as a key implementation

outcome, as it is an indicator of implementation success (i.e. how well a new treatment, practice, or program is implemented; Mihalic & Irwin, 2003; Proctor et al., 2011).

The predominant approach to measuring treatment adherence has been to develop instruments specific to a manualized treatment protocol, and these instruments are rarely used in multiple studies (Garland & Schoenwald, 2013). This approach requires a significant allocation of resources to the development and use of redundant instruments (i.e., multiple instruments designed for the same treatment protocol). Furthermore, most existing instruments of treatment adherence lack evidence of score reliability and validity, and efforts to establish such instruments are largely absent in the literature (Baer et al., 2007; Garland & Schoenwald, 2013; Perepletchikova et al., 2007). In fact, in a review of treatment adherence instruments, of the 249 unique instruments identified by Schoenwald and Garland (2013), evidence for score validity (e.g., convergent, discriminant, predictive) was presented for fewer than 5%. Developing instruments of treatment adherence with strong evidence of score reliability and validity may require considerable time and effort (e.g. for observational instruments —developing a rating protocol, training coders in the protocol, assessing inter-rater reliability) and the cost and resources associated with establishing an instrument may very well deter researchers from adequately assessing treatment adherence (Perepletchikova et al., 2007).

Researchers have recently argued that a conceptual shift in treatment adherence measurement to a practice elements approach is required to best fit the needs of implementation researchers (McLeod, Southam-Gerow, Tully, Rodriguez, & Smith, 2013; Schoenwald et al., 2011). With this shift, adherence would no longer focus on the delivery of a specific evidence-based treatment protocol, but rather, evidence-based practice elements for a particular problem area (e.g. exposure for anxiety, behavioral activation for depression). This approach to defining

treatment adherence may benefit implementation research in several ways. First, it may be more efficient. Rather than needing a separate adherence instrument for each treatment protocol, this approach only requires a single instrument that contains practice elements for a given problem area (Schoenwald et al., 2011). For example, in evaluating 25 separate anxiety protocols for youth, Chorpita, Daleiden, and Weisz (2005) identified a common set of approximately six practice elements. With a practice elements-based approach to adherence measurement, 25 separate instruments of adherence may be reduced down to six items, eliminating redundancies. Second, practice elements-based treatment adherence instruments may be more flexible and able to be used across CBT protocols, and would therefore be more widely applicable across service settings wherein the specific CBT protocols used may differ (McLeod et al., 2013). Third, since such an instrument could be used across protocols it could also be used across studies, better allowing for cross-treatment and cross-study comparisons (Malik, Beutler, Alimohamed, Gallagher-Thompson, & Thompson, 2003). Thus, the development of a practice elements-based instrument that demonstrates score reliability and validity is needed.

Taking a step to address this need, Southam-Gerow et al. (2016) developed an observational instrument designed to capture adherence to common practice elements found in individual cognitive-behavioral therapy (ICBT) for youth anxiety (the Cognitive-Behavioral Therapy Adherence Scale for Youth Anxiety; CBAY-A). Scores on the instrument showed evidence of inter-rater reliability, convergent validity, and discriminant validity for use with an ICBT protocol (*Coping Cat*; Kendall & Hedtke, 2006a, 2006b). Taking the next step toward the goal of developing a practice elements-based instrument that would benefit implementation researchers, the current study aimed to evaluate the performance of the CBAY-A across two

separate CBT protocols for youth anxiety: a standard manualized treatment (*Coping Cat*; Kendall & Hedtke, 2006a, 2006b), and a modular treatment (*MATCH*; Chorpita & Weisz, 2005).

To be useful for implementation science research, the CBAY-A must produce scores that can be interpreted in ways intended by implementation researchers, requiring an examination of the instrument's score reliability and validity (Cook & Beckman, 2006). For scores produced by an observational instrument to be interpretable, observers must be able to code with adequate item-level reliability across both treatment protocols (Foster & Cone, 1995). While reliability alone is not a sufficient indicator of validity, it is a necessary condition (Devellis, 2017); therefore, the inter-rater reliability of the CBAY-A was examined.

The current study also examined the construct validity of CBAY-A scores to increase confidence that it is able to capture adherence to ICBT for youth anxiety. Researchers' ability to interpret and generalize scores produced by an instrument is contingent upon the instrument's construct validity (Foster & Cone, 1995). Scores produced by an adherence instrument that is not, in fact, assessing adherence (but instead another related construct) cannot be interpreted as intended, rendering it ineffectual for its intended use in implementation research. The current study explored two aspects of construct validity: convergent validity and discriminant validity. Convergent validity examines the relation between the instrument under examination and alternative instruments or ways of measuring the same construct, with high correlations exhibiting evidence for convergent validity (Campbell & Fiske, 1959; Foster & Cone, 1995). To use the CBAY-A for its intended purpose of capturing adherence to ICBT for youth anxiety across treatment protocols, it must be able to "replace" adherence instruments designed for individual protocols (e.g., *Coping Cat* and *MATCH*), requiring that the CBAY-A produce scores comparable to these instruments. Therefore, the current study examined the convergence

between the CBAY-A and two separate observational instruments of adherence – one designed to capture adherence to *Coping Cat* and the other designed to capture adherence to *MATCH*.

In contrast to convergent validity, discriminant validity determines whether the instrument is related to instruments of other independent constructs (Foster & Cone, 1995). The current study examined the extent to which the CBAY-A demonstrated the ability to discriminate between adherence and two related but distinct constructs: competence (Barber, Sharpless, Klostermann, & McCarthy, 2007; Perepletchikova & Kazdin, 2005) and alliance (Carroll et al., 2000; Hogue, Dauber et al., 2008; McLeod et al., 2018). Conceptually, adherence is distinct from both competence (McLeod et al., 2018; Perepletchikova & Kazdin, 2005) and alliance (Hogue, Dauber, et al., 2008; Southam-Gerow et al., 2016); however, past research has found that the degree of overlap between adherence and these related constructs varies (e.g., Barber, Liese, & Abrams, 2003; Carroll et al., 2000; Hogue, Dauber, et al., 2008). Thus, critical to establishing the score validity of the CBAY-A is to ensure that it is differentiated from competence and alliance ratings for the same treatment (Garland, Hurlburt, & Hawley, 2006; Hogue, Dauber, et al., 2008). The relations between the CBAY-A, an instrument of competence, and an instrument of alliance were investigated. Together, convergent and discriminant validity increase our ability to be confident that we are measuring the construct we intend to assess, and that we are not measuring any independent, unintended constructs (Cook & Beckman, 2006).

An additional aspect of validity that was examined in the current study is discriminative validity. Adherence instruments are often used to evaluate the effectiveness of an implementation strategy (e.g. training, supervision, consultation, treatment packaging; Powell et al., 2014), and to be useful for this purpose, scores should display evidence of discriminative validity. Discriminative validity of an adherence instrument, distinct from discriminant validity,

requires that an instrument is able to discriminate between groups expected to differ on scores of adherence, such as an intervention group (e.g., received training) and a control group (e.g., did not receive training; Foster & Cone, 1995) or two separate intervention groups (e.g. *Coping Cat* and *MATCH*). Past research has exposed treatment group differences in levels of adherence. For example, Weisz et al. (2012) found that therapists delivering *Coping Cat* displayed higher levels of adherence than therapists delivering *MATCH,* and that both therapists delivering *Coping Cat* and therapists delivering *MATCH* displayed higher levels of adherence than therapists delivering usual care. To assess the ability of the CBAY-A to discriminate between treatment groups and thus have utility in implementation science, the current study examined treatment group differences in levels of adherence.

In sum, this study evaluated the ability of the CBAY-A to: (1) Provide reliable scores of adherence across separate treatment protocols; (2) Capture adherence across two separate treatment protocols in ways that are convergent with instruments designed specifically for each protocol; (3) Assess adherence independently from commonly associated constructs (i.e., competence, alliance); and (4) Identify group differences in levels of adherence consistent with past research.

Literature Review

Youth Mental Health Problems

Most mental health disorders have their roots in childhood or adolescence, adversely impacting important developmental tasks such as establishing interpersonal relationships, completing academic requirements, and transitioning into the workforce (National Research Council and Institute of Medicine, 2009). Given that prevalence rates of mental health problems among children and adolescents are estimated at approximately 20-25% (Merikangas et al., 2010), youth mental health is deserving of substantial public health attention. Additionally, youth mental health problems have a significant economic impact in the United States, with associated costs approaching $250 billion per year (National Research Council, 2009). Despite the high prevalence rates and sizeable economic burden, only about 36% of affected youth receive mental health care (Merikangas et al., 2011), with even fewer (about 10%) receiving mental health services based in evidence (Bickman, 2008). Thus, there has been a push from a number of agencies and organizations (e.g. American Academy of Child and Adolescent Psychiatry, 2006; American Psychological Association Task Force on Evidence-Based Practice for Children and Adolescents, 2008; Institute of Medicine, 2015) to increase the use of evidence-based treatments in treating youth mental health problems.

Evidence-based Treatments for Youth

To improve the quality of care received by youth with mental health problems, recent efforts have focused on the development and delivery of evidence-based psychosocial treatments (American Psychological Association Task Force on Evidence-Based Practice for Children and Adolescents, 2008; Institute of Medicine, 2015; Weisz, Jensen-Doss, & Hawley, 2005). Evidence-based treatments (EBTs) are those that are based directly in scientific evidence

(Kendall & Beidas, 2007). Encouragingly, a large number of EBTs exist for a wide range of youth mental health problems including anxiety (Higa-McMillan, Francis, Rith-Najarian, & Chorpita, 2016), depression (Weersing, Jeffreys, Do, Schwartz, & Bolano, 2016), disruptive behaviors (Kaminski & Claussen, 2017), post-traumatic stress disorder (Dorsey et al., 2016), and attention-deficit/hyperactivity disorder (Evans, Owens, & Bunford, 2014).

While there are a considerable number of EBTs that have displayed efficacy in treating youth mental health problems (Weisz et al., 2017), the effectiveness of such treatments in community-based practice settings has been less consistent. Direct comparison of EBTs with usual care in community settings has shown modest mean effect sizes (Weisz et al., 2013). Further, some effectiveness studies have failed to find a difference in youth outcomes between the EBT being tested and usual care, and others have found that usual care produced more positive outcomes than the EBT (see Weisz et al., 2013 for a review). For example, CBT is the best-supported treatment for youth anxiety, with more than 45 published randomized-controlled trials demonstrating positive effects (Higa-McMillan et al., 2016). Yet, in multiple effectiveness studies CBT, was found to be no more effective than the usual clinical care provided within community settings (e.g., Barrington, Prior, Richardson, & Allen, 2005; Southam-Gerow et al., 2010). In order to understand this inconsistency, researchers have increasingly focused on the translation of EBTs to community settings, giving rise to the field of implementation science.

Implementation Science

Implementation science research has emerged with the purpose of understanding and closing the gap between clinical science and its application in community care settings. Implementation science identifies factors and methods involved in translating EBTs demonstrated to be efficacious in research settings to community-based practice settings (Proctor

et al., 2009). Through promoting the uptake of research findings and other evidence-based practices into routine practice, implementation science aims to improve the quality and effectiveness of community-based care (Eccles & Mittman, 2006).

An ongoing debate among researchers is how to best define implementation "success." Researchers have taken numerous approaches to measuring implementation success, including the measurement of clinical treatment outcomes (e.g. function and symptomatology), service system outcomes (e.g., efficiency), and implementation outcomes (e.g. feasibility, sustainability; Proctor et al., 2011). Proctor et al. (2011) argue for the use of implementation outcomes, defined as "the effects of deliberate and purposive actions to implement new treatments, practices and services" (p. 65) to conceptualize implementation success. Proctor et al. maintain that focusing solely on clinical outcomes ignores the process of implementation and the role of contextual factors, while the measurement of implementation outcomes allows for the identification of mechanisms and causal relationships within implementation processes (Proctor et al., 2011).

The implementation outcomes set forth by Proctor and colleagues (2011) include acceptability (the degree to which stakeholders view an intervention as agreeable, palatable, or satisfactory), adoption (intention to employ an intervention), appropriateness (perceived fit of an intervention for a given setting, provider, or consumer), cost (the fiscal impact of an implementation effort), feasibility (the extent to which an intervention can practically be employed), fidelity (often referred to as treatment integrity or treatment adherence; the degree to which an intervention was delivered as intended), penetration (the integration of an intervention within a setting and its subsystems), and sustainability (the extent to which a newly implemented intervention can be maintained within a service setting). Proctor and colleagues argue that core to implementation research, then, is the measurement of these implementation outcomes.

Implementation Outcome Measurement

Implementation outcomes have been lauded as potentially the most critical factor in implementation science; however, the field is lacking in high quality instruments of implementation outcomes (Lewis, Stanick, et al., 2015). High quality instrumentation (i.e. instruments that have displayed evidence of score reliability and validity across multiple studies, samples, and settings; Martinez et al., 2014) is critical to the advancement of any field, particularly new fields such as implementation science (Cook & Beckman, 2006; Lewis, Stanick, et al., 2015; Rabin et al., 2012). Without such instruments, we are unable to empirically test and compare the effectiveness of various implementation strategies and thus, we are unable to infer implementation "success" (Proctor et al., 2009; Proctor et al., 2011). Additionally, measurement aids in the refinement and understanding of constructs within implementation science (Cook et al., 2012). While the development and utilization of high quality instruments for the field of implementation science has been designated as a high priority (Lewis, Stanick, et al., 2015; Rabin et al., 2012), recent reviews have demonstrated a lack of existing instruments with established psychometric properties, potentially limiting the identification and use of various implementation strategies for advancing the implementation of EBTs.

A number of reviews exploring the score reliability and validity of existing implementation-related instruments have been conducted that underscore the lack of high quality instruments of implementation-related constructs (e.g., Emmons, Weiner, Fernandez, & Tu, 2012; Scott, Mannion, Davies, & Marshall, 2003; Weiner, Amick, & Lee, 2008). In a systematic review of instruments designed to assess constructs that may predict implementation outcomes (e.g. organizational level factors, provider level factors), Chaudoir, Dugan, and Barr (2013) found that of the 62 identified instruments, 30 did not display any evidence of criterion validity

(i.e., whether scores on the measure are predictive of implementation outcomes). Similarly, Chor et al. (2014) conducted a systematic review of 118 instruments designed to predict the implementation outcome of adoption, revealing that only 52.5% exhibited any established score reliability and validity.

A more recent review, conducted by Clinton-McHarg et al. (2016) examined the score reliability and validity of instruments which have been developed for use in community-based non-clinical settings. The instruments included in this review were specifically designed to assess constructs aligned with the Consolidated Framework for Implementation Research (CFIR), a framework that contains common constructs from implementation theories (see Damschroder et al., 2009 for a detailed description). Assessing for multiple psychometric dimensions including content validity, construct validity (convergent, discriminant), criterion validity, internal consistency, and test-retest reliability, Clinton-McHarg et al. found that of 51 unique instruments, almost all (n = 47) had undergone a process of content validation, but that only two instruments reported having examined convergent and/or discriminant validity. Further, it was found that only one measure produced scores that displayed evidence of adequate convergent validity ($r > 0.40$) and discriminant validity ($r < 0.30$; Cohen, 1988) with external instruments. Eight instruments were examined for scores that supported criterion validity for sub-populations with known differences. While 50 of the 51 instruments reported on the internal consistency of *either* the total score or individual domains, the internal consistency of the total scale was reported for only five instruments, and the internal consistency of *both* the total scale and the domains was reported for only four instruments. Additionally, 50% of the instruments did not meet the acceptable threshold (Cronbach's alpha (α) > 0.70; Lohr et al., 1996) for all of

their domains. Finally, only three instruments were examined for test-retest reliability. No single instrument reported on all key psychometric quality indicators.

As part of the Society for Implementation Research Collaboration's (SIRC) Instrument Review Project, Lewis, Stanick, et al. (2015) conducted a systematic review assessing the score reliability and validity of 104 instruments of implementation outcomes directly applicable to mental and behavioral healthcare settings. Lewis, Stanick, et al. used the previously developed evidence-based assessment rating criteria (see Lewis, Fischer, et al., 2015 for a description) to rate the psychometric quality of each instrument. They found that 47% of the identified instruments were rated as "good" or "excellent" for reliability, 17% for structural validity, and 9% for predictive ability. Overall, across the 104 identified instruments, 46% were missing information on four or more of the psychometric properties under examination.

It is clear that implementation science is in need of more well-established instruments of implementation outcomes that display strong psychometric properties. A lack of these instruments presents "a formidable barrier to the efforts to advance implementation science" (Chaudoir et al., 2013, p. 14), preventing accurate interpretations of data and limiting our ability to characterize and understand implementation strategy effectiveness and implementation "success" (Martinez et al., 2014).

Treatment Integrity

One of the most commonly studied implementation outcomes is treatment integrity (also termed treatment fidelity, implementation integrity, implementation fidelity; Powell et al., 2014), which refers to the degree to which an intervention was delivered as designed (Allen et al., 2012). Treatment integrity comprises adherence (the delivery of therapeutic interventions specified by a particular treatment), competence (a therapist's skill and responsiveness in

delivering an intervention), and differentiation (the delivery of interventions not prescribed by a specific treatment; Bellg et al., 2004; McLeod et al., 2013; Perepletchikova et al., 2007).

Historically, treatment integrity measurement has been used as an independent variable check which is critical in achieving adequate internal validity and allowing researchers to make valid inferences from clinical trials, such that conclusions about the efficacy or effectiveness (or lack thereof) of a particular intervention can be justifiably made (Perepletchikova & Kazdin, 2005; Weisz et al., 2005). In fact, it has been suggested that low treatment integrity is the most common reason for unfavorable clinical outcomes (Mills & Ragan, 2000). Making conclusions about the effectiveness of treatment A based on clinical outcomes when, in reality, treatment B was delivered leads to what Dobson and Cook (1980) refer to as "Type III error." Additionally, if the delivered treatments are not well-specified, well-tested, and carried out as intended, the ability of researchers to replicate study findings is hindered, limiting their ability to evaluate external validity (Freeman & Rossi, 1985; Kazdin, 1994; Perepletchikova et al., 2007).

Rather than measuring treatment integrity to support internal validity, the field of implementation science considers treatment integrity to be an outcome of interest, often conceptualized as an indicator of implementation success (Proctor et al., 2011). For example, if a treatment failed to produce desired clinical outcomes when transferred from a controlled research setting to a community setting, assessing the extent to which an intervention was delivered as intended aids implementation scientists in determining whether the lack of intervention effectiveness was a result of insufficient treatment or insufficient treatment integrity (McLeod et al., 2013). This gives implementation scientists the ability to try various implementation strategies (e.g., training, supervision) if, in fact, insufficient treatment integrity is to blame. Additionally, the use of treatment integrity instruments can identify and prevent program drift

during effectiveness trials (Bond et al., 2000), and within multi-site trials, treatment integrity measurement on an ongoing basis can help ensure that treatment delivery is consistent across sites (Paulson, Post, Herinckx, & Risser, 2002). Because competence presupposes adherence (that is, treatment must be delivered before its quality can be judged), adherence is the aspect of treatment integrity most commonly assessed by implementation researchers (Hagermoser Sanetti, Gritter, & Dobey, 2011; Perepletchikova et al., 2007). Therefore, this study will focus specifically on instruments of adherence.

Lack of Adherence Measurement

Despite receiving greater attention that the other components of treatment integrity (i.e. competence and differentiation), the frequency of adherence measurement within clinical trials has been low, as indicated by several reviews of treatment outcome studies (Cox, Martinez, & Southam-Gerow, in press; Goense, Boendermaker, van Yperen, Stams, & van Laar, 2014; Hagermoser Sanetti et al., 2011; Perepletchikova et al., 2007). A review of 226 youth treatment outcome studies published between the years 1962 and 2002 revealed that only 32% of studies noted any procedures for monitoring adherence (Weisz et al., 2005). This review included randomized trials involving the treatment of youth anxiety, depression, ADHD and related conditions, and conduct-related problems and disorders. Within these four problem areas, trials of treatment for youth anxiety had the fewest studies with adherence checks (24.4%).

Similarly, a review of 147 child and adult treatment studies published between the years 2000 and 2004 conducted by Perepletchikova et al. (2007) resulted in comparable findings, revealing that very few studies adequately measured treatment integrity. “Adequacy” was determined using the Implementation Treatment Integrity Procedures Scale (ITIPS), developed by Perepletchikova et al., which was designed to evaluate the extent to which randomized

controlled trials established, assessed, evaluated, and reported adherence and competence. This review revealed that the measurement of treatment integrity procedures (represented by the total score on the ITIPS) was adequate for 3.5% of studies. Treatment integrity was: (1) established adequately for 15.8% of studies; (2) assessed adequately for 3.5% of studies; (3) evaluated adequately for 2.0% of studies; and (4) reported adequately for 6.4% of studies. When evaluated separately from competence, treatment adherence procedures were implemented across treatments 8.9% adequately (vs. 1.5% for competence). Several prior reviews of treatment outcome studies exhibited the same pattern in the measurement of adherence (e.g., Kazdin, Bass, Ayers, & Rogers, 1990; Moncher & Prinz, 1991; Yeaton & Sechrest, 1981). Although methodologically challenging, measurement of treatment adherence is critical to advancing the field of implementation science.

Existing Treatment Adherence Instruments

The current and most common approach to treatment adherence measurement is to use an adherence instrument designed for a specific treatment (often an evidence-based manualized treatment) that is being investigated within an individual study, with the purpose of capturing the degree to which a treatment is delivered (Garland & Schoenwald, 2013). However, there are several shortcomings to this approach.

First, this approach can require substantial resources for the creation and utilization of redundant instrumentation (i.e., there are existing instruments of the same constructs). In order to develop an adherence instrument with strong score reliability and validity, researchers need to dedicate considerable time, effort, and financial resources. For example, the creation of an adherence instrument requires a researcher to (1) define treatment adherence and generate items, often through expert consultation or qualitative research (Moncher & Prinz, 1991; e.g., Barber et

al., 2003; Southam-Gerow et al., 2016); (2) examine the score reliability and validity of the instrument; and if the psychometric properties are insufficient, (3) make adjustments to the instrument and repeat the process. Further, the development and use of observational adherence instruments (considered the gold standard; Hogue, Liddle, & Rowe, 1996) may require additional steps including (1) developing a rating protocol which often involves multiple iterations; (2) training coders to use the protocol; (3) establishing adequate initial inter-rater reliability; and (4) evaluating ongoing inter-rater reliability and providing additional training if necessary. The cost and resources associated with the development and use of such instruments may deter researchers from sufficiently measuring treatment adherence (Perepletchikova et al., 2007).

Additionally, because the majority of adherence instruments are rarely used outside of the study for which they were developed, little attention is paid to the establishment and evaluation of psychometric properties. In fact, it has been purported that few existent adherence instruments have evidence supporting their score reliability and validity, and that the literature is lacking in attempts to establish such instruments (Baer et al., 2007; Garland & Schoenwald, 2013; Perepletchikova et al., 2007). In a review of 249 unique adherence instruments, Schoenwald and Garland (2013) revealed that researchers reported evidence for score reliability and validity for fewer than 5% of the instruments. This is problematic for the field of implementation science because without evaluation of the psychometric properties of adherence instruments implementation researchers cannot be confident in their interpretations of study findings, hindering progress in the field (Martinez et al., 2014). Furthermore, to prevent the continued development of redundant instrumentation or the selection and use of inadequate instruments, it

is helpful to have evidence of score reliability and validity for existing adherence instruments widely available to researchers (Lewis, Stanick, et al., 2015).

Another disadvantage of the current approach is that it is not conducive to conducting cross-study and cross-treatment comparisons, critical for building the implementation science knowledge base and moving the field forward (Drake et al., 2001; Martinez et al., 2014; Schoenwald et al., 2011). For example, if two studies are conducted in two different settings (e.g., community mental health agency and primary care setting) examining the effectiveness of "Defiant Children" (an evidence-based Behavioral Parent Training treatment protocol designed for the treatment of disruptive behaviors in children; Barkley, 2013), the levels of adherence between the two settings could not be justifiably compared unless they were to use the same instrument of adherence. Similarly, for comparisons of adherence to two separate Behavioral Parent Training protocols to be legitimately interpreted, the use of a single adherence instrument would be required. Thus, the field would benefit from available adherence instruments that can produce reliable and valid scores across settings and treatment protocols to promote knowledge accumulation resulting from cross-treatment and cross-study comparisons.

The first step in the development of such instruments may be to identify commonalities among existing instruments for a given problem area. Schoenwald et al. (2011) took this approach when they examined the extent to which treatment adherence instruments for a given problem area overlapped and therefore potentially facilitated cross-treatment and cross-study comparisons. Specifically, they assessed the similarities within and differences between instruments of adherence to EBTs designed to target disruptive behavior disorders in children. Authors compiled items from existing adherence instruments with hopes of developing a classification scheme to evaluate whether and how frequently functionally similar or identical

items were found across instruments. Schoenwald et al. (2011) found such variability between instruments that they were unable to establish dimensions along which the items could be classified, nor could they evaluate the extent to which adherence to any elements or procedures exemplified in the treatment programs could be understood using a single frame of reference (i.e., conceptual consensus regarding criteria against which adherence instruments can be judged) or shared language, restricting the ability to make comparisons between instruments. Consequently, implementation science research would benefit from the development of instruments for a given problem area using a single frame of reference.

Design. While the majority of existing adherence instruments have the same purpose (i.e., capturing the degree to which an intervention was delivered as intended), they vary considerably with respect to their design. For example, the level of specification of such instruments can be molar (e.g. items consistent with the components found in a specific treatment manual) or molecular (e.g. items representing discrete principles or skills found across treatments; McLeod et al., 2013). Schoenwald (2011) maintains that adherence instruments often use a "form follows function" rule whereby treatments with session-by-session manuals often use adherence instruments with similar amounts of detail, while treatments consisting of component parts or classes of therapeutic interventions often give rise to adherence instruments with fewer and less detailed items. Furthermore, some instruments of adherence include only items representing prescribed interventions, while others include the addition of non-prescribed interventions (Schoenwald, 2011).

Assessment method. Adherence instruments also vary in the way adherence data are collected. Two of the most common approaches involve the use of direct observation (usually via

audio or video recordings) and self-report instruments (Mowbray, Holter, Teague, & Bybee, 2003).

Direct observation. Direct observation uses trained coders, preferably independent from the intervention team, to observe and quantify an individual's behavior (Sheridan, Swanger-Gagne, Welch, Kwon, & Garbacz, 2009). The use of direct observation using coders who have been trained in a coding system has been considered the gold standard in clinical trials, as it is argued that direct observation produces ratings of therapist behavior that are objective and thus more accurate than self-report instruments (Hogue et al., 1996; Mowbray et al., 2003). However, due to the costs and resources associated with training and using multiple coders who most often conduct several observations, some argue for the use of self-report instruments (Sheridan et al., 2009).

Self-report. Self-report instruments are used to assess treatment adherence as perceived by the individual who delivered the intervention, and are often in the form of checklists wherein the reporter records the completion of treatment components on an intervention-specific checklist (Sheridan et al., 2009). Although some research has suggested that self-report instruments of adherence result in overestimation of adherence (e.g. Carroll, Martino, & Rounsaville, 2010; Hurlburt, Garland, Nguyen, & Brookman-Frazee, 2010), there are advantages to self-report instruments including increased efficiency and decreased utilization of human and financial resources (Sheridan et al., 2009). Despite the practical benefits of self-report instruments, direct observational instruments are recommended for providing reliable and valid data (Hogue et al., 1996).

Scoring. Adherence instruments also vary in terms of their scoring systems. Various scoring approaches include presence or absence of a component of an intervention, frequency of

component delivery, and Likert-type quantitative ratings of the delivery extensiveness (Hogue et al., 1996; Waltz, Addis, Koerner, & Jacobson, 1993). It has been contended that rating the extensiveness of the delivery of an intervention more accurately captures treatment adherence than presence/absence or frequency ratings (Hogue et al., 1996; McLeod et al., 2013).

Extensiveness has been described as the degree to which therapists deliver an intervention during a session (Hogue et al., 1996; Southam-Gerow et al., 2016). Extensiveness, generally rated on a 7-point Likert-type scale, consists of both thoroughness (i.e., the persistence and intensity with which an intervention is delivered) and frequency (i.e., the number of instances in which a given intervention occurs; Hogue et al., 1996). While there is not an established algorithm for how to incorporate both frequency and thoroughness in a rating of extensiveness, it has been suggested that coders rely on item descriptions in the coding manual, the training they received, and familiarity with the coding system to make decisions about weighting the two dimensions in scoring extensiveness (Hogue et al., 1996). It is largely agreed that more complex methods (e.g. extensiveness ratings rather than presence/absence) are more comprehensive and more likely to capture contingent and situational influences (Julien, Markman, & Lindahl, 1989; Markman, Leber, Cordova, & St. Peters, 1995). Thus, extensiveness ratings are often considered to be the gold-standard (Hogue et al., 1996).

Conceptual Shift in Measurement

Recent proposals in the field have emerged which suggest a shift in the definition of treatment integrity, wherein the focus shifts from capturing components of specific manualized treatments to capturing broader evidence-based principles, often termed “practice elements” (McLeod et al., 2013; Schoenwald et al., 2011). The suggestion that treatment adherence instruments, which have historically captured adherence at a molar-level (e.g., whether

prescribed components of a specific treatment protocol are delivered), should instead be crafted using a more molecular definition of treatment adherence (e.g., focusing on the extent to which various practice elements are present; McLeod et al. 2013), is consistent with the distillation and matching model set forth by Chorpita, Daleiden, and Weisz (2005), also referred to as the "common elements approach" (e.g., Barth et al., 2012).

Distillation and Matching Model. The Distillation and Matching Model (DMM) is a literature analysis procedure wherein specific strategies and techniques that are found across a set of treatments are defined and extracted (or distilled) and subsequently "matched" to the unique characteristics and context of the client (Chorpita, Daleiden, & Weisz, 2005). The distilled practices are often referred to as "practice elements," defined by Chorpita et al. (2005) as "a discrete clinical technique or strategy used as part of a larger intervention plan" (p. 11). The DMM involves the aggregation of information across all treatments relevant to a particular problem area and group (e.g. anxiety in middle childhood) rather than relying only on individual specific treatment protocols (Boustani et al., 2017).

The argument set forth for a practice elements-based approach to treatment adherence is based on the distillation portion of the model, designating the need for adherence instruments whereby interventions are conceptualized as an amalgam of practice elements rather than single units of analysis specific to a unique treatment protocol (McLeod et al., 2013). A focus on individual practices rather than treatment programs would allow implementation researchers to construct a more flexible adherence instrument for use across treatment programs, increasing efficiency and reducing cost. Instead of defining adherence as the extent to which the prescribed components of a specific treatment protocol are delivered, it is suggested that an instrument that

captures data at the practice element level is more appropriate for the current goals of implementation science.

Practice elements in adherence measurement. It has been maintained that the conceptual shift from a molecular to molar understanding of treatment adherence and associated instruments will better fit the needs of implementation researchers. Specifically, implementation researchers would be able to capture adherence to a wide array of practice elements for a particular problem area (e.g., cognitive restructuring for anxiety, praise for disruptive behavior) rather than being limited only to the interventions found in a specific treatment manual, consistent with recent developments in the field (Chorpita & Daleiden, 2009). For example, there are a number of EBTs for childhood anxiety (e.g. *Coping Cat*; Kendall, 1994; *Coping Koala*; Barrett, Dadds, & Rapee, 1996; *Cool Kids*; Schniering et al., 2006), with most involving cognitive components; although they vary in how this component is delivered. If an instrument was to capture the different cognitive components from each treatment manual, there would be an unmanageable number of items and thus would be less useful to implementation researchers. Instead, then, is the suggestion to define adherence at the practice element level. In the current example, the practice element that may capture the cognitive component across several treatment manuals might be "cognitive strategy" or "cognitive restructuring," producing a more efficient instrument that is conducive to use in implementation research, which is increasingly concerned with whether treatment delivered in community practice settings adhere to evidence-based principles rather than specific treatment programs or protocols (McLeod et al., 2013).

Benefits to implementation science. Defining treatment integrity in terms or practice elements rather than components of a specific treatment program offers several advantages to implementation research.

Efficiency. An approach which defines adherence as the presence of practice elements as opposed to individual components of a specific treatment protocol would require fewer adherence instruments, reducing redundancy (Schoenwald et al., 2011). For example, Chorpita and colleagues (2005) identified six common practice elements across 25 separate treatment protocols for youth anxiety; using the current approach to treatment adherence measurement, this would require 25 separate instruments. However, with a practice elements-based approach, this may be reduced down to one instrument with six items. Additionally, a significant provision of resources is involved in the development and use of adherence instruments, including the establishment of coder competence which requires training on all items present on the observational instrument. In a clinical trial comparing two separate anxiety treatments, a coder may be able to be trained on one practice elements-based instrument rather than two protocol-specific instruments, reducing the associated time and costs as well as inconsistency in similar items across instruments that may lead to confusion and coding inaccuracy.

Flexibility. The design of practice elements-based treatment adherence instruments allows for more flexible use across CBT protocols, making them more widely applicable across a variety of service settings where the specific CBT protocols delivered often differ (McLeod et al., 2013). Additionally, contrary to instruments for individual treatment protocols, a practice elements-based adherence instrument can more easily be adapted to reflect developments in the evidence base (McLeod et al., 2013). Treatment protocols are often revised and updated with small changes, consequently requiring the same of any associated program-specific adherence instruments. However, due to its broad nature, a practice elements-based instrument is less likely to require modification with each iteration of a treatment manual. For example, if the specific cognitive strategy prescribed by a treatment protocol changes, a protocol-specific adherence

instrument would need to be altered to reflect those changes, while a practice elements-based instrument containing the item "cognitive strategy" would not require modification.

Comparisons. Instruments that are designed to be specific to one intervention within a specific setting are not broadly relevant to the field of implementation science. Inferences that can be made from such instruments are restricted to a single treatment in a single study, limiting our ability to make cross-study and cross-treatment comparisons (Malik et al., 2003). The use of practice elements-based adherence instrument can be used more generally across treatment protocols, enhancing comparability of findings across studies and enhancing our ability to make inferences from aggregated data that are more widely applicable to the field of implementation science. (Moser et al., 2011; Rabin et al., 2012).

Psychometric properties. It has been established that adherence instruments with evidence of score reliability and validity are lacking in the field of implementation science, threatening the quality of implementation research (see Cox et al., in press; Goense et al., 2014; Perepletchikova et al., 2007). Additionally, even if an instrument has displayed evidence of score reliability and validity, it is often investigated in only one study, limiting our confidence in the quality of the instrument. A practice-elements based instrument of adherence that can be used across multiple studies and settings would provide more opportunities for psychometric evaluation across contexts and populations, increasing our confidence in the performance of the instrument.

In sum, the development of an adherence instrument that specifies treatment components sufficiently broadly such that they apply across multiple treatment protocols is needed for the progression of implementation science.

Yale Adherence and Competence Scale. The Yale Adherence and Competence Scale (YACS; Carroll et al., 2000) exemplifies such an adherence instrument. The YACS is a rating system that was designed to assess general interventions for substance use disorders, with the intention of being used across treatment protocols and studies. The YACS includes three scales containing items that are common across most behavioral treatments for substance abuse disorders, as well as three scales that assess components of specific EBTs that are often utilized as control or comparison treatments in clinical trials, allowing for use across a wide range of studies. All six scale scores on the YACS displayed excellent inter-rater reliability, with intraclass correlation coefficients (ICCs) ranging from .80 to .95 for the adherence ratings and from 0.71 to 0.97 for the competence ratings. At the item level, ICCs ranged from .28 to .84 for the adherence ratings and from .06 to .81 for competence ratings. Additionally, the six scales fit the hypothesized factor structure, and evidence supported concurrent validity, discriminant validity (with competence scores), and discriminative validity (finding expected differences between treatment groups) of scale scores (Carroll et al., 2000). While the YACS is an example the type of instrument argued for throughout this review, it is only applicable to substance abuse research with adults. There is a need for similar instruments created across various problem areas; therefore, the problem area of focus of the current study will be youth anxiety.

Necessary Psychometric Properties

The development of practice elements-based adherence instruments may facilitate the progression of implementation science; however, for such an instrument to be useful, it must produce scores that can be interpreted as intended, requiring the establishment of certain psychometric properties, including reliability and validity dimensions (Cook & Beckman, 2006; Downing & Haladyna, 1997). Martinez et al. (2014) provide recommendations for how to best

evaluate psychometric properties to maximize the confidence that can be placed in study findings and interpretations. First, they recommend that researchers report on the appropriate measure of reliability (i.e. the consistency of scores obtained from an administered instrument) given the design of the instrument. Reliability assessment may include internal consistency, test-retest reliability, or inter-rater reliability. For instruments involving multiple observational coders, inter-rater reliability is considered to be the most relevant measure of reliability (Martinez et al., 2014). Second, Martinez et al. discuss the importance of validity (e.g., convergent, discriminant) assessment, noting that it has been considered a crucial, and perhaps the most important, consideration in instrument evaluation (Downing & Haladyna, 1997).

Inter-rater reliability. First, for instruments that require ratings from observers, inter-rater reliability (i.e. consistency among observational ratings provided by multiple coders; Hallgren, 2012) should be examined. The assessment of inter-rater reliability identifies how much score variance is due to the actual behaviors being observed and how much is a result of coder characteristics (Hallgren, 2012; Novick & Lewis, 1966). Thus, high inter-rater agreement indicates that the instrument will perform equally when rated by different coders, as only a small portion of the variance is attributable to coder differences.

Convergent Validity. Convergent validity refers to the extent to which scores produced by the instrument under investigation are related to scores on alternate instruments measuring the same construct, with high correlations representing convergence (Foster & Cone, 1995). For a practice elements-based adherence instrument to be useful to implementation researchers, it must be able to be substituted for existing adherence instruments that are designed for individual treatment protocols. Specifically, a researcher must be able to have confidence that a practice elements-based adherence instrument is capturing adherence to various treatment protocols in

ways comparable to instruments designed specifically for those protocols. High convergence between a practice elements-based instrument and instruments for individual protocols, then, would suggest that a researcher can make similar interpretations of scores produced by each instrument. Low convergence, however, may indicate that the practice elements-based instrument is not capturing adherence in a way similar to the individual adherence instruments, suggesting that interpretations made from such an instrument cannot be applied across individual treatment protocols, restricting its relevance to implementation researchers.

Discriminant Validity. Discriminant validity refers to the ability of an instrument to discriminate between independent constructs (Foster & Cone, 1995). In order for implementation researchers to have confidence in their interpretations of an adherence instrument, the scores produced must be a representation of adherence rather another related, but distinct, construct. Most relevant to instruments of adherence is their relation to instruments of competence and the alliance.

Competence. Competence, another component of treatment integrity, is considered to be conceptually distinct from adherence as it represents a therapist's skill and responsiveness in delivering an intervention, whereas adherence represents the frequency and thoroughness with which an intervention is delivered (Perepletchikova et al., 2007). However, the extent to which these two components overlap has varied in past research (e.g. Barber et al., 2003; Carroll et al., 2000; McLeod et al., 2018). To be useful to implementation researchers, an adherence instrument should produce scores that can be interpreted as adherence, rather than the related construct of competence.

Alliance. Past research has also found relations, albeit smaller than those with competence, between instruments of adherence and instruments of the alliance (Hogue, Dauber,

et al., 2008; Southam-Gerow et al., 2016). The alliance is commonly considered to be made up of bond (i.e., the affective aspects of the client-therapist relationship) and task (i.e., client participation in the activities of treatment; McLeod & Weisz, 2005), and is considered to be a related but distinct aspect of the therapy process (Hogue, Dauber, et al., 2008; Southam-Gerow et al., 2016). Past research has found small to moderate correlations with instruments of adherence (e.g., Carroll et al., 2000; Hogue, Dauber, et al., 2008; Southam-Gerow et al., 2016). Thus, similar patterns of correlations would provide evidence for the conclusion that an adherence instrument is not, in fact, capturing alliance rather than adherence.

Discriminative Validity. Within the field of implementation science, adherence instruments are commonly used to evaluate the effectiveness of an implementation strategy (Powell et al., 2014). For example, to evaluate the effectiveness of a training program, implementation researchers may use an adherence instrument to evaluate the extent to which a therapist delivered an intervention in which they were trained in a clinical setting (Beidas & Kendall, 2010). To be used in this way, an instrument should be able to discriminate between therapists who received training and therapists who did not. If an instrument is able to discriminate between groups expected to differ, it is said to display evidence of discriminative validity (Foster & Cone, 1995). Past research using established instruments of adherence to cognitive-behavioral interventions have found group differences in levels of adherence between therapists delivering cognitive-behavioral interventions and therapists delivering usual care (McLeod et al., 2017; Southam-Gerow et al., 2010; Southam-Gerow et al., 2016; Weisz et al., 2009). Levels of adherence have also been shown to differ between separate treatment protocols delivered within a single clinical trial. For example, using two separate instruments of adherence (one designed for *Coping Cat*, an individual CBT protocol for youth anxiety, and the other

designed for *MATCH*, a modular manualized treatment designed to address anxiety, depression, and conduct problems; Chorpita & Weisz, 2005), Weisz et al. (2012) found that in community settings, therapists delivering *Coping Cat* obtained higher scores of treatment adherence than therapists delivering *MATCH*. For a practice elements-based adherence instrument to be equally as useful as adherence instruments designed for individual treatment protocols, scores should display similar evidence of discriminative validity. Specifically, scores should reveal group differences similar to those found using treatment protocol specific adherence instruments.

Current Study

Cognitive-Behavioral Therapy Adherence Scale for Youth Anxiety (CBAY-A). The CBAY-A is an observational adherence instrument created by Southam-Gerow and colleagues (2016) with the purpose of capturing the delivery of common practice elements found in individual cognitive behavioral therapy (ICBT) for youth anxiety. The scoring approach of the CBAY-A is consistent with the recommendation of Hogue et al. (1996), producing an extensiveness score on a 7-point Likert-type scale. The CBAY-A item and scale scores have displayed initial evidence of inter-rater reliability, convergent validity, and discriminant validity for use with *Coping Cat* (Kendall & Hedtke, 2006a, 2006b), an evidence-based manualized treatment (ICBT) of youth anxiety, delivered in an efficacy (Kendall et al., 2008) and an effectiveness trial (Southam-Gerow et al., 2010). Additionally, the utility of the CBAY-A was supported by its use in determining whether therapists in community settings achieved benchmark levels of adherence to *Coping Cat* set by therapists in a research clinic, revealing expected differences wherein community therapists did not achieve benchmark scores (McLeod et al., 2017).

As previously argued, implementation science is often interested in cross-treatment and cross-study comparisons. Thus, the current study evaluated the performance of the CBAY-A across two separate ICBT protocols for youth anxiety in order to determine its potential to be used across treatments and thus, across studies. The CBAY-A was used with a standard manualized treatment (*Coping Cat*; Kendall & Hedkte, 2006), and a modular manualized treatment (*MATCH*; Chorpita & Weisz, 2005).

Using data from a large randomized controlled effectiveness trial comparing two separate protocols for youth anxiety (*Coping Cat* and *MATCH*) and usual clinical care, this study examined the score reliability and validity of the CBAY-A. To be useful for implementation research, the CBAY-A must produce scores that researchers can interpret for their intended use, requiring an examination of the instrument's score reliability and validity (Cook & Beckman, 2006). Thus, the inter-rater reliability, convergent validity, discriminant validity, and discriminative validity of the CBAY-A were examined across the two protocols. For the purpose of the current study (i.e., to examine the performance of the CBAY-A across two separate treatment protocols) two subscales were created: one to represent adherence to *Coping Cat* and one to represent adherence to *MATCH*, and all analyses were conducted for the *Coping Cat* and *MATCH* treatment groups separately.

Hypotheses

It was expected that the CBAY-A, when used to assess levels of adherence to two separate ICBT treatment protocols for youth anxiety delivered in an effectiveness trial, would display score reliability and validity comparable to those found by Southam-Gerow et al. (2016) in their investigation of the performance of the CBAY-A when used with *Coping Cat* delivered in efficacy and effectiveness trials.

Hypothesis 1. To make valid interpretations of scores produced by an observational instrument, it is critical that variance due to coder characteristics is minimal; thus, the CBAY-A was examined for evidence of inter-rater reliability. Because the CBAY-A is rated on a 7-point Likert-type extensiveness scale, the most appropriate measure of inter-rater reliability is intra-class correlations (ICCs; Shrout & Fleiss, 1979). It was hypothesized that all item and subscale scores on the CBAY-A would demonstrate at least *good* levels of inter-rater reliability (ICC(2,2) ≥ .60; Cicchetti, 1994) for both the *Coping Cat* group and the *MATCH* group.

Hypothesis 2. Correlations between the CBAY-A *Coping Cat* and *MATCH* subscale scores and two adherence instruments designed specifically for *Coping Cat* (Standard Consultation Record) and *MATCH* (MATCH Consultation Record) were examined as support for convergent validity. It was hypothesized that the *Coping Cat* subscale scores produced by the CBAY-A would be correlated at high ($r \geq .36$; Rosenthal & Rosnow, 1984) levels with scale scores produced by the Standard Consultation Record when used with the *Coping Cat* group. It was also hypothesized that *MATCH* subscale scores produced by the CBAY-A would be correlated at high ($r \geq .36$; Rosenthal & Rosnow, 1984) levels with scale scores produced by the MATCH Consultation Record when used with the *MATCH* group.

Hypothesis 3a. Correlations between the CBAY-A *Coping Cat* and *MATCH* subscale scores, and corresponding subscale scores on the Cognitive-Behavioral Treatment for Anxiety in Youth Competence Scale (CBAY-C; McLeod et al., 2018), an instrument of therapist competence, were examined for evidence supporting the ability of scores produced by the CBAY-A subscales to discriminate between adherence and competence. Past research has found moderate to high correlations (ranging from .22 to .65; $M = .46$, $SD = .17$) between original CBAY-A subscale and scale scores (i.e., skills, exposure, and total scale) and the corresponding

CBAY-C subscale and scale scores (McLeod et al., 2018). Based on past findings, it was hypothesized that both the *Coping Cat* and *MATCH* CBAY-A subscale scores would correlate at moderate levels ($0.24 \leq r < 0.36$; Rosnow & Rosenthal, 1984) with corresponding CBAY-C subscale scores when used with their respective treatment groups (*Coping Cat* or *MATCH*). It was further hypothesized that the correlations between instruments of adherence and competence would be significantly lower than the correlations between the two instruments of adherence in both treatment groups.

Hypothesis 3b. Correlations between the CBAY-A subscale scores and scale scores produced by the Therapy Process Observational Coding System for Child Psychotherapy—Alliance (TPOCS-A; McLeod & Weisz, 2005), an instrument designed to assess the client-therapist alliance, were examined for evidence supporting the ability of scores produced by the CBAY-A subscales to discriminate between adherence and alliance. Southam-Gerow et al. (2016) found moderate correlations (ranging from .26 to .45; $M = .36$, $SD = .07$) between original CBAY-A scale scores and TPOCS-A scale scores. Based on past findings, it was hypothesized that both the *Coping Cat* and *MATCH* CBAY-A subscale scores would correlate at moderate levels ($0.24 \leq r < 0.36$; Rosnow & Rosenthal, 1984) with TPOCS-A scale scores when used with their respective groups (*Coping Cat* or *MATCH*). It was also hypothesized that the correlations between instruments of adherence and alliance would be significantly lower than the correlations between the two instruments of adherence in both treatment groups.

Hypothesis 3c. As further support for discriminate validity, I examined whether the CBAY-A *Coping Cat* and *MATCH* subscales display levels of discriminant validity (examined through correlations with CBAY-C subscales and the TPOCS-A) similar to those of instruments designed specifically for each protocol (Standard Consultation Record and MATCH

Consultation Record). It was hypothesized that the CBAY-A subscales would display similar levels and patterns of discriminant validity as the Standard and MATCH Consultation Records.

Hypothesis 4. To examine if the CBAY-A subscales are able to discriminate between groups expected to differ on levels of adherence, group mean differences were explored. It was hypothesized that when each treatment group (*Coping Cat* and *MATCH*) was rated using their respective CBAY-A subscales, group differences would reflect those found using the Standard and MATCH Consultation Records (Weisz et al., 2012); specifically, the *Coping Cat* group would have higher levels of adherence than the *MATCH* group.

Method

Data Source

The current study used data from a larger randomized effectiveness trial (Child STEPs; Weisz et al., 2012) that included 174 youth (*M* age = 10.59; 30% female, 45% Caucasian) and 84 therapist (*M* age = 40.60; 80% female; 56% Caucasian) participants. The Child STEPs study was a project conducted in 10 outpatient clinical service organizations in Massachusetts and Hawaii to compare the effectiveness of two separate treatment program designs in treating youth diagnosed with anxiety, depression, or disruptive behavior problems. Study therapists and youth participants were randomized to one of three groups: (1) usual care treatment, (2) a modular approach, or (3) a standard manualized treatment approach (the latter two will be described in greater detail in the next section).

Weisz et al. (2012) found that youth in the modular treatment group showed significantly faster improvement in symptom severity and lower rates of posttreatment diagnoses than youth in the standard manualized and usual care treatment groups. In contrast, there were no significant differences in outcomes between youth in the standard manualized treatment group and youth in

the usual care treatment group. Further, Weisz et al. (2012) found group differences in treatment adherence. Independent observers coded treatment sessions using adherence instruments that included all model-specific elements found in the SMT groups (*Coping Cat, PASCET,* and *Defiant Children*), and items for each of the 31 *MATCH* modules. Usual care recordings were randomly coded with either the SMT or MMT adherence measure to identify the presence/absence of treatment elements in the UC group, with low instance of standard or modular elements demonstrating treatment differentiation. The modular treatment sessions contained more evidence-based content than usual care treatment sessions (83% vs 8%), and more non-evidence-based content than standard manualized treatment sessions (17% vs 7%). In the standard manualized treatment group, 93% of session content was consistent with the standard treatment manuals, and in the modular manualized treatment group, 83% of the content was consistent with modular treatment protocol. In the usual care treatment group, only 8% of content was found in any of the standard (*Coping Cat, PASCET, Defiant Children*) or modular (*MATCH*) treatment manuals.

Participants

Youth participants. Youth were included in the parent study if they met the following criteria: (1) met diagnostic criteria for a Diagnostic and Statistical Manual of Mental Disorders, 4th ed. (DSM-IV-TR; American Psychiatric Association, 2000) disorder in the areas of anxiety, depression, or conduct problems, as determined by the Children's Interview for Psychiatric Syndromes (ChIPS; Weller, Weller, Rooney, & Fristad, 1999, Weller, Weller, Fristad, Rooney, & Schecter, 2000), or (2) displayed clinically elevated problem levels ($T > 65$) in one or more of the three problem areas based on relevant scales of the Child Behavior Checklist (CBCL) and the Youth Self Report (YSR; Achenbach & Rescorla, 2001); and (3) family sought treatment (as

opposed to recruitment or advertising). The primary clinical problem in each case was determined by considering diagnoses, scale scores from the CBCL and YSR, and youth- and parent-identified top problems on the Top Problems Assessment (Weisz et al., 2011). Youth were excluded from the parent study if: (1) they presented evidence of intellectual disability, pervasive developmental disorder, psychotic symptoms, or bipolar disorder; or (2) their primary problem was inattention or hyperactivity.

The parent study consisted of 174 youths aged 7 to 13 (M = 10.59 years, SD = 1.76), of whom 121 (70.0%) were male; 45.0% were Caucasian, 32.0% were multiracial, 9.0% were African American, 6.0% were Latino/a, 4.0% were Asian American or Pacific Islander, 2.0% were classified as "other," and 2.0% chose not to identify their ethnicity. The current study included only youth participants whose principal problem area was determined to be anxiety (31.6% of the parent study) and assigned to one of the two manualized treatment groups (i.e. not usual care). Additionally, youth participants were included in the current study only if they had a minimum of two audible treatment sessions and received treatment from a single therapist. The present study included 38 youths aged 8 to 13 (M = 9.84 years, SD = 1.65), of whom 52.6% were male; 60.5% were Caucasian, 26.3% were multiracial, 5.3% were African American, 2.6% were Latino/a, 2.6% were Asian American or Pacific Islander, and 2.6% were classified as "other."

The SMT group consisted of 22 youth (M age = 9.77 years, SD = 1.51; 50.0% male; 72.5% Caucasian, 18.2% multiracial, 4.5% Asian American, 4.5% Latino/a). CBCL T-scores ranged from 54 to 84 (M = 70, SD = 6.72) for the internalizing subscale, and from 33 to 77 (M = 59, SD = 11.28) for the externalizing subscale. The MMT group consisted of 16 youth (M age = 9.94 years, SD = 1.88; 56.3% male; 43.8% Caucasian, 37.5% multiracial, 12.5% African American, and 6.3% other). CBCL T-scores ranged from 48 to 84 (M = 69.56, SD = 9.33) for the

internalizing subscale, and from 33 to 73 (*M* = 55.06, *SD* = 11.64) for the externalizing subscale. See Table 1 for demographic and clinical characteristics of the youth participants.

Therapist participants. Therapist characteristics were collected using the Therapist Background Questionnaire (TBQ), a therapist self-report form that captures information about therapist demographic and training characteristics including age, ethnicity/race, sex, educational and training background (e.g., specialty degree, years of training, years practicing) and primary theoretical orientation. The parent study consisted of 84 therapists (*M* age = 40.60 years, *SD* not reported; *M* years full-time clinical experience (post-training) = 7.60, *SD* not reported) from 10 outpatient clinical service organizations in Massachusetts and Hawaii. Therapists were 80% female; 56% were Caucasian, 23% were Asian American, 6% were African American, and 6% were Pacific Islander; ethnicity was not reported for 9% of the sample.

The present study included 26 therapists who delivered individual SMT or MMT, and treated youth with a principal problem area of anxiety. Therapists were 80.8% female; 57.6% were Caucasian, 23% were Asian American, 7.6% were African American, and 3.8% were Pacific Islander; ethnicity was not reported for 8.0% of the sample. The mean age of the therapists was 40.35 years (*SD* = 9.67), and the mean number of years of clinical experience was 6.50 (*SD* = 6.82). Therapists included social workers (42.3%), behavior specialists/behavior health specialists (23.1%), psychologists (15.4%), mental health counselors (15.4%); professional specialty was not reported for 3.8% of the sample.

The SMT group consisted of 16 therapists (*M* age = 43.56, *SD* = 9.96*; M* years of experience = 7.17, *SD* = 7.75; 81.3% female; 56.3% Caucasian, 12.5% African American, 12.5% Asian American, 6.3% Pacific Islander, and 12.5% not reported). The MMT group consisted of 10 therapists (*M* age = 35.2, *SD* = 6.81; *M* years of experience = 5.25, *SD* = 4.83*;*

80% female; 60% Caucasian, and 40% Asian American. See Table 1 for demographic and training characteristics of the therapist participants.

Table 1

Youth and Therapist Descriptive Data and Group Comparisons

Variable	M *(SD)* or %		*t* or *χ2 (p-value)*
Youth	SMT (*N* = 22)	MMT (*N* = 16)	
Age	9.77 *(1.51)*	9.94 *(1.88)*	0.30 *(.766)*
Sex			0.15 *(.703)*
Female	50	43.7	
Race/Ethnicity			8.18 *(.147)*
Caucasian	72.5	43.8	
African American	-	12.5	
Asian American	4.5	-	
Latino/a	4.5	-	
Multiracial	18.2	37.5	
Other	-	6.3	
CBCLT-scores			
Total (pre)	65.27 *(7.49)*	63.63 *(10.39)*	-0.57 *(.573)*
Internalizing (pre)	70 *(6.72)*	69.56 *(9.33)*	-0.17 *(.868)*
Externalizing (pre)	59 *(11.28)*	55.06 *(11.64)*	-1.05 *(.301)*
Anxiety (pre)	69.64 *(7.20)*	69.88 *(7.76)*	0.10 *(.923)*
Therapist	SMT (*N* = 16)	MMT (*N* = 10)	
Age	43.56 *(9.96)*	35.2 *(6.81)*	-2.33 *(.029)**
Sex			0.01 *(.937)*
Female	81.3	80.0	
Years of experience	7.17 *(7.75)*	5.25 (*4.83)*	-0.63 *(.534)*
Race/Ethnicity			5.25 *(.272)*
Caucasian	56.3	60.0	
African American	12.5	-	
Asian American	12.5	40.0	
Pacific Islander	6.3	-	
Not reported	12.5	-	
Specialty			2.92 *(.712)*
Social worker	43.8	40.0	
Behavior specialist	18.7	30.0	
Psychologist	12.5	20.0	
Mental health counselor	18.7	10.0	
Not reported	6.3	-	

Note. SMT = Standard Manualized Treatment (*Coping Cat*), MMT = Modular Manualized Treatment (*MATCH*) ICBT in Weisz et al. (2012) study. CBCL = Child Behavior Checklist.
*$p < .05$

Treatment Groups

In the parent study, therapists were randomized into one of three treatment groups using a cluster randomization design (Campbell, Elbourne, & Altman, 2004) wherein a blocked randomization stratified by therapist educational level (master's vs doctoral degree) was utilized; the allocation ratio for each block was 1:1:1. Youths and caregivers knew they would be randomized, but were blind to treatment group.

Standard manualized treatment (SMT). The standard manualized treatment (SMT) included three treatment protocols with manualized instructions and a prescribed order of treatment sessions. These protocols included: (1) *Coping Cat*, an individual cognitive behavioral therapy protocol for anxiety (Kendall, 1994; Kendall & Hedtke, 2006a, 2006b) (2) *Primary and Secondary Control Enhancement Training (PASCET)*, an ICBT protocol for depression (Weisz, Weersing, Valeri, & McCarty, 1999); and (3) *Defiant Children*, a Behavioral Parent Training protocol for conduct problems (Barkley, 1997). Therapists used diagnostic information, CBCL/YSR scores, and the Top Problems Assessment (Weisz et al., 2011) to determine whether treatment should begin with a focus on anxiety, depression, or conduct. If anxiety was determined to be the principal problem area, therapists first delivered *Coping Cat*; if the principal problem was depression, therapists first delivered *PASCET*; and if the principal problem was conduct, therapists first delivered *Defiant Children*. Because the current study was limited to youth with a principal problem area of anxiety, it included only therapists who first delivered *Coping Cat.*

Coping Cat. *Coping Cat* consists of 16-20 sessions designed to address anxiety through skill-building (e.g. cognitive restructuring, relaxation, problem solving), graduated exposure to

feared objects or situations, and continued practice of skills both in (e.g. role plays) and out (i.e. homework assignments) of session.

PASCET. *PASCET* consists of 10-15 sessions of ICBT that targets depressive symptoms through the use of cognitive (e.g. cognitive restructuring) and behavioral (e.g. behavioral activation) skills. The skills are learned through in-session practice and out-of-session homework.

Defiant Children. *Defiant Children* is a 10-step behavioral parent training protocol that teaches parenting skills (e.g. praise, one-on-one time) shown to be efficacious in reducing disruptive and noncompliant behaviors in children. Parents participate in in-session role plays with the therapist and child, as well as at-home practice with the child.

Modular manualized treatment (MMT). Therapists in the modular manualized treatment (MMT) group used the *Modular Approach to Therapy for Children with Anxiety, Depression, and Conduct Problems* (*MATCH*; Chorpita & Weisz, 2005) protocol. *MATCH* consists of modules designed to address three problem areas: anxiety, depression, and conduct problems, and correspond to treatment procedures delivered in the *Coping Cat*, *PASCET*, and *Defiant Children* protocols. The *MATCH* protocol includes flowcharts for each problem area that specify a default sequence of modules. In the MMT group, therapists focused on the initial problem area based on the CHIPs, CBCL/YSR, and Top Problems Assessment, and select its designated flowchart. The flowcharts suggest a default module sequencing, however, if the default sequence was impeded by a crisis, stressor, or comorbid condition (e.g. depressive symptoms are interfering with anxiety treatment) the therapist was able to alter the sequence of modules within a flow chart, incorporate modules from another flowchart, or change to another flowchart altogether. For example, the flowchart for anxiety includes “Getting Acquainted,”

"Fear Ladder," "Learning About Anxiety-Child," and "Learning about Anxiety-Parent." The flowchart then diverges depending on if the provider determines that the child is "ready to practice." If the therapist determines that the child is "ready to practice," they proceed to graduated exposures. In contrast, if the therapist determines that the child is not "ready to practice," the therapist may follow the "interference" pathway of the flowchart wherein modules related to conduct problems, trauma-related problems, and mood-related problems are available for use. While the current study will focus only on the treatment of youth with anxiety, therapists had access to all modules.

Therapist training and consultation. Therapists randomized into either SMT or MMT groups received six days of training together, with two days designated for each of the three problem areas. Both SMT and MMT assigned therapists received weekly consultation on study cases from project consultants. To maintain treatment adherence, the parent study utilized a feedback system that allowed consultants to track the delivery of treatment practices. This feedback system involved a semi-structured interview in each consultation meeting wherein consultants used a checklist of treatment practices to gather information from therapists about what content was delivered (e.g. relaxation), and the techniques employed in delivering the content (e.g. role play). This feedback system was utilized in both the SMT and MMT groups. Based on the gathered information, consultants then provided guidance and support to therapists. Study consultants also led conversations about measurement feedback and client progress.

Treatment Adherence Instrument

Cognitive Behavioral Therapy Adherence Scale for Youth Anxiety (CBAY-A; Southam-Gerow et al., 2016). The current study evaluated the performance of the CBAY-A for use across treatment groups. The CBAY-A is a 33-item observational instrument intended to

capture adherence to common practice elements found in ICBT for youth anxiety. The CBAY-A consists of 6 standard (i.e. prescribed interventions standard to many CBT programs such as homework assignment), 22 model (i.e. interventions specific to ICBT for youth anxiety, such as exposure), and 7 delivery (i.e. how specific model items were delivered, such as modeling or rehearsal) items (for a detailed description of CBAY-A item generation, see Southam-Gerow et al., 2016). Scores on a previous version of the CBAY-A (consisting of 5 standard items, e.g. Homework Review; 12 model items, e.g., Exposure; and 6 delivery items, e.g., Rehearsal), have displayed evidence of item- and subscale-level score reliability (ICCs ranged from .48 to .80; *M* = .77, *SD* = .15) and construct validity for use with an ICBT protocol (*Coping Cat*; Kendall & Hedtke, 2006a, 2006b; Southam-Gerow et al., 2016). For a full list of items, see Appendix A. Coders produced extensiveness ratings for each item on a 7-point Likert-type scale with the following anchors: 1 = *not at all*, 3 = *somewhat*, 5 = *considerably*, and 7 = *extensively*. Extensiveness consists of both the frequency (the number of intervals in which an item is observed) and thoroughness (the depth and comprehensiveness) with which a therapist delivers an intervention. Coders considered both frequency and thoroughness equally when assigning an extensiveness rating. Because the current study focused on common practice elements of ICBT for youth anxiety, only model items will be used in the analyses.

Instruments Used for Validity Analyses

MATCH and Standard Consultation Records (Ward et al., 2013). The MATCH and Standard Consultation Records are two separate consultant-rated instruments designed to capture adherence to a standard ICBT protocol (*Coping Cat*) and the *MATCH* protocol used in the Child STEPs trial (Ward et al., 2013). The instruments consist of matrix checkboxes, with rows representing specific treatment practices and columns representing activities or procedures

related to each session. The records are specific to either the *Coping Cat* protocol or the *MATCH* protocol. For example, the rows on the Standard Consultation Record consist of *Coping Cat* elements such as "FEAR Plan" while the rows on the MATCH Consultation Record consist of *MATCH* modules such as "Cognitive STOP." Some of the rows are consistent across the two instruments because they are elements found in both protocols, such as "Psychoeducation," and "Practice."

Because there is not an exact one-to-one match between the elements of the Standard and Modular treatments, there are 47 rows in the Standard Consultation Record and 35 rows in the MATCH Consultation Record. The columns were consistent across the two instruments and included items such as "Rehearsal," "Homework Assigned," "Receipt" (a client's comprehension of the session content), and "Coverage" (full or partial coverage of the session content). The consultation records were filled out by project consultants in collaboration with therapists for each session discussed during weekly consultation meetings.

The consultant-rated records have shown initial evidence of convergent validity when compared to consultation records scored by independent observers (Ward et al., 2013). Additionally, scores on both the Standard Consultation Record and MATCH Consultation Record have demonstrated good inter-rater reliability (ICCs ranged from .50 to 1.0; $M = .80$, *SD* not reported) when rated by independent observers (Ward et al., 2013).

Cognitive-Behavioral Treatment for Anxiety in Youth Competence Scale-Revised (CBAY-C; McLeod et al., 2018). The CBAY-C is an observational instrument designed to capture therapist competence (i.e. skill and responsiveness) in the delivery of core practice elements found in ICBT for youth anxiety. The CBAY-C was designed to parallel the content found in the CBAY-A and consists of 4 standard items (e.g., Homework Assignment), 22 model

items (e.g., Emotion Education), and 7 delivery items (e.g., Modeling). Scores on a previous version of the CBAY-C (consisting of 5 standard items, 12 model items, 6 delivery items, and 2 global items) have shown initial evidence of item-level inter-rater reliability (ICCs ranged from .37 to .80; *M* = .67, *SD* = .11) and construct validity (McLeod et al., 2018). For the present study, two subscale scores were created; one for the SMT group and one for the MMT group. Procedures for creating these scores followed those used to create subscales for the CBAY-A described below.

Therapy Process Observational Coding System-Alliance Scale (TPOCS-A; McLeod & Weisz, 2005). The TPOCS-A is an observational measure of youth-therapist alliance. The TPOCS-A consists of 9 items designed to objectively capture two commonly emphasized dimensions of alliance (Shirk & Russell, 1998): bond (i.e., affective aspects of the youth-therapist relationship) and task (i.e., client participation in the activities of treatment). Items are rated on a 6-point Likert-type scale ranging from 0 = *not at all* to 5 = *a great deal*. Scores on the TPOCS-A have displayed evidence of inter-rater reliability (ICCs ranged from .40 to .75; *M* = .59, *SD* = .10), internal consistency (α = .95), convergent validity with a self-report alliance instrument (TASC; Shirk & Saiz, 1992), and predictive validity with child outcomes (Liber et al., 2010; McLeod & Weisz, 2005). The TPOCS-A scale score was created by averaging the nine items on the scale. In the present study, internal consistency of the TPOCS-A was α = .89 in the SMT group and α = .85 in the MMT group. Scale-level inter-rater reliability was *excellent* in both the SMT group (ICC(2,2) = .86), and the MMT group (ICC(2,2) = .82; Cicchetti, 1994).

Symptom and Diagnostic Instruments

Youth clinical characteristics. Youth eligibility for the parent study (Weisz et al., 2012) and primary clinical problem area were determined using the Children's Interview for

Psychiatric Syndromes (ChIPs; Weller et al., 1999, Weller et al., 2000), the Child Behavior Checklist (CBCL) and Youth Self-Report (YSR; Achenbach & Rescorla, 2001), and the Top Problems Assessment (Weisz et al., 2011).

Children's Interview for Psychiatric Syndromes—Child and Parent Report. The parent study (Weisz et al., 2012) determined diagnoses using the Children's Interview for Psychiatric Syndromes (ChIPS), a structured interview designed to assess DSM-IV diagnoses (Weller, et al., 1999; Weller et al., 2000). A blinded interviewer administered the ChIPS to both the parent and the child, and a diagnosis of anxiety was given if generated by both parent and child interviews, or if generated only by the child interview (Silverman & Nelles, 1988). The ChIPS has demonstrated evidence of score reliability and validity across several studies and samples (Fristad et al., 1998; Weller et al., 1999; Weller et al., 2000).

Child Behavior Checklist and Youth Self-Report. The parent study used the Child Behavior Checklist (CBCL) and Youth Self-Report (YSR) to identify elevated problem levels in anxiety, depression, and/or conduct problems (Weisz et al., 2012). The CBCL and YSR are widely used caregiver and youth self-report measures of emotional and behavioral symptoms. A problem area was considered to be elevated if the reporter achieved a T-score equal to or greater than 65. Reliability and validity of the CBCL and YSR are well documented (Achenbach, Dumenci, & Rescorla, 2003; Achenbach & Rescorla, 2001).

Top Problems Assessment. The Top Problems Assessment (Weisz et al., 2011) obtains ratings from participants about the top three problems they identify as most important to them. The top three problems were identified separately by youth and their caregivers in structured interviews. The Top Problems Assessment has evidence of score reliability, validity, and sensitivity to change (Weisz et al., 2011).

Observational Coding Procedures

Coders. Three female clinical psychology doctoral students, one identifying as Asian-American and two identifying as Caucasian, comprised the CBAY-A and TPOCS-A coding team. Two female clinical psychology doctoral students, one identifying as Mexican-American and one identifying as Caucasian, comprised the CBAY-C coding team. One coder served on both coding teams.

Coding procedures. One coding team was created for the CBAY-A and TPOCS-A, and another was created for the CBAY-C. Coding teams were trained separately in a group format by the PIs over the course of approximately three months to reach adequate reliability at the item level (ICC(2,2) > .60; Cicchetti, 1994). Coders were naïve to treatment group and coded sessions in a randomly assigned order.

Initially, coders received didactic instruction in the form of reading and discussing coding manual as well as reviewing coded sessions with PIs. Coders then proceeded to code recordings independently and participated in weekly meetings in which results of the practice coding were discussed. Next, coders entered the certification phase wherein they were required to reach an adequate level of reliability (ICC > .60; Cicchetti, 1994) across 32 recordings.

Upon completion of the certification phase, coders began to independently code randomly assigned sessions. Coders met regularly with the PIs throughout the independent coding phase to prevent coder drift (Margolin et al., 1998), which was assessed through the ongoing examination between-coder reliability (i.e. reliability coefficients). If an item fell below an acceptable level of reliability (ICC < .60, Cicchetti, 1994), additional training was provided, consisting of a reexamination of the coding manual, group discussions about the discrepancies, and group coding of challenging items.

Sampling of treatment sessions. All available recordings, except for the first and last sessions for each client, were selected from each case for coding and randomly assigned to coders. The first and last sessions were excluded because as they were more likely to contain intake or termination content. The final sample used in the present study consisted of a total of 602 recordings (359 SMT recordings and 243 MMT recordings). While all 602 recordings were coded using the TPOCS-A, the child was not present in 50 sessions, resulting in a final TPOCS-A sample of 552 recordings. Consultation Record data were not available for a total of 22 sessions, resulting in a final sample of 580 (350 SMT and 230 MMT) recordings.

Data Analysis Plan

The purpose of the current study was to examine whether the CBAY-A could be used across two types of ICBT protocols for youth anxiety, so two separate CBAY-A subscales were created for each treatment protocol: a *Coping Cat* subscale (CBAY-A SMT) used for the *Coping Cat* group and a *MATCH* subscale (CBAY-A MMT) used for the *MATCH* group. Score reliability, convergent validity, discriminant validity, and discriminative validity were examined separately for each subscale and thus, each group. Subscale creation is described in detail below.

Preliminary Analyses

Patterns of missing data were assessed following procedures outlined by Schafer and Graham (2002) using Little's Missing Completely at Random (MCAR) test. Sample bias analyses were conducted to determine if the samples in the SMT and MMT groups differed, and if the sample used in this study differed significantly from the parent study (Weisz et al., 2012) sample.

Inter-rater reliability and data distribution. Normality was evaluated for the items of each instrument used in the primary analyses (CBAY-A, CBAY-C, TPOCS-A, Standard

Consultation Record, and MATCH Consultation Record) by examining skewness and kurtosis. Normality analyses were conducted separately for the *Coping Cat* group and the *MATCH* group. Inter-rater reliability for the CBAY-A and CBAY-C item scores and TPOCS-A scale score was calculated using intra-class correlation coefficients (ICCs; Shrout & Fleiss, 1979). The model ICC(2,2) based on a two-way random effects model was used because it provides a reliability estimate of the average score of all coders and allows for generalizability of the findings to other samples (Shrout & Fleiss, 1979). Based on recommendations by Cicchetti (1994), ICCs below .40 were considered *poor*, between .40 and .59 were considered *fair*, between .60 and .74 were considered *good*, and .75 and above were considered *excellent*. ICCs were calculated separately for the *Coping Cat* and *MATCH* groups.

Subscale generation. To evaluate the score validity of the CBAY-A in two separate ICBT protocols for youth anxiety, two subscales were created to represent adherence to each protocol (see Figure 1). Specifically, one subscale (i.e., the *Coping Cat* subscale, hereafter referred to as the CBAY-A SMT subscale) included items that represent adherence to *Coping Cat*, and the other subscale (i.e., the *MATCH* subscale, hereafter referred to as the CBAY-A MMT subscale) included items that represent adherence to *MATCH*. Subscales were generated based on (1) an item's inclusion in a specific protocol's treatment manual; (2) whether an item met acceptable levels of inter-rater reliability (ICC > .60; Cicchetti, 1994) and used the full range of scores; and (3) expert consultation. Corresponding subscales were created for the CBAY-C (CBAY-C SMT, CBAY-C MMT). The CBAY-A MMT subscale was used for the primary analyses involving the *MATCH* treatment group, and the CBAY-A SMT subscale was used for the primary analyses involving the *Coping Cat* treatment group. Normality was evaluated for each of the scales and subscales (CBAY-A SMT, CBAY-A MMT, CBAY-C SMT, CBAY-C

MMT, TPOCS-A, Standard Consultation Record, MATCH Consultation Record) by an examination of skewness and kurtosis. Normality analyses were conducted separately for the *Coping Cat* group and the *MATCH* group.

Standard Manualized Treatment (SMT) subscale	Modular Manualized Treatment (MMT) subscale
1. Psychoeducation 2. Emotion Education 3. Fear Ladder 4. Relaxation 5. Cognitive 6. Problem Solving 7. Self-Reward 8. Coping Plan 9. Exposure: Prep 10. Exposure 11. Exposure: Debrief	1. Psychoeducation 2. Emotion Education 3. Fear Ladder 4. Cognitive 5. Coping Plan 6. Exposure: Prep 7. Exposure 8. Exposure: Debrief 9. Maintenance

Figure 1. CBAY-A subscales for Standard Manualized Treatment (SMT; *Coping Cat*) and Modular Manualized Treatment (MMT; *MATCH*)

Construct validity: Convergent and discriminant validity

The CBAY-A SMT subscale scores were compared to scores on instruments of adherence (Standard Consultation Record) for evidence of convergent validity. The CBAY-A MMT subscale scores were compared to scores on instruments of adherence (MATCH Consultation Record) for evidence of convergent validity using Pearson product-moment correlations. Pearson product-moment correlations were also calculated among scores on the CBAY-A SMT subscale, an instrument of adherence (Standard Consultation Record), an instrument of competence (CBAY-C SMT), and an instrument of the alliance (TPOCS-A) to examine discriminant validity in the SMT group. Similarly, Pearson product-moment correlations were calculated among scores on the CBAY-A MMT subscale, an instrument of adherence (MATCH Consultation Record), an instrument of competence (CBAY-C MMT), and

an instrument of the alliance (TPOCS-A) to examine discriminant validity in the MMT group. Correlation magnitudes were interpreted following Rosenthal and Rosnow's (1984) guidelines: correlations are small if $.10 \leq r < .24$, moderate if $.24 \leq r < .36$, and high if $r \geq .36$. Follow-up contrasts were calculated using Fisher *r*-to-*z* transformation.

Discriminative validity

We evaluated the discriminative validity of the CBAY-A SMT and CBAY-A MMT subscale scores. To test the discriminative validity of the CBAY-A SMT and CBAY-A MMT subscale scores, group mean differences between the SMT and MMT groups were compared in two ways. First, adjusted least square means (LSMs) were computed to account for the nested design of the data (Barber, Foltz, Crits-Christoph, & Chittams, 2004) and the overall *F* test for study group was examined. The LSMs account for the influence of study group, therapist, client, and coder. Group differences were also explored by conducting an independent-samples t-test to compare the percent of practice elements that were delivered at some point during the course of treatment (i.e., out of the 11 practice elements on the CBAY-A SMT subscale, how many were delivered to a client at least once over the coded treatment sessions?). Weisz et al. (2012) compared group levels of adherence by comparing the percent of treatment delivery that was consistent with the assigned treatment protocol (e.g., what percent of treatment delivered can be found in the *Coping Cat* protocol?). Because our data did not allow us to replicate Weisz et al.'s (2012) analyses, we chose to use two separate methods in an attempt to enhance our ability to make interpretations about discriminative validity.

Results

Preliminary Analyses

Sample bias. Six youth participants from the parent study were not included in the current study because they did not have more than two recorded sessions. The six excluded youth did not differ from the included sample in age ($t[41] = 1.45$, $p = .168$), sex ($\chi^2 [1, N = 44] = 1.99$, $p = .158$), race/ethnicity ($\chi^2 [5, N = 42] = 5.73$, $p = .333$), CBCL total scores ($t[42] = -1.13$, $p = .276$), CBCL internalizing subscale scores ($t[42] = -1.14$, $p = .261$), CBCL externalizing subscale scores ($t[42] = -1.30$, $p = .201$), or CBCL anxiety scores ($t[42] = -1.78$, $p = .082$).

Client-level demographics were compared between the SMT group and the MMT group (see Table 1). The samples did not significantly differ in age ($t[36] = 0.30$, $p = .766$), sex ($\chi^2 [1, N = 38] = 0.15$, $p = .703$), race/ethnicity ($\chi^2 [5, N = 38] = 8.18$, $p = .147$), CBCL total scores ($t[36] = -0.57$, $p = .573$), CBCL Internalizing scale scores ($t[36] = -0.17$, $p = .868$), CBCL Externalizing Subscale scores ($t[36] = -1.05$, $p = .301$), or Anxiety subscale scores ($t[36] = 0.10$, $p = .923$).

All therapists who delivered either SMT or MMT to youth with a principal problem area of anxiety in the parent study were included in the current sample. A comparison of therapist-level demographics between the SMT group and MMT group (see Table 1) revealed that the mean age of therapists was significantly higher in the SMT group ($M = 43.56$ years; $SD = 9.96$) than the MMT group ($M = 35.20$ years; $SD = 6.81$; $t[24] = -2.33$, $p = .029$), but the groups did not significantly differ in years of experience ($t[21] = -0.63$, $p = .534$). The SMT and MMT groups did not significantly differ in regards to sex ($\chi^2 [1, N = 26] = 0.01$, $p = .937$), race/ethnicity ($\chi^2 [4, N = 26] = 5.26$, $p = .272$), or specialty ($\chi^2 [5, N = 26] = 2.92$, $p = .712$).

Missing data. The CBAY-A was used to code a sample of treatment sessions from the parent study. To ensure that the coded *Coping Cat* (SMT) sessions and the coded *MATCH* (MMT) sessions were equally representative of overall treatment, the percent of sessions coded

was compared across groups using an independent samples t-test. There was not a significant difference in percent of sessions coded between the SMT group (M = .72, SD = .18), and the MMT group (M = .73, SD = .14; $t(36)$ = .155; p = .880).

Patterns of missing data for the consultation records were assessed using Little's MCAR test. A data point was considered missing if an item (e.g., *Relaxation*) was noted as occurring within a session (e.g. a rating was given for level of understanding, indicating that the item was delivered; see Appendices B and C for a full list of possible ratings), but was not rated as either "covered-part" or "covered-full." For both the MMT group (Little's MCAR test Chi Square = 4.11, df = 10, p = .942) and the SMT group (Little's MCAR test Chi Square = 4.65, df = 7, p = .703), data were considered to be MCAR. In the SMT group, 0.03% of data were missing, and in the MMT group, 0.4% of data were missing.

A total of 2.3% of therapist-level demographic information (i.e. race/ethnicity information for two therapists, specialty area for one therapist) were missing, and was considered to be MCAR (Little's MCAR test Chi Square = 17.92, df = 11, p = .083). No youth-level demographic information was missing.

Inter-rater reliability and data distribution. CBAY-A means, standard deviations, ranges, and normality were explored at both the item and subscale level. All items had a range of at least 5 in the SMT group, and at least 3.5 in the MMT group. In both groups, all items were positively skewed. SMT item skewness ranged from 1.49 to 4.90, and MMT item skewness ranged from 1.19 to 4.67. Kurtosis of items ranged from 1.46 to 26.41 in the SMT group, and from -0.05 to 24.23 in the MMT group. Skewness and kurtosis for multiple items fell outside the range of a normal distribution (i.e., -2 to 2; George & Mallery, 2016), though these patterns are consistent with previous evaluations of the CBAY-A items (Southam-Gerow et al., 2016). Scores

were considered to be outliers if they had a *z*-score > 3.00. In the SMT group there was an average of 12 outliers per CBAY-A SMT item, and in the MMT group there was an average of 11 outliers per CBAY-A MMT item. Outliers were distributed across sessions and were determined to be scores at the high end of the range rather than data error, so they were not removed. Inter-rater reliability (ICC(2,2)) was evaluated for each item in the CBAY-A SMT subscale in the SMT group and for each item in the MMT subscale in the MMT group. In the SMT group, ICCs for CBAY-A items ranged from .66 to .94 (M = .83, SD = .07), and in the MMT group, ICCs for CBAY-A items ranged from .60 to .91 (M = .80, SD = .09). Table 2 displays descriptive statistics and ICCs for CBAY-A items.

Normality was also explored for the Standard and MATCH Consultation Records. Standard Consultation Record item skewness ranged from 0.68 to 4.53, and kurtosis ranged from -1.47 to 19.04. MATCH Consultation Record item skewness ranged from 0.27 to 5.69 and kurtosis ranged from -1.76 to 31.91. Skewness and kurtosis for multiple items fell outside the range of a normal distribution (i.e., -2 to 2; George & Mallery, 2016), although these patterns are expected given that items that were not present in a session were scored a 0, and items are not expected to occur in all, or even multiple, sessions.

Subscale generation. Subscale scores were generated for the CBAY-A, CBAY-C, Standard Consultation Record, and MATCH Consultation Record.

CBAY-A. CBAY-A SMT and CBAY-A MMT subscales were created based on an item's inclusion in a specific protocol's treatment manual. The content of the SMT and MMT subscales (see Figure 1) were presented to, and approved by, expert consultants. Thus, the CBAY-A SMT subscale included all CBAY-A items found in the *Coping Cat* treatment protocol, and the

CBAY-A MMT subscale included all CBAY-A items found in the *MATCH* treatment protocol for anxiety.

Table 2 displays descriptive statistics for CBAY-A SMT and CBAY-A MMT subscales for the SMT and MMT groups. For each subscale, all items showed at least *good* inter-rater reliability (ICC > .60; Cicchetti, 1994) in the SMT and MMT groups. Only two CBAY-A items in the SMT subscale did not display the full range of scores in the SMT group (i.e., maximum range was 6; *Exposure Prep* and *Exposure Debrief* had a range of 5). In the MMT group, six CBAY-A items in the MMT subscale had a range of at least 5, two items (i.e., *Emotion Education*, *Cognitive*) had a range of 4.5, and one item (i.e., *Coping Plan*) had a range of 3.5. Overall, CBAY-A SMT and MMT subscale items were coded reliably by trained coders and were able to capture a range of adherence-related therapist behaviors, and thus no items were removed from either subscale.

CBAY-A SMT and CBAY-A MMT subscale scores were created by first averaging items across coders, and then averaging all items in each subscale. The CBAY-A SMT and CBAY-A MMT subscale scores had low skewness (SMT = 0.32, MMT = 0.55) and kurtosis (SMT = -0.33, MMT = -0.21) in the SMT and MMT groups, respectively, which fall within the range of a normal distribution (George & Mallery, 2016). Inter-rater reliability was ICC(2,2) = .79 for the CBAY-A SMT subscale and ICC(2,2) = .85 for the CBAY-A MMT subscale, indicating that both subscales can be coded reliably by independent coders.

Table 2

CBAY-A Item and Subscale Descriptive Data and Inter-rater Reliability

Item	*N*	Range	*M*	*SD*	ICC (2,2)	Skewness	Kurtosis
SMT							
Psychoeducation	219	6	1.96	1.16	.663	1.49	2.07
Emotion Education	161	6	1.93	1.50	.858	1.74	1.90
Fear Ladder	121	6	1.66	1.18	.827	1.92	3.00
Relaxation	102	6	1.57	1.24	.898	2.60	6.15
Cognitive	149	6	1.82	1.40	.871	1.96	3.00
Problem Solving	42	6	1.20	0.76	.771	4.90	26.41
Self-Reward	37	6	1.23	0.90	.940	4.71	22.09
Coping Plan	157	6	1.90	1.39	.849	1.68	2.06
Exposure: Prep	107	5	1.62	1.15	.843	1.88	2.60
Exposure	91	6	1.72	1.36	.894	1.68	1.57
Exposure: Debrief	83	5	1.40	0.86	.764	2.42	6.17
Subscale	359	1.55	1.64	0.31	.791	0.32	-0.33
MMT							
Psychoeducation	154	6	2.17	1.44	.856	1.49	1.62
Emotion Education	34	4.5	1.18	0.61	.754	4.54	23.17
Fear Ladder	115	5.5	1.81	1.29	.807	2.06	3.87
Cognitive	49	4.5	1.30	0.80	.753	3.45	12.51
Coping Plan	22	3.5	1.13	0.48	.601	4.67	24.23
Exposure: Prep	103	5	1.74	1.10	.789	1.54	1.68
Exposure	88	5.5	1.99	1.51	.911	1.19	-0.05
Exposure: Debrief	78	5	1.59	1.06	.865	1.80	2.36
Maintenance	36	6	1.35	1.03	.883	3.43	11.81
Subscale	243	1.67	1.58	0.39	.847	0.55	-0.21

Note. N represents the number of times an item was rated as present at least once during a treatment session by either coder.

CBAY-C. Descriptive statistics for CBAY-C SMT and CBAY-C MMT subscale scores can be found in Table 3. The CBAY-C SMT subscale contained the same items as the CBAY-A SMT subscale. Inter-rater reliability of the CBAY-C SMT subscale items ranged from *fair* (ICC(2,2) = .54) to *excellent* (ICC(2,2) = .84) with a mean ICC(2,2) = .71, which is in the *good* range (Cicchetti, 1994). Because all item-level ICCs were at least *fair*, all items were retained. Item-level inter-rater reliability of the CBAY-C MMT subscale ranged from *poor* (ICC(2,2) = .34; *Coping Plan*) to *excellent* (ICC(2,2) = .89) with a mean ICC(2,2) = .59 in the *fair* range (Cicchetti, 1994). The MMT subscale of the CBAY-C also mirrored the MMT subscale of the CBAY-A with the exception of one item (*Coping Plan*) which was not included in the CBAY-C MMT subscale due to poor inter-rater reliability (ICC(2,2) = .34). CBAY-C SMT and CBAY-C MMT subscale scores were created by first averaging items across coders, and then averaging all items in each subscale. Item scores of 0 were considered missing, and only items given a score of at least 1 by both coders were averaged across coders and included in the subscale scores. Inter-rater reliability was excellent for both the CBAY-C SMT subscale (ICC(2,2) = .77) and the CBAY-C MMT subscale with the *Coping Plan* item removed (ICC(2,2) = .76; Cicchetti, 1994).

Table 3

CBAY-C Item and Subscale Descriptive Data and Inter-rater Reliability

Item	*N*	Range	*M*	*SD*	ICC(2,2)	Skewness	Kurtosis
SMT							
Psychoeducation	53	3.5	3.72	0.85	.671	0.46	-0.003
Emotion Education	63	5	4.31	1.03	.775	0.43	0.47
Fear Ladder	50	4	3.66	1.07	.758	0.39	-0.68
Relaxation	37	4	4.00	0.83	.541	0.79	1.21
Cognitive	33	4.5	4.41	0.99	.646	-0.14	-0.12
Problem Solving	11	4	4.50	1.30	.674	0	-1.23
Self-Reward	14	3.5	4.18	1.12	.841	0.77	-0.52
Coping Plan	90	4.5	4.03	0.93	.697	0.44	-0.17
Exposure: Prep	68	5	3.35	1.23	.797	0.58	-0.64
Exposure	59	4.5	3.39	1.01	.674	1.01	0.62
Exposure: Debrief	41	4.5	3.65	1.20	.729	0.42	-0.72
Subscale	351	4.5	3.71	0.94	.768	0.33	-0.49
MMT							
Psychoeducation	66	4	3.70	0.74	.467	.270	0.26
Emotion Education	12	2	3.50	0.64	.444	.313	-0.86
Fear Ladder	34	4	3.66	1.05	.832	.023	-0.84
Cognitive	3	1.5	3.50	0.87	.889	1.73	-
Coping Plan*	9	2	3.94	0.73	.336	-0.09	-1.48
Exposure: Prep	59	3.5	2.66	0.72	.616	0.90	0.61
Exposure	57	4	2.97	0.86	.412	1.67	0.62
Exposure: Debrief	41	3	3.05	0.84	.646	0.70	0.72
Maintenance	7	2	3.93	0.67	.684	0.35	1.59
Subscale	222	4.12	3.19	0.80	.763	0.47	-0.27

Note. N represents the number of times an item was given a rating of competence by both coders. Competence ratings were only given if an item was identified as being present during a treatment session.

*Not included in the final CBAY-C MMT subscale due to lower inter-rater reliability.

Standard and MATCH Consultation Records. To create consistency in scoring strategies across the CBAY-A SMT and MMT subscales and the Standard and MATCH Consultation Records (which originally consisted of presence/absence ratings; see Appendices B

and C), Consultation Record scores were recoded into a 3-point extensiveness scale. For each item on the consultation records the consultant rater could select "covered-part," "covered-full," or make no selection. Item scores were recoded such that if neither "covered-part" nor "covered-full" were selected, the item value was 0, if "covered-part" was selected, the item value was 1, and if "covered full" was selected, the item value was 2. Subscale scores for each session were created by averaging all model items. If an item was missing for a given session, the average of the remaining items comprised the subscale score. Both the Standard Consultation Record and the MATCH Consultation Record subscale scores had skewness (-.48 and .21 respectively) and kurtosis (.92 and .23 respectively) scores that fell within the range of a normal distribution (George & Mallery, 2016).

Construct validity: Convergent and discriminant validity

CBAY-A SMT subscale. As can be seen in Table 4, the magnitude of the correlations among the scores on the four instruments in the SMT group (CBAY-A SMT, Standard Consultation Record, CBAY-C SMT, and TPOCS-A) ranged from .017 to .352. The strongest correlation was observed between scores on the CBAY-A SMT subscale and scores on the Standard Consultation Record (r = .352). Scores on the CBAY-A SMT subscale correlated at a moderate level with scores on the CBAY-C SMT subscale (r = .284), and at a low level with scores on the TPOCS-A (r = .137). Scores on the Standard Consultation Record evidenced a correlation below Rosenthal and Rosnow's (1984) cutoff for a small correlation with scores on the CBAY-C SMT subscale (r = -.03) and scores on the TPOCS-A (r = .017). These findings suggest that the CBAY-A SMT subscale and the Standard Consultation Record displayed similar patterns of correlations with instruments of adherence, competence, and alliance. Specifically,

they were both most highly correlated with another instrument of adherence, second most highly correlated with an instrument of competence, and least correlated with an instrument of alliance.

Table 4

Correlations Between the CBAY-A SMT Subscale, Standard Consultation Record, CBAY-C SMT Subscale, and TPOCS-A Scale

	2.	3.	4.
1. CBAY-A SMT	.352 ** (N = 350)	.284** (N = 351)	.137* (N = 332)
2. Standard CR		-.030 (N = 342)	.017 (N = 324)
3. CBAY-C SMT			--
4. TPOCS-A			--

Note. CR = Consultation Record. Sample sizes varied based on available data. For correlations involving the CBAY-C, sample size was contingent upon the presence of an item in a given session. For correlations involving the TPOCS-A, sample size was contingent upon the presence of the child in the session. Not all sessions rated with the CBAY-A SMT had corresponding scores on the Standard Consultation Record. All available data across constructs was retained to maximize sample size for each correlation.
*$p < .05$
**$p < .01$

Follow-up contrasts demonstrated that correlations between scores on the CBAY-A SMT subscale were significantly stronger for the Standard Consultation Record (r = .352) than either the correlation with scores on the CBAY-C SMT subscale (r = .284; z = 1.00, p = .032) or the TPOCS-A (r = .137; z = 2.99, p = .003). Further, scores on the CBAY-A SMT subscale were significantly more highly correlated with scores on the CBAY-C SMT subscale (r = .284) than the TPOCS-A (r = .137; z = 3.54, $p < .001$). The Standard Consultation Record was more highly correlated with the CBAY-A SMT subscale (r = .352) than either the CBAY-C SMT subscale (r = -.03; z = 5.21, $p < .001$) or the TPOCS-A scale (r = .017; z = 2.53, p = .011), and there was no

significant difference between its correlations with the CBAY-C SMT subscale and the TPOCS-A ($z = 0.60$, $p = .274$). These contrasts indicated that the CBAY-A SMT subscale and the Standard Consultation Record evidenced a significantly higher correlation with an instrument of adherence than instruments of competence or alliance.

CBAY MMT subscale. Seen in Table 5, the magnitude of the correlations among the scores on the four instruments in the MMT group (CBAY-A MMT, Standard Consultation Record, CBAY-C MMT, and TPOCS-A) ranged from -.055 to .474. The highest correlation was observed between scores on the CBAY-A MMT subscale and scores on the MATCH Consultation Record ($r = .474$). Scores on the CBAY-A MMT subscale correlated at a low level with scores on the CBAY-C MMT subscale ($r = .154$) and at a moderate level with scores on the TPOCS-A ($r = .337$). Scores on the MATCH Consultation Record were correlated at a level below Rosenthal and Rosnow's (1984) cutoff for a small correlation with scores on the CBAY-C MMT subscale ($r = -.055$), but had a small and positive (though non-significant) correlation with the TPOCS-A ($r = .114$).[1] These findings indicate that the CBAY-A MMT subscale and the MATCH Consultation Record display similar patterns of correlations with instruments of adherence, competence, and alliance. They are both most highly correlated with an instrument of adherence, second most highly correlated with an instrument of alliance, and least correlated with an instrument of competence.

[1] Because the Coping Plan item was removed from the CBAY-C MMT subscale, correlations were reexamined with the Coping Plan item removed from the CBAY-A MMT subscale, and no significant differences were found (MATCH Consultation Record; $r = .469$, $z = 0.07$, $p = .944$; CBAY-C MMT subscale; $r = .151$, $z = 0.03$, $p = .976$; TPOCS-A ($r = .325$, $z = 0.14$, $p = .889$), so the Coping Plan item was retained in the CBAY-A MMT subscale.

Table 5

Correlations Between the CBAY-A MMT Subscale, MATCH Consultation Record, CBAY-C MMT Subscale, and TPOCS-A Scale

	2.	3.	4.
1. CBAY-A MMT	.474** (N = 229)	.154* (N = 221)	.337** (N = 219)
2. MATCH CR		-.055 (N = 210)	.114 (N = 206)
3. CBAY-C MMT			--
4. TPOCS-A			--

Note. CR = Consultation Record. Sample sizes varied based on available data. For correlations involving the CBAY-C, sample size was contingent upon the presence of an item in a given session. For correlations involving the TPOCS-A, sample size was contingent upon the presence of the child in the session. Not all sessions rated with the CBAY-A MMT had corresponding scores on the MATCH Consultation Record. All available data across constructs was retained to maximize sample size for each correlation.
*$p < .05$
**$p < .01$

Follow-up contrasts demonstrated that the correlation between scores on the CBAY-A MMT subscale and the MATCH Consultation Record (r = .474) were significantly stronger than the correlation between the CBAY-A MMT subscale and CBAY-C MMT subscale (r = .154; z = 3.79, $p < .001$), and did not significantly differ from the correlation between the CBAY-A MMT subscale and the TPOCS-A (r = .337; z = 1.73, p = .084). Further, the correlation between scores on the CBAY-A MMT subscale and the TPOCS-A (r = .337) was significantly higher than the correlation between the CBAY-A MMT subscale and the CBAY-C MMT subscale (r = .154; z = 2.04; p = .041). The MATCH Consultation Record was more highly correlated with the CBAY-A MMT subscale (r = .474) than either the CBAY-C MMT subscale (r = -.055; z = 5.93, $p < .001$) or the TPOCS-A (r = .114; z = 2.53, p = .011), and there was no significant difference between its correlations with the CBAY-C MMT subscale and the TPOCS-A (z = 1.72, p = .085). In sum, these contrasts indicate that both the CBAY-A MMT subscale and the MATCH

Consultation Record are more highly correlated with an instrument of adherence than an instrument of competence. The magnitude of the correlations for the CBAY-A MMT subscale and the MATCH Consultation Record displayed the same pattern (i.e., from highest to lowest: adherence, alliance, competence). However, contrasts revealed a different pattern. The correlation between the MATCH Consultation Record and an instrument of adherence was significantly higher than the correlation between the MATCH Consultation Record and an instrument of alliance, but the same pattern was not seen for the CBAY-A MMT subscale.

Taken together, findings indicate that the CBAY-A SMT subscale, CBAY-A MMT subscale, the Standard Consultation Record, and the MATCH Consultation Record have a greater degree of overlap with other instruments of adherence than with instruments of competence or alliance. However, the patterns of competence and alliance correlations differ across the four adherence instruments. Specifically, the CBAY-A SMT subscale and the Standard Consultation Record had a greater degree of overlap with an instrument of competence than an instrument of alliance, while the CBAY-A MMT subscale and the MATCH Consultation Record displayed the opposite pattern.

Discriminative validity

To examine group differences in levels of adherence, adjusted LSMs were computed to account for the nested design of the data, considering the influence of treatment group, therapist, and client (see Table 6). The mean CBAY-A SMT subscale score was 1.64 (SE = 0.02), and the mean CBAY-A MMT subscale score was 1.58 (SE = 0.02). An F test for treatment group failed to find a significant difference between mean CBAY-A SMT and MMT subscale scores; $t(574) = -1.74$, $p = .080$.

Table 6

Least Square Means of CBAY-A SMT Subscale and CBAY-C MMT Subscale Scores Across Groups

SMT *M (SE)*	MMT *M (SE)*	*t*	*p*
1.64 *(0.02)*	1.58 *(0.02)*	-1.74	.080

Note. SMT means were generated using the CBAY-A SMT subscale and MMT means were generated using the CBAY-A MMT subscale.

We also conducted independent-samples t-tests to compare groups on percent of practice elements in each subscale (11 in SMT, 9 in MMT) delivered over the course of treatment (see Table 7); no significant group differences were found (SMT $M = 0.84$, $SD = 0.23$; MMT $M = 0.85$, $SD = 0.17$; $t(36) = 0.17$, $p = .870$). Taken together, these findings indicate that the SMT and MMT groups did not differ in levels of adherence.

Table 7

Group Differences in Percent of Practice Elements Delivered Across Treatments

SMT *M (SD)*	MMT *M (SD)*	*t*	*p*
0.84 *(0.23)*	0.85 *(0.17)*	.165	.870

Note. SMT means were generated using the CBAY-A SMT subscale and MMT means were generated using the CBAY-A MMT subscale.

Discussion

The present study examined the score reliability and validity of the Cognitive-Behavioral Therapy Adherence Scale for Youth Anxiety (CBAY-A; Southam-Gerow et al., 2016), an instrument designed to assess adherence to common practice elements found in ICBT for youth anxiety, for use with two types of ICBT protocols: *Coping Cat* (i.e., a standard manualized treatment; SMT) and *MATCH* (i.e., a modular manualized treatment; MMT). Scores on the CBAY-A SMT and MMT subscales were compared to scores on two instruments of adherence

designed specifically for *Coping Cat* and *MATCH* to determine the ability of the CBAY-A to be used across different types of ICBT treatment protocols. Findings largely supported the score reliability and construct validity of the CBAY-A SMT and MMT subscales across both treatment groups. Independent raters were able to reliably code a variety of practice elements for ICBT for youth anxiety. The subscale scores also demonstrated convergent validity in both treatment groups, with moderate to large correlations with separate adherence instruments. Discriminant validity was likewise supported for the CBAY-A SMT subscale which produced higher correlations with another instrument of adherence than instruments of competence and alliance. In contrast, discriminant validity was partially supported for the CBAY-A MMT subscale, which produced higher correlations with another instrument of adherence than an instrument of competence, but not alliance. Further, scores on the CBAY-A SMT and MMT subscales displayed similar patterns of convergent and discriminant validity as the Standard and MATCH Consultation Records. Contrary to our hypotheses, results did not support discriminative validity as group differences in levels of adherence were not found.

Reliability

All CBAY-A SMT and CBAY-A MMT subscale items demonstrated *excellent* inter-rater reliability (ICC ≥ .75; Cicchetti, 1994) in each treatment group with the exception of one CBAY-A SMT subscale item (*Problem Solving*) and one CBAY-A MMT subscale item (*Coping Plan*) which each had *good* inter-rater reliability. These findings indicate that the CBAY-A items can be reliably rated in both groups, supporting its ability to be used across multiple ICBT treatment protocols for youth anxiety. Further, inter-rater reliability was *excellent* for both CBAY-A SMT and CBAY-A MMT subscale scores, suggesting that score variance was not due to coder characteristics, and that the CBAY-A SMT and MMT subscales would likely perform equally

when rated by different coders (Hallgren, 2012). The ability of the CBAY-A to be rated reliably across different ICBT protocols and coders suggests that it may be a useful tool for measuring and comparing levels of adherence across both research and clinical contexts where treatment protocols and coders differ.

Construct validity: Convergent and discriminant validity

CBAY-A SMT and MMT subscale scores were hypothesized to demonstrate strong correlations with two other instruments designed to assess adherence to *Coping Cat* and *MATCH* respectively. The correlation between scores on the CBAY-A SMT subscale and the Standard Consultation Record was in the moderate range, and the CBAY-A MMT subscale score evidenced a strong correlation with the *MATCH* Consultation Record.

The magnitudes of the correlations between the CBAY-A SMT and MMT subscales and the Standard and MATCH Consultation Records (*rs* .35 to .47) are lower than those demonstrated by McLeod et al. (2018), who found an average correlation of $r = .52$ when comparing the CBAY-A Skills and Exposure subscales to a Coping Cat subscale on the Therapy Process Observational Coding System for Child Psychotherapy—Revised Scale (TPOCS-RS; McLeod et al., 2015). A potential explanation for this difference is inconsistency in method (i.e., raters, coding procedures, instrument design, and scoring strategy). The instruments compared by McLeod et al. (2018; i.e., the CBAY-A and the TPOCS-RS), were both rated by independent coders who underwent very similar training (i.e., learned a coding manual, reached and maintained ICC > .60, regular coding meetings). The Standard and MATCH Consultation Records, however, were rated by study consultants in weekly supervision during which therapists reported on what practices they delivered. It is expected for scores on two instruments of the same construct that utilize the same methods to demonstrate a stronger correlation than scores on

two instruments of the same construct that utilize differential methods (Campbell & Fiske, 1959). Thus, the patterns of correlations observed in this study support the convergent validity of the CBAY-A SMT and CBAY-A MMT subscales. More generally, the findings support the ability of the CBAY-A to capture adherence across two ICBT protocols for youth anxiety.

To be used in place of protocol-specific adherence instruments, the CBAY-A must display a similar ability to discriminate between adherence and distinct constructs as the protocol-specific instruments. The CBAY-A SMT and MMT subscales and the Standard and MATCH Consultation Records (i.e., the protocol-specific adherence instrument for the SMT and MMT groups respectively), all overlapped to a greater degree with another instrument of adherence than with instruments of competence and alliance. Further, the pattern of correlations observed for the CBAY-A subscale and the Standard Consultation Records across both treatment groups support the discriminant validity of the subscale scores. The patterns of these correlations were similar to those found in previous research comparing the relation between adherence, competence, and alliance (Carroll et al., 2000; Hogue et al., 2008; McLeod et al., 2018; Southam-Gerow et al., 2016).

It is important to note that while the magnitude of the correlation between the CBAY-A MMT subscale and an instrument of adherence was larger than the magnitude of the correlation between the CBAY-A MMT subscale and an instrument of alliance, follow-up contrasts revealed that the correlations were not significantly different. It is possible that that the design of the CBAY-A MMT subscale contributes to its inability to differentiate between adherence and alliance; however, given that this pattern did not emerge with the SMT subscale and that the MMT scale was able to successfully differentiate between adherence and competence, alternative explanations should be considered.

It is possible that in practice, adherence and alliance are more strongly associated in *MATCH* than *Coping Cat*, given that *MATCH* allows therapists to flexibly deliver treatment to meet the individual needs of a client (Chorpita & Weisz, 2005). If this were the case, findings would not indicate that the CBAY-A MMT subscale is unable to differentiate between adherence and alliance, but rather that they are highly related in practice. Past research has demonstrated that the use of manualized treatments does not undermine alliance and in some cases, may enhance it (Langer, McLeod, & Weisz, 2011; McLeod et al., 2016); but to our knowledge, alliance has not been compared between a standard manualized treatment and a modular manualized treatment. Overall, findings are largely supportive of the ability of the CBAY-A MMT subscale to discriminate between adherence and competence, but its ability to discriminate between adherence and alliance is less clear. To expand on these findings, future studies should examine differences in alliance across treatment modalities, and specifically whether the added flexibility of modular treatments increases alliance.

Additionally, unlike the CBAY-A subscales, neither consultation record had a significant correlation with either competence or alliance. Common method variance (i.e., "variance that is attributable to the measurement method rather than to the constructs the measures represent;" Podsakoff, MacKenzie, Lee, & Podsakoff, 2003, p. 879) is likely to have contributed to these findings. The CBAY-A SMT and MMT subscale correlations with instruments of competence (CBAY-C SMT and MMT subscales) and alliance (TPOCS-A) may have been inflated due to common scale formats and scoring methods with the CBAY-C subscales and the TPOCS-A, and item overlap with the CBAY-C (Podsakoff et al., 2003). Conversely, the lack of method overlap between the Standard and MATCH Consultation Records and the CBAY-C and TPOCS-A may have introduced error, thereby reducing power and producing deflated correlations. While results

should be interpreted with these limitations in mind, the CBAY-A SMT and MMT subscales did display some ability to differentiate between adherence and distinct constructs, and produced correlations with these constructs in a similar pattern to those produced by the Standard and MATCH Consultation Records.

Overall, the findings support the score convergent and discriminant validity of the CBAY-A SMT and MMT subscales when compared to a separate adherence instrument, further supporting its ability to be used in place of protocol-specific adherence instruments.

Discriminative Validity

When examining group differences in adherence between the SMT and MMT groups using the Standard and MATCH Consultation Records, Weisz et al. (2012) found that therapists in the SMT group had higher rates of adherence than therapists in the MMT group. Contrary to their findings and our hypotheses, we found no group differences in adherence when each group was scored using the CBAY-A SMT and MMT subscales.

A number of factors may explain our findings. First, Weisz et al. (2012) did not examine adherence differences separately for the three primary problem areas (i.e., anxiety, depression, and conduct), so it is possible that when looking at the anxiety group only, the group differences they found would not have held. Second, because the Standard and MATCH Consultation Records were protocol specific, they included every proscribed treatment element for each treatment, whereas the practice element approach of the CBAY-A subscales runs the risk of misrepresenting the delivery of some treatment elements. For example, the MATCH Consultation Record included a number of items representing depression- and conduct-focused practice elements. Because the focus of the current study was on ICBT for youth anxiety, these items were not included in our analyses, but were included in the analyses done by Weisz et al.

(2012). Finally, due to differences in data collection group differences in adherence could not be assessed in the same way they were by Weisz et al. (2012), as described in the data analysis plan. We attempted to reduce this limitation by assessing group differences in two ways: the first in a way more consistent with the design of the CBAY-A (comparing average extensiveness ratings), and the second in a way more consistent with the design of the consultation records (comparing percent of practice elements delivered). Despite these efforts, our analyses did not produce group differences in CBAY-A SMT and MMT subscale scores; thus, discriminative validity was not supported.

Implications

The current study represents an important first step in evaluating the ability of a practice elements-based adherence instrument to be used in place of instruments designed for individual protocols. Such an instrument could be used more widely due to its flexibility, which promotes the accumulation of psychometric information across a range of settings and samples, as well as knowledge accumulation resulting from cross-treatment and cross-study comparisons (Martinez et al., 2014; Schoenwald et al., 2011; Malik et al., 2003). Further, it would reduce the resources required by individual research teams to create and utilize new adherence instruments (Perepletchikova et al., 2007).

It is important to recognize that in addition to providing psychometric evidence for the CBAY-A SMT and MMT subscales, convergent validity of the Standard and MATCH Consultation Records was supported. The moderate to high relation between scores on the consultation records and the CBAY-A subscales provide validity evidence that can supplement the reliability evidence found by Ward et al. (2013). Because the consultation records rely on self- or consultant-report and are rated as presence/absence, they may be more pragmatic than

the CBAY-A for use in clinical practice. However, the CBAY-A provides some distinct benefits. For example, unlike the consultation records, the CBAY-A was designed to be used across ICBT protocols for youth anxiety beyond *Coping Cat* and *MATCH* and could therefore be more widely used across studies and clinical sites where the specific protocols delivered may differ (McLeod et al., 2013). Additionally, the CBAY-A items are rated on a Likert-type extensiveness rating scale, and thus provide information that may be useful when evaluating implementation success (McLeod et al., 2013). For example, to evaluate whether a training was successful, researchers likely would want to know the frequency and thoroughness with which an intervention is delivered rather than simply whether or not a therapist reports delivering the intervention. An extensiveness rating scale is also conducive to utilizing treatment integrity data as quality indicators used to guide quality improvement efforts in clinical settings, such as using benchmarking as a feedback system (McLeod et al., 2013). An alternative approach, then, may be to create and utilize instruments that combine strengths of both instruments (e.g., it is pragmatic and can be rated as self-report, and also includes non-protocol-specific practice elements for use across protocols, studies, and settings).

To my knowledge, this is the first examination of the use of practice elements-based adherence instruments across two separate ICBT protocols. Results provide preliminary evidence that the CBAY-A can produce scores that can be reliably and validly interpreted across separate ICBT protocols for youth anxiety. If our findings are replicated across settings and samples, the CBAY-A can be a useful tool for implementation researchers as well as for evaluating service quality in practice settings.

Limitations

While the current study has several strengths and contributes to the treatment integrity literature, a number of limitations warrant discussion. First, as previously noted, methodological inconsistencies in design, coding procedures, and scoring strategy between the CBAY-A SMT and MMT subscales, the Standard and MATCH Consultation Records, the CBAY-C SMT and MMT subscales, and the TPOCS-A may reduce the comparability of the instruments. For example, it is possible that rater effects influenced the results, as coders differed in important ways between the current study and the Weisz et al. (2012) study. Specifically, coders in the current study were doctoral students who were trained to reliability on an observational instrument, whereas the raters in the Weisz et al. (2010) study were consultants who made ratings based on therapist report. Coders also differed in their knowledge of treatment session number and expected content, as CBAY-A coders did not code treatment sessions in order, and thus may have had difficulty identifying treatment elements out of context that were not explicitly stated (Hogue et al., 2008). These methodological differences are likely to increase error and decrease power to detect small effects (i.e., less pronounced relations between constructs; Podsakoff et al., 2003).

An additional issue of comparability is that the CBAY-A SMT and MMT subscales define adherence differently than do the consultation records. The consultation records include every item that is found in their respective treatment protocols (including depression and conduct items for *MATCH*), but only those items. Conversely, the CBAY-A SMT and MMT subscales do not include all items found in the individual treatment protocols, but do include additional practice elements that may not be found in every treatment protocol; however, we attempted to reduce this limitation by creating CBAY-A subscales that only included items found in their

respective treatment protocols. Moreover, therapists in the SMT group were assigned a manualized treatment to deliver based on the client's primary problem area (i.e., *Coping Cat* for the current study), but therapists were able to move on to an additional manualized treatment (e.g., *Defiant Children*) after completion of *Coping Cat* if they considered it necessary. It is possible, then, that some coded sessions in the SMT group were not *Coping Cat* sessions.

Lastly, although all available participant sessions were selected for coding with the exception of the first and last sessions, some videos could not be coded due to audio or visual issues, or because the client was out of the room for the majority of the session. While the percent of coded sessions did not differ across groups, important treatment content could have been missed.

Future Directions

The current study provides initial evidence that the CBAY-A can produce reliable and valid scores across two separate ICBT protocols for youth anxiety. To be considered "high quality," instruments should undergo psychometric evaluations across multiple studies, samples, and settings (Martinez et al., 2014). Thus, while replication in similar samples provides useful information, future studies should also attempt to replicate reliability and validity findings across a wide variety of samples and settings. Additionally, because the CBAY-A is not intended to be used only with *Coping Cat* and *MATCH* treatment sessions, future research should also evaluate psychometric properties of the CBAY-A when used with alternative ICBT treatment protocols for youth anxiety.

The CBAY-A should also be compared to instruments with a range of common method variance. Because the current study evaluated convergent validity with an instrument that differed in design, rating strategy, and scoring strategy, the CBAY-A should be compared to an

instrument that is more equivalent in these areas (e.g., rated by independent observes, rated on a Likert-type scale). However, triangulation of findings resulting from various data sources and methods provides additional validity evidence, so comparisons of the CBAY-A with instruments that differ in method or scoring are useful (Mathison, 1988). Finally, the current study provided preliminary evidence for representative validity, so future research should focus on elaborative validity (Foster & Cone, 1995).

References

Achenbach, T. M., Dumenci, L., & Rescorla, L. A. (2003). DSM-oriented and empirically based approaches to constructing scales from the same item pools. *Journal of Clinical Child and Adolescent Psychology*, *32*(3), 328-340.

Achenbach, T. M., & Rescorla, L. (2001). ASEBA school-age forms & profiles.

American Academy of Child and Adolescent Psychiatry (2006). *AACAP Workforce Fact Sheet.* Washington, D.C: American Academy of Child and Adolescent Psychiatry.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. Washington, DC: American Psychiatric Association.

American Psychological Association Task Force on Evidence-Based Practice for Children and Adolescents. (2008). Disseminating evidence-based practice for children & adolescents: A systems approach to enhancing care: APA task force on evidence-based practice with children and adolescents. Washington, D.C.: American Psychological Association.

Allen, J. D., Linnan, L. A., & Emmons, K. M. (2012). Fidelity and its relationship to implementation effectiveness, adaptation, and dissemination. In R. C. Brownson, G. A. Colditz & E. K. Proctor (Eds.), *Dissemination and implementation research in health: Translating science to practice* (pp. 281-304). New York: Oxford University Press.

Baer, J. S., Ball, S. A., Campbell, B. K., Miele, G. M., Schoener, E. P., & Tracy, K. (2007). Training and fidelity monitoring of behavioral interventions in multi-site addictions research. *Drug and Alcohol Dependence, 87*(2), 107-118. doi:10.1016/j.drugalcdep.2006.08.028

Barber, J. P., Foltz, C., Crits-Christoph, P., & Chittams, J. (2004). Therapists' adherence and

competence and treatment discrimination in the NIDA Collaborative Cocaine Treatment Study. *Journal of Clinical Psychology, 60*(1), 29-41. doi: 10.1002/jclp.10186

Barber, J. P., Liese, B. S., & Abrams, M. J. (2003). Development of the cognitive therapy adherence and competence scale. *Psychotherapy Research, 13*(2), 205-221. doi:10.1093/ptr/kpg019

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. doi:10.1037/0735-7028.38.5.493

Barkley, R. A. (2013). *Defiant children: A clinician's manual for assessment and parent training*. Guilford press.

Barrett, P. M., Dadds, M. R., & Rapee, R. M. (1996). Family treatment of childhood anxiety: A controlled trial. *Journal of Consulting and Clinical Psychology, 64*(2), 333-342. doi:10.1037/0022-006X.64.2.333

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*(1), 29-43. doi:10.1375/bech.22.1.29.66786

Barth, R. P., Lee, B. R., Lindsey, M. A., Collins, K. S., Strieder, F., Chorpita, B. F., … Sparks, J. A. (2012). Evidence-based practice at a crossroads: The timely emergence of common elements and common factors. *Research on Social Work Practice, 22*(1), 108-119. doi:10.1177/1049731511408440

Beidas, R. S., & Kendall, P. C. (2010). Training therapists in evidence-based practice: A critical

review of studies from a systems-contextual perspective. *Clinical Psychology: Science and Practice, 17*(1), 1-30. doi:10.1177/1077559512436674

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., … Czajkowski, S. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology, 23*(5), 443-451. doi:10.1037/0278-6133.23.5.443

Bickman, L. (2008). Why don't we have effective mental health services? *Administration and Policy in Mental Health and Mental Health Services Research, 35*, 437-439. doi:10.1007/s10488-008-0192-9

Boustani, M., Gellatly, R., & Chorpita, B. F. (2017). Advances in cognitive behavioral treatment design: Time for a Glossary. *The Behavior Therapist, 40*(6), 199-208.

Campbell, M. K., Elbourne, D. R., & Altman, D. G. (2004). CONSORT statement: Extension to cluster randomised trials. *BMJ*, *328*(7441), 702-708.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105. doi:10.1037/h0046016

Carroll, K. M., Martino, S., & Rounsaville, B. J. (2010). No train, no gain? *Clinical Psychology: Science and Practice, 17*(1), 36-40. doi:10.1111/j.1468-2850.2009.01190.x

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. doi:10.1016/S0376-8716(99)00049-6

Chaudoir, S. R., Dugan, A. G., & Barr, C. H. (2013). Measuring factors affecting implementation

of health innovations: A systematic review of structural, organizational, provider, patient, and innovation level measures. *Implementation Science, 8*, 22. doi:10.1186/1748-5908-8-22

Chor, K. H. B., Wisdom, J. P., Olin, S. C. S., Hoagwood, K. E., & Horwitz, S. M. (2015). Measures for predictors of innovation adoption. *Administration and Policy in Mental Health and Mental Health Services Research, 42*(5), 545-573. doi: 10.1007/s10488-014-0551-7

Chorpita, B. F., & Daleiden, E. L. (2009). Mapping evidence-based treatments for children and adolescents: Application of the distillation and matching model to 615 treatments from 322 randomized trials. *Journal of Consulting and Clinical Psychology, 77*(3), 566-579. doi:10.1037/a0014565

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. doi:10.1007/s11020-005-1962-6

Chorpita, B. F., & Weisz, J. R. (2005). *MATCH-ADC: A Modular Approach to Treatment for Children with Anxiety, Depression, and Conduct Problems*. Honolulu, HI, and Boston, MA: University of Hawaii at Manoa and Judge Baker Children's Center, Harvard Medical School.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290. doi: 10.1037/1040-3590.6.4.284

Clinton-McHarg, T., Yoong, S. L., Tzelepis, F., Regan, T., Fielding, A., Skelton, E., …

Wolfenden, L. (2016). Psychometric properties of implementation measures for public health and community settings and mapping of constructs against the Consolidated Framework for Implementation Research: a systematic review. *Implementation Science, 11*, 148. doi:10.1186/s13012-016-0512-5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine, 119*(2), 166.e7-166.e16. doi:10.1016/j.amjmed.2005.10.036

Cook, J. M., O'Donnell, C., Dinnen, S., Coyne, J. C., Ruzek, J. I., & Schnurr, P. P. (2012). Measurement of a model of implementation for health care: Toward a testable theory. *Implementation Science, 7*, 59. doi:10.1186/1748-5908-7-59

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. *Implementation Science, 4,* 50. doi:10.1186/1748-5908-4-50

DeVellis, R. F. (2017). *Scale development: Theory and applications*. Los Angeles: Sage Publications, Inc.

Dobson, D., & Cook, T. J. (1980). Avoiding type III error in program evaluation: Results from a field experiment. *Evaluation and Program Planning, 3*(4), 269-276. doi:10.1016/0149-7189(80)90042-7

Dorsey, S., McLaughlin, K. A., Kerns, S. E., Harrison, J. P., Lambert, H. K., Briggs, E. C., …

Amaya-Jackson, L. (2017). Evidence base update for psychosocial treatments for children and adolescents exposed to traumatic events. *Journal of Clinical Child & Adolescent Psychology, 46*(3), 303-330. doi:10.1080/15374416.2016.1220309

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*(1), 61-82. doi:10.1207/s15324818ame1001_4

Drake, R. E., Goldman, H. H., Leff, H. S., Lehman, A. F., Dixon, L., Mueser, K. T., & Torrey, W. C. (2001). Implementing evidence-based practices in routine mental health service settings. *Psychiatric Services, 52*(2), 179-182. doi:10.1176/appi.ps.52.2.179

Eccles, M. P., Armstrong, D., Baker, R., Cleary, K., Davies, H., Davies, S., … Sibbald, B. (2009). An implementation research agenda. *Implementation Science, 4*, 18. doi:10.1186/1748-5908-4-18

Eccles, M. P., & Mittman, B. S. (2006). Welcome to implementation science. *Implementation Science, 1*, 1.

Emmons, K. M., Weiner, B., Fernandez, M. E., & Tu, S. P. (2012). Systems antecedents for dissemination and implementation: A review and analysis of measures. *Health Education & Behavior, 39*(1), 87-105. doi:10.1177/1090198111409748

Evans, S. W., Owens, J. S., & Bunford, N. (2014). Evidence-based psychosocial treatments for children and adolescents with attention-deficit/hyperactivity disorder. *Journal of Clinical Child & Adolescent Psychology, 43*(4), 527-551. doi:10.1080/15374416.2013.850700

Field, A. (2009). *Discovering statistics using SPSS*. Los Angeles: Sage Publications, Inc.

Foster, S. L., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment, 7*(3), 248-260. doi:10.1037/1040-3590.7.3.248

Freeman, H. E., & Rossi, P. H. (1985). *Collecting evaluation data: Problems and solutions*. Sage Publications, Inc.

Fristad, M. A., Glickman, A. R., Verducci, J. S., Teare, M., Weller, E. B., & Weller, R. A. (1998). Study V: Children's Interview for Psychiatric Syndromes (ChIPS): Psychometrics in two community samples. *Journal of Child and Adolescent Psychopharmacology*, *8*(4), 237-245.

Garland, A. F., Hurlburt, M. S., & Hawley, K. M. (2006). Examining psychotherapy processes in a services research context. *Clinical Psychology: Science and Practice, 13*(1), 30-46. doi:10.1111/j.1468-2850.2006.00004.x

George, D., & Mallery, P. (2016). *IBM SPSS Statistics 23 Step by Step: A Simple Guide and Reference (13th ed.)*. New York: Routledge.

Garland, A., & Schoenwald, S. K. (2013). Use of effective and efficient quality control methods to implement psychosocial interventions. *Clinical Psychology: Science and Practice, 20*(1), 33-43. doi:10.1111/cpsp.12021

Goense, P., Boendermaker, L., van Yperen, T., Stams, G. J., & van Laar, J. (2014). Implementation of treatment integrity procedures: An analysis of outcome studies of youth interventions targeting externalizing behavioral problems. *Zeitschrift für Psychologie, 222*(1), 12-21. doi:10.1027/2151-2604/a000161

Hagermoser Sanetti, L. M., Gritter, K. L., & Dobey, L. M. (2011). Treatment integrity of interventions with children in the school psychology literature from 1995 to 2008. *School Psychology Review, 40*(1), 72-84.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34.

Higa-McMillan, C. K., Francis, S. E., Rith-Najarian, L., & Chorpita, B. F. (2016). Evidence base update: 50 years of research on treatment for child and adolescent anxiety. *Journal of Clinical Child & Adolescent Psychology, 45*(2), 91-113. doi:10.1080/15374416.2015.1046177

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. doi:10.1016/j.jsat.2007.09.002

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), 544-555. doi:10.1037/0022-006X.76.4.544

Hogue, A., Liddle, H. A., & Rowe, C. (1996). Treatment adherence process research in family therapy: A rationale and some practical guidelines. *Psychotherapy: Theory, Research, Practice, Training, 33*(2), 332-345. doi:10.1037/0033-3204.33.2.332

Hurlburt, M. S., Garland, A. F., Nguyen, K., & Brookman-Frazee, L. (2010). Child and family therapy process: Concordance of therapist and observational perspectives. *Administration and Policy in Mental Health and Mental Health Services Research, 37*(3), 230-244. doi:10.1007/s10488-009-0251-x

Insel, T. R. (2008). *The national institute of mental health strategic plan.* Bethesda, Maryland: US Department of Health and Human Services.

Institute of Medicine (2015). *Improving the quality of health care for mental and substance-use conditions: Quality chasm series.* Washington, DC: National Academies Press.

Julien, D., Markman, H. J., & Lindahl, K. M. (1989). A comparison of a global and a microanalytic coding system: Implications for future trends in studying interactions. *Behavioral Assessment, 11*(1), 81-100.

Kaminski, J. W., & Claussen, A. H. (2017). Evidence base update for psychosocial treatments for disruptive behaviors in children. *Journal of Clinical Child & Adolescent Psychology*, *46*(4), 477-499. doi:10.1080/15374416.2017.1310044

Kazdin, A. E. (1994). Methodology, design, and evaluation in psychotherapy research. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (pp. 19-71). Oxford, England: John Wiley.

Kazdin, A. E., Bass, D., Ayers, W. A., & Rodgers, A. (1990). Empirical and clinical focus of child and adolescent psychotherapy research. *Journal of Consulting and Clinical Psychology, 58*(6), 729-740. doi:10.1037/0022-006X.58.6.729

Kendall, P. C. (1994). Treating anxiety disorders in children: Results of a randomized clinical trial. *Journal of Consulting and Clinical Psychology, 62*(1), 100-110. doi:10.1037/0022-006X.62.1.100

Kendall, P. C., & Beidas, R. S. (2007). Smoothing the trail for dissemination of evidence-based practices for youth: Flexibility within fidelity. *Professional Psychology: Research and Practice, 38*(1), 13-20. doi:10.1037/0735-7028.38.1.13

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. A. (2006b). *Coping cat workbook (2nd ed.)*. Ardmore, PA: Workbook Publishing.

Langer, D. A., McLeod, B. D., & Weisz, J. R. (2011). Do treatment manuals undermine youth–

therapist alliance in community clinical practice? *Journal of Consulting and Clinical Psychology*, *79*(4), 427-432. doi: 0.1037/a0023821

Lewis, C. C., Fischer, S., Weiner, B. J., Stanick, C., Kim, M., & Martinez, R. G. (2015). Outcomes for implementation science: An enhanced systematic review of instruments using evidence-based rating criteria. *Implementation Science, 10*, 155. doi:10.1186/s13012-015-0342-x

Lewis, C. C., Stanick, C. F., Martinez, R. G., Weiner, B. J., Kim, M., Barwick, M., & Comtois, K. A. (2015). The society for implementation research collaboration instrument review project: A methodology to promote rigorous evaluation. *Implementation Science, 10*, 2. doi:10.1186/s13012-014-0193-x

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. 10.1016/j.beth.2009.02.003

Lohr, K. N., Aaronson, N. K., Alonso, J., Burnam, M. A., Patrick, D. L., Perrin, E. B., & Roberts, J. S. (1996). Evaluating quality-of-life and health status instruments: Development of scientific review criteria. *Clinical Therapeutics, 18*(5), 979-992. doi:10.1016/S0149-2918(96)80054-3

Malik, M. L., Beutler, L. E., Alimohamed, S., Gallagher-Thompson, D., & Thompson, L. (2003). Are all cognitive therapies alike? A comparison of cognitive and noncognitive therapy process and implications for the application of empirically supported treatments. *Journal of Consulting and Clinical Psychology, 71*(1), 150-158. doi:10.1037/0022-006X.71.1.150

Margolin, G., Oliver, P., Gordis, E., O'Hearn, H., Medina, A., Ghosh, C., & Morland, L. (1998). The nuts and bolts of behavioral observation of marital and family interaction. *Clinical Child & Family Psychology Review, 1*, 195–213. doi:10.1023/A:1022608117322

Markman, H. J., Leber, B. D., Cordova, A. D., & St Peters, M. (1995). Behavioral observation and family psychology: Strange bedfellows or happy marriage? Comment on Alexander et al. (1995). *Journal of Family Psychology, 9*(4), 371-379. doi:10.1037/0893-3200.9.4.371

Martinez, R. G., Lewis, C. C., & Weiner, B. J. (2014). Instrumentation issues in implementation science. *Implementation Science, 9*, 118. doi: 10.1186/s13012-014-0118-8

Mathison, S. (1988). Why triangulate? *Educational Researcher*, *17*(2), 13-17. https://doi.org/10.3102/0013189X017002013

McLeod, B. D., Jensen-Doss, A., Tully, C. B., Southam-Gerow, M. A., Weisz, J. R., & Kendall, P. C. (2016). The role of setting versus treatment type in alliance within youth therapy. *Journal of Consulting and Clinical Psychology*, *84*(5), 453-464. http://dx.doi.org/10.1037/ccp0000081

McLeod, B. D., Southam-Gerow, M. A., Jensen-Doss, A., Hogue, A., Kendall, P. C., & Weisz, J. R. (2017). Benchmarking treatment adherence and therapist competence in individual cognitive-behavioral treatment for youth anxiety disorders. *Journal of Clinical Child & Adolescent Psychology*, *22*, 1-13. doi:10.1080/15374416.2017.1381914

McLeod, B. D., Southam-Gerow, M. A., Rodríguez, A., Quinoy, A. M., Arnold, C. C., Kendall, P. C., & Weisz, J. R. (2018). Development and initial psychometrics for a therapist competence instrument for CBT for youth anxiety. *Journal of Clinical Child & Adolescent Psychology*, 1-14. doi:10.1080/15374416.2016.1253018

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. doi:10.1111/cpsp.12020

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. doi:10.1037/0022-006X.73.2.323

Merikangas, K. R., He, J. P., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., … Swendsen, J. (2010). Lifetime prevalence of mental disorders in US adolescents: Results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry, 49*(10), 980-989. doi:10.1016/j.jaac.2010.05.017

Merikangas, K. R., He, J. P., Burstein, M., Swendsen, J., Avenevoli, S., Case, B., … Olfson, M. (2011). Service utilization for lifetime mental disorders in US adolescents: Results of the National Comorbidity Survey–Adolescent Supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry, 50*(1), 32-45. doi:10.1016/j.jaac.2010.10.006

Mihalic, S. F., & Irwin, K. (2003). Blueprints for violence prevention: From research to real-world settings—factors influencing the successful replication of model programs. *Youth Violence and Juvenile Justice, 1*(4), 307-329. doi:10.1177/1541204003255841

Mills, S. C., & Ragan, T. J. (2000). A tool for analyzing implementation fidelity of an integrated

learning system. *Educational Technology Research and Development, 48*(4), 21-41. doi:10.1007/BF02300498

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review, 11*(3), 247-266. doi:10.1016/0272-7358(91)90103-2

Moser, R. P., Hesse, B. W., Shaikh, A. R., Courtney, P., Morgan, G., Augustson, E., … Dunn, M. (2011). Grid-enabled measures: Using Science 2.0 to standardize measures and share data. *American Journal of Preventive Medicine, 40*(5), S134-S143. doi:10.1016/j.amepre.2011.01.004

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. doi:10.1177/109821400302400303

National Research Council. (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities*. National Academies Press.

Novick, M. R., & Lewis, C. (1966). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1-13.

Paulson, R. I., Post, R. L., Herinckx, H. A., & Risser, P. (2002). Beyond components: Using fidelity scales to measure and assure choice in program implementation and quality assurance. *Community Mental Health Journal, 38*(2), 119-128. doi:10.1023/A:1014591020400

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice, 12*(4), 365-383. doi:10.1093/clipsy.bpi045

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), 829-841. doi:10.1037/0022-006X.75.6.829

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879-903. doi:10.1037/0021-9010.88.5.879

Proctor, E. K., Landsverk, J., Aarons, G., Chambers, D., Glisson, C., & Mittman, B. (2009). Implementation research in mental health services: An emerging science with conceptual, methodological, and training challenges. *Administration and Policy in Mental Health and Mental Health Services Research, 36*(1), 24-34. doi:10.1007/s10488-008-0197-4

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., … Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(2), 65-76. doi:10.1007/s10488-010-0319-7

Powell, B. J., Proctor, E. K., & Glass, J. E. (2014). A systematic review of strategies for implementing empirically supported mental health interventions. *Research on Social Work Practice, 24*(2), 192-212. doi:10.1177/1049731513505778

Rabin, B. A., Purcell, P., Naveed, S., Moser, R. P., Henton, M. D., Proctor, E. K., … Glasgow, R. E. (2012). Advancing the application, quality and harmonization of implementation science measures. *Implementation Science, 7*, 1-11. doi:10.1186/1748-5908-1-1

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177. doi: 10.1037/1082-989X.7.2.147

Schniering, C. A., Rapee, R. M., Lyneham, H. J., Wuthrich, V., Hudson J. L., & Wignall, A.

(2006). *The Cool Kids® adolescent anxiety & depression program therapist manual.* Centre for Emotional Health, Macquarie University University: Sydney, Australia.

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. doi:10.1007/s10488-008-0181-z

Schoenwald, S. K., & Garland, A. F. (2013). A review of treatment adherence measurement methods. *Psychological Assessment, 25*(1), 146-156. doi:10.1037/a0029715

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(1), 32-43. doi:10.1007/s10488-010-0321-0

Scott, T., Mannion, R., Davies, H., & Marshall, M. (2003). The quantitative measurement of organizational culture in health care: A review of the available instruments. *Health Services Research, 38*(3), 923-945. doi:10.1111/1475-6773.00154

Sheridan, S. M., Swanger-Gagné, M., Welch, G. W., Kwon, K., & Garbacz, S. A. (2009). Fidelity measurement in consultation: Psychometric issues and preliminary examination. *School Psychology Review, 38*(4), 476-495.

Shirk, S. R., & Russell, R. L. (1998). Process issues in child psychotherapy. In A. Bellack & M. Hersen (Eds.), *Comprehensive clinical psychology, Vol. 15*, pp. 57-82. Oxford: Pergamon

Shirk, S. R., & Saiz, C. (1992). The therapeutic alliance in child therapy: Clinical, empirical, and developmental perspectives. *Development and Psychopathology, 4*(4), 713-728.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability.

Psychological Bulletin, *86*(2), 420-428.

Silverman, W. K., & Nelles, W. B. (1988). The anxiety disorders interview schedule for children. *Journal of the American Academy of Child & Adolescent Psychiatry*, *27*(6), 772-778.

Southam-Gerow, M. A., McLeod, B. D., Arnold, C. C., Rodríguez, A., Cox, J. R., Reise, S. P., … Kendall, P. C. (2016). Initial development of a treatment adherence measure for cognitive–behavioral therapy for child anxiety. *Psychological Assessment, 28*(1), 70-80. doi:10.1037/pas0000141

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. doi:10.1016/j.jaac.2010.06.009

Tabachnik, B. G., & Fidell, S. L. (2007). *Using multivariate statistics.* Boston: Pearson Education Inc.

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.-630. doi:10.1037/0022-006X.61.4.620

Ward, A. M., Regan, J., Chorpita, B. F., Starace, N., Rodriguez, A., Okamura, K., … Research Network on Youth Mental Health. (2013). Tracking evidence based practice with youth: Validity of the MATCH and Standard Manual Consultation Records. *Journal of Clinical Child & Adolescent Psychology, 42*(1), 44-55. doi:10.1080/15374416.2012.700505

Weersing, R., Jeffreys, M., Do, M., Schwartz, K. G., & Bolano, C. (2016). Evidence-based update: Treatment of depression in children and adolescents. *Journal of Clinical Child & Adolescent Psychology, 46*(1), 11-43). doi:10.1080/15374416.2016.1220310

Weiner, B. J., Amick, H., & Lee, S. Y. D. (2008). Conceptualization and measurement of organizational readiness for change: A review of the literature in health services research and other fields. *Medical Care Research and Review, 65*(4), 379-436. doi:10.1177/1077558708317802

Weisz, J. R., Chorpita, B. F., Frye, A., Ng, M. Y., Lau, N., Bearman, S. K., ... & Hoagwood, K. E. (2011). Youth Top Problems: Using idiographic, consumer-guided assessment to identify treatment needs and to track change during psychotherapy. *Journal of Consulting and Clinical Psychology, 79*(3), 369-380. doi:10.1037/a0023307

Weisz, J. R., Chorpita, B. F., Palinkas, L. A., Schoenwald, S. K., Miranda, J., Bearman, S. K., … Gray, J. (2012). Testing standard and modular designs for psychotherapy treating depression, anxiety, and conduct problems in youth: A randomized effectiveness trial. *Archives of General Psychiatry, 69*(3), 274-282. doi:10.1001/archgenpsychiatry.2011.147

Weisz, J. R., Doss, A. J., & Hawley, K. M. (2005). Youth psychotherapy outcome research: A review and critique of the evidence base. *Annual Review of Psychology, 56*, 337-363. doi:10.1146/annurev.psych.55.090902.141449

Weisz, J. R., Kuppens, S., Eckshtain, D., Ugueto, A. M., Hawley, K. M., & Jensen-Doss, A. (2013). Performance of evidence-based youth psychotherapies compared with usual clinical care: A multilevel meta-analysis. JAMA Psychiatry, 70(7), 750-761. doi:10.1001/jamapsychiatry.2013.1176

Weisz, J. R., Kuppens, S., Ng, M. Y., Eckshtain, D., Ugueto, A. M., Vaughn-Coaxum, R., … Weersing, V. R. (2017). What five decades of research tells us about the effects of youth psychological therapy: A multilevel meta-analysis and implications for science and practice. *American Psychologist, 72*(2), 79-117. doi:10.1037/a0040360

Weisz, J. R., Southam-Gerow, M. A., Gordis, E. B., Connor-Smith, J. K., Chu, B. C., Langer, D. A., … Weiss, B. (2009). Cognitive–behavioral therapy versus usual clinical care for youth depression: An initial test of transportability to community clinics and clinicians. *Journal of Consulting and Clinical Psychology, 77*(3), 383-396. doi:10.1037/a0013877

Weisz, J. R., Weersing, V. R., Valeri, S. M., & McCarty, C. A. (1999). *Therapist's Manual PASCET: Primary and secondary control enhancement training program*. Los Angeles: University of California.

Weller, E. B., Weller, R. A., Fristad, M. A., Rooney, M. T., & Schecter, J. (2000). Children's Interview for Psychiatric Syndromes (ChIPS). *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*(1), 76-84.

Weller, E. B., Weller, R. A., Rooney, M. T., & Fristad, M. A. (1999). Children's Interview for Psychiatric Syndromes: Parent version. *Washington, DC: American Psychiatric Association*.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49*(2), 156-167. doi:10.1037/0022-006X.49.2.156

Appendix A: CBAY-A

	1 Not at all	2	3 Somewhat	4	5 Considerably	6	7 Extensively			
	Item							**Freq**	**Ext**	
Standard	1. HW Review									1.
	2. HW Assigned									2.
	3. Agenda Setting									3.
	4. Rapport Building									4.
	5. Weekly Rating									5.
	6. Assessment									6.
	Problem Item: A D C									
Model	7. Psychoed									7.
	8. Emotion Ed									8.
	9. Fear Ladder									9.
	10. Relaxation									10.
	11. Cognitive									11.
	12. Problem Solving									12.
	13. Self-Reward									13.
	14. Coping Plan									14.
	15. Exposure: Prep									15.
	16. Exposure									16.
	17. Exposure: Debrief									17.
	18. Maintenance									18.
	19. Activity Selection									19.
	20. Talents and Skills									20.
	21. Positive Self									21.
	22. One-on-One Time									22.
	23. Praise									23.
	24. Active Ignoring									24.
	25. Effective Instructions									25.
	26. Rewards									26.
	27. Time Out									27.
	28. Daily Report Card									28.
Delivery	29. Didactic teaching									29.
	30. Collaborative teaching									30.
	31. Modeling									31.
	32. Rehearsal									32.
	33. Coaching									33.
	34. Self-disclosure									34.
	35. Play/Creative									35.

Appendix B: Standard Consultation Record

	Before Session			In Session								In Supervision				
	HW Done Full	HW Done Part	HW Not Done	Covered Full	Covered Part	Role Played	In Vivo	Understood Full	Understood Part	Understood No	HW Assigned	Reviewed Tape	Discussed	Next Session	Modeled	Role Played
Anxiety Skill																
Engagement-Parent	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Engagement-Child	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Psychoeducation-C	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Psychoeducation-P	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Relaxation	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Coping Thoughts	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Problem Solving	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Results & Rewards	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
FEAR Plan	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Parent Involvement	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Practice	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Review, Commercial, Party	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Depression Skill																
Formulation & Orientation-P	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Home Visit	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
School Visit	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Getting Acquainted-C	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Psychoeducation-C	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
STEPS	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Activity Selection	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Relaxation	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Secret Calming	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Positive Self	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Talents & Skills	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Formulation Review-P	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Think Positive	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
H-I-N	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Keep Thinking	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
3-Step Plan	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Best Fit ACT & THINK	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
ACT & THINK in Real Life	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Final Summation-P	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Review, Commercial, Party	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Conduct Skill																
Assessment & Engagement	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Why Children Misbehave	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Pay Attention	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Attending to Compliance	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Independent Play	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Effective Instructions	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Compliance Training	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Home Point System	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Response Cost	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Time Out (Noncompliance)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Time Out (House Rules)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Anticipating Problems	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Daily Report Card	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Future Problems	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Booster	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Appendix C: MATCH Consultation Record

	Before Session			In Session								In Supervision				
	HW Done Full	HW Done Part	HW Not Done	Covered Full	Covered Part	Role Played	In Vivo	Understood Full	Understood Part	Understood No	HW Assigned	Reviewed Tape	Discussed	Next Session	Modeled	Role Played
Anxiety Skill																
Getting acquainted	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Fear Ladder	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Psychoeducation-Child	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Psychoeducation-Parent	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Practicing	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Maintenance	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Cognitive STOP	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Wrap-up	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Depression Skill																
Getting acquainted	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Psychoeducation-Child	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Psychoeducation-Parent	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Problem Solving	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Activity Selection	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Relaxation	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Secret Calming	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Talents & Skills	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Positive Self	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Cognitive BLUE	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Cognitive FUN	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Three Step Plan	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Wrap-up	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Home Visit	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
School Visit	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Conduct Skill																
Engaging Parents	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Psychoeducation-Parent	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
One on One Time	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Praise	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Active Ignoring	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Effective Instructions	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Rewards	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Time Out	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Anticipating Problems	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Daily Report Card	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Future Problems	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Booster	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Vita

Stephanie Violante was born on November 19, 1991 in Seattle, Washington. She graduated with great distinction from the University of Washington in June 2013 with a Bachelor of Science in Psychology. She then worked for three years as a research assistant at the University of Washington and Seattle Children's Research Institute before entering the Clinical Psychology doctoral program at Virginia Commonwealth University in Richmond, Virginia.