Initial Development of a Treatment Adherence Measure for Cognitive–Behavioral Therapy for Child Anxiety

Michael A. Southam-Gerow, Bryce D. McLeod,
Cassidy C. Arnold, Adriana Rodríguez,
and Julia R. Cox
Virginia Commonwealth University

Steven P. Reise and Wesley E. Bonifay
University of California-Los Angeles

John R. Weisz
Harvard University

Philip C. Kendall
Temple University

The measurement of treatment adherence (a component of treatment integrity defined as the extent to which a treatment is delivered as intended) is a critical element in treatment evaluation research. This article presents initial psychometric data for scores on the Cognitive–Behavioral Therapy Adherence Scale for Youth Anxiety (CBAY-A), an observational measure designed to be sensitive to common practice elements found in individual cognitive–behavioral therapy (ICBT) for youth anxiety. Therapy sessions ($N = 954$) from 1 efficacy and 1 effectiveness study of ICBT for youth anxiety were independently rated by 2 coders. Interrater reliability (as gauged by intraclass correlation coefficients) for the item scores averaged 0.77 ($SD = 0.15$; range .48 to .80). The CBAY-A item and scale (skills, model, total) scores demonstrated evidence of convergent and discriminant validity with an observational measure of therapeutic interventions and an observational measure of the alliance. The CBAY-A item and scale scores also discriminated between therapists delivering ICBT in research and practice settings and therapists delivering nonmanualized usual clinical care. We discuss the importance of replicating these psychometric findings in different samples and highlight possible application of an adherence measure in testing integrity-outcome relations.

Keywords: adherence, treatment integrity, child anxiety, CBT

Treatment outcome research requires well-specified treatments that are delivered as designed (Comer & Kendall, 2013). Thus, treatment integrity, a term that encompasses adherence (how closely treatment delivered matches the intended plan), differentiation (the extent to which nonprescribed treatment content is present), and competence (the quality of treatment delivery) represents a critical focus for clinical science (e.g., Gresham, 2009; Hagermoser Sanetti & Kratochwill, 2009). However, reviews of treatment studies have concluded that evidence of treatment integrity is lacking. Integrity measurement is often used as a manipulation check (i.e., are the levels of the independent variable different from one another?), with integrity conceptualized in a binary fashion (e.g., Perepletchikova, Treat, & Kazdin, 2007), measured at a superficial level (Schoenwald et al., 2011), and rarely facilitating analysis of the relationship between adherence and treatment outcome (Schoenwald et al., 2011). Such shortcomings limit conclusions that can be drawn about the effects of treatments and importance of their components.

Most development of integrity instruments has focused on adult therapy and recent reviews in the child therapy, school psychology, and applied behavior analysis literatures concluded that few randomized controlled trials (RCTs) adequately measure treatment integrity in youth therapy (e.g., Hagermoser Sanetti, Fallon, & Collier-Meeka, 2011; Perepletchikova et al., 2007). Perepletchikova, Treat, and Kazdin (2007) identified several key dimensions of treatment integrity measurement, including: (a) clear definition of integrity (e.g., treatment manual present, training present); (b) integrity measurement (e.g., use of a measure with psychometric properties); (c) evaluation of integrity data (e.g., training of coders, assessing interrater reliability); and (d) reporting integrity data (e.g., reporting a variety of scores). In a review of 147 randomized controlled trials (RCTs), they found that only 3.5% of the studies achieved adequate measurement across these

four dimensions. We updated their review, considering an additional 112 RCTs and arrived at the same conclusion: Measurement of treatment adherence was largely inadequate, with the scores from very few measures of youth treatment integrity supported by published reliability and validity data.

Further, for most prior measures, adherence has been assessed only in terms of the percentage of treatment components covered and calculated using a nominal scale (presence/absence of treatment components). Although this approach has intuitive appeal as a manipulation check, a measure using an interval extensiveness scale to gauge the dose of each treatment component has been advocated (Carroll et al., 2000; Hogue, Liddle, & Rowe, 1996). From a measurement perspective, interval rating scales have benefits relative to nominal scales as they allow one to average scores across raters, entire sessions, or the course of treatment (McLeod, Islam, & Wheat, 2013). Capturing the breadth and depth of therapeutic interventions found in treatments has important benefits for treatment integrity research relative to nominal scales. Namely, interval extensiveness scales can detect if therapists vary in the extent to which they deliver specific interventions, which is important as therapists have been found to vary in how much they employ different interventions (e.g., Garland, Hurlburt, Brookman-Frazee, Taylor, & Accurso, 2010; McLeod, Smith, Southam-Gerow, Weisz, & Kendall, 2015). Interval extensiveness scales could therefore be used to answer questions such as "How much exposure was delivered in treatment Group A versus treatment Group B?" or "How much of the cognitive intervention was needed to produce a favorable outcome?"

Past adherence measurement focused solely on measuring which specific and prescribed therapeutic interventions were present; in other words, the measures focused on the content (or the what) of treatment. This is appropriate for manipulation checks, but insufficient for the study of *how* treatments are delivered (e.g., Beidas & Kendall, 2010). To address the "how" question, we designed an instrument that measures content along with the method of content delivery (e.g., didactically or via rehearsal; see Garland et al., 2010).

In this study, we report interrater reliability data and validity data, including how well the item and scale scores: (a) relate to an established measure of therapeutic interventions (Therapy Process Observational Coding System for Child Psychotherapy—Revised Scale (TPOCS-RS); McLeod et al., 2015); (b) relate to theoretically distinct measures of alliance; and (c) discriminate groups of therapists providing (or not) individual CBT (ICBT) for youth anxiety. The measure includes (a) items that capture specific practice elements (i.e., "discrete clinical technique . . . used as part of larger intervention plan," Chorpita & Daleiden, 2009, p. 569) across multiple ICBT programs; (b) items that gauge *how* practice elements were delivered (e.g., didactically, via rehearsal); and (c) three proposed scale scores.

Method

Data Sources, Participants, and Recording Data Used

Data sources. Therapy process data were collected on 89 youth participants who participated in one of two RCTs. The Kendall Coping Cat Study (Kendall, Hudson, Gosch, Flannery-Schroeder, & Suveg, 2008) compared the relative efficacy of

ICBT, family-CBT, and an active control condition. The Youth Anxiety Study (YAS; Southam-Gerow et al., 2010) compared the effectiveness of ICBT with usual care (UC). The present study focused on the ICBT conditions from both studies (ICBT, YAS-ICBT) and the UC (YAS-UC) condition from the Southam-Gerow et al. (2010) study. The primary data were archival video- or audio-taped therapy sessions.

Treatments. From both studies, we focused on therapists who delivered *Coping Cat* (ICBT or YAS-ICBT) or therapists who delivered therapy in their usual way (YAS-UC). *Coping Cat*, an ICBT program designed to treat youth diagnosed with anxiety disorders (Kendall & Hedtke, 2006), includes 16–20 sessions conducted individually with the youth. The program teaches youth skills to manage anxiety (e.g., cognitive restructuring; changing self-talk) using a FEAR acronym: (a) *F*eeling frightened? (identify symptoms of anxiety); (b) *E*xpecting bad things to happen? (recognize anxious thoughts); (c) *A*ctions and attitudes that can help (identify coping skills by changing negative self-talk and promoting coping behavior); and (d) *R*esults and rewards (reward youth for effort and teach him/her to self-reinforce). Therapeutic interventions such as graduated exposure tasks and role-playing are provided, and homework is regularly assigned to the youth. ICBT and YAS-ICBT therapists were both trained and supervised by experts in the *Coping Cat* program. YAS-UC therapists received no training as part of the study and were instructed to provide therapy in the manner to which they were accustomed.

Youth participants. The 89 youth participants (51 ICBT, 17 YAS-ICBT; 21 YAS-UC) met the following criteria: (a) a minimum of two audible sessions; and (b) received treatment from a single therapist (vs. multiple therapists). Youth participants in the ICBT group were treated in an outpatient setting at a research clinic at a large university in the mid-Atlantic region of the United States (Kendall et al., 2008). Recruitment for this study occurred via community sources. The youth participants in the YAS-ICBT and YAS-UC groups were clinically referred and treated in community-based outpatient settings in a large metropolitan area in southern California (see Southam-Gerow et al., 2010). Table 1 summarizes demographic and clinical data for these participants. Client participants were blind to treatment condition.

Therapist participants. There were 45 therapist participants (16 ICBT, 13 YAS-ICBT, 16 YAS-UC; 55.60% Caucasian, 11.11% did not report ethnicity; 13.33% male, 8.89% did not report sex). Therapists in the Kendall, Hudson, Gosch, Flannery-Schroeder, and Suveg (2008) study (N = 16; 12.50% male) were 81.25% Caucasian, 6.25% Latino, and 6.25% Asian/Pacific Islander (6.25% did not report). In YAS, therapists were clinic employees (N = 29) who volunteered to participate and were randomly assigned to groups. Therapists assigned to YAS-ICBT (N = 13; 15.38% male) were 53.85% Caucasian, 15.38% Latino, 15.38% Asian/Pacific Islander, and 15.38% mixed/other. Therapists assigned to YAS-UC (N = 16; 12.50% male, 25.00% did not report sex) were 31.25% Caucasian, 37.50% Latino, 6.25% and Asian/Pacific Islander, 25.00% did not report ethnicity. Therapist participants were not blind to treatment condition.

Adherence Measure Development Steps

Overview and preliminary steps. The development of the CBT for Anxiety in Youth Adherence Scale (CBAY-A) was

Table 1
Client Descriptive Data and Comparisons Across Groups

	M (SD) or %			
Variable	ICBT	YAS-ICBT	YAS-UC	F or Chi Square
Age	10.36 (1.90)	11.32 (2.32)	10.44 (1.91)	1.56
Male	60.80	29.40	52.40	5.04
Ethnicity				29.91[*]
Caucasian	86.30[a]	41.20	33.30	
African American	9.80	—	9.50	
Latino	2.00	17.60[d]	42.90[e]	
Mixed/other	2.00	5.90	9.50	
Not reported	—	35.30[b]	4.80	
CBCL				
Total	63.18 (8.44)	64.19 (7.34)	65.06 (6.46)	0.39
Internalizing	67.40 (8.37)	66.38 (8.33)	66.82 (8.33)	0.10
Externalizing	52.96 (10.08)	60.81 (7.49)[d]	59.41 (9.67)	5.61[*]
Primary diagnoses				22.73[*]
GAD	37.30[c]	5.90	14.30	
SAD	29.40	35.30	38.10	
SOP	33.30	23.50	28.60	
SP	—	35.30[b]	19.00	
Family income				15.66[*]
0 to $60K per year	35.30	70.60[d]	76.20[e]	
Number of sessions	15.92 (1.43)	16.82 (5.02)	15.71 (9.34)	0.26
Weeks in treatment	19.52 (3.97)	26.38 (10.41)[d]	26.84 (15.53)[e]	6.45[*]

Note. ICBT = individual cognitive–behavioral therapy (ICBT) delivered in Kendall et al. (2008) study; YAS-ICBT = ICBT delivered in YAS; YAS-UC = usual care delivered in YAS. For continuous variables, an ANOVA was conducted. For categorical variables, chi square analyses were conducted. CBCL = Child Behavior Checklist; GAD = generalized anxiety disorder; SAD = separation anxiety disorder; SOP = social phobia; SP = specific phobia.
[a] = ICBT > YAS-ICBT, YAS-UC. [b] = YAS-ICBT > YAS-UC. [c] = ICBT > YAS-ICBT. [d] = YAS-ICBT > ICBT. [e] = YAS-UC > ICBT.
[*] $p < .01$.

modeled after exemplar observer-rated treatment integrity measures, such as the *Therapist Behavior Rating Scale* (Hogue, Rowe, Liddle, & Turner, 1994) and *Rater's Manual for Yale Adherence and Competence Rating Scale* (Sifry et al., 1994). For the CBAY-A, we used the following sequence: (a) item generation and refinement; (b) scoring strategy, wherein a scoring strategy was determined for the items; and (c) scoring manual development and pilot coding, wherein a draft of the scoring manual was produced and refined via pilot coding.

Item generation and refinement. Our primary goal was to develop an instrument to measure adherence to ICBT for youth anxiety, rather than adherence to a specific ICBT treatment manual. First, we developed item categories conceptually and in consultation with the developers of ICBT approaches. The first category was *standard items*. These items were prescribed interventions that were standard to many CBT programs (i.e., not unique to ICBT for youth anxiety) and were expected to occur across many sessions. Examples of such items include *homework review*, an item reflecting efforts by the therapist to discuss a therapy homework assignment the client has completed, and *rapport building*, an item reflecting therapist efforts to develop a positive relationship with the client, often through informal conversation (e.g., discussing favorite meals or recent vacations) and games. The second category was *model items*, interventions specific to ICBT for youth anxiety and expected to be the focus of one or more sessions. Examples include *cognitive*, interventions designed to help a client develop skill in identifying and modifying anxiety-provoking thoughts; and *exposure*, interventions designed to facilitate the client's

engagement with safe but anxiety-provoking stimuli, with the goal being habituation/extinction. We identified a third set of items called *delivery items*; these items represent *how* specific model items were delivered. Examples include *didactic teaching* (i.e., teaching through direct instruction or explanation) and *rehearsal* (i.e., teaching via encouraging the client practice the skill[s] being taught in staged or actual situations).

Item development drew from three sources. First, we used Chorpita et al.'s (2011) review of youth evidence-based treatment to identify common practice elements (i.e., "discrete clinical technique or strategy used as part of a larger intervention plan used in CBT for youth anxiety"; Chorpita & Daleiden, 2009, p. 569). The Chorpita et al. review distilled the ingredients (i.e., interventions) of various evidence-based treatments for youth anxiety and identified the most common practice elements. We used the list of ICBT practice elements for youth anxiety as an initial pool of items. Second, we included all prescribed content from the *Coping Cat* manual (Kendall & Hedtke, 2006) and the Modular Approach to Therapy for Children[1] manual (MATCH; Chorpita & Weisz, 2009). *Coping Cat* is one of the first CBT programs for child anxiety developed and was the basis for many of the subsequent treatment programs studied since the first study (Kendall, 1994). MATCH represents similar items derived from either program

---

[1] We included MATCH because our project involves (later) testing our measures on recordings of therapists using MATCH.

were combined into a single item. Finally, experts in ICBT for youth anxiety, including the developers for *Coping Cat* and MATCH, reviewed the items and had the opportunity to generate additional items. The resulting measure had 22 items: four standard, 12 model, and six delivery. Scale development and scoring are described later. The full set of items, along with brief descriptions, appears in Table 2.

Scoring strategy. Extensiveness ratings, a widely used approach (e.g., Carroll et al., 2000; Hogue, Henderson et al., 2008), are used to measure the degree to which therapists use each intervention during a session. In making extensiveness ratings, coders estimate the extent to which a therapist engages in each intervention during the entire session using a 7-point Likert-type scale with the following anchors: 1 = *not at all*, 4 = *considerably*, 7 = *extensively*. Extensiveness ratings are comprised of two components: thoroughness and frequency. Thoroughness refers to the depth, complexity, or persistence with which the therapist engages in a given intervention whereas frequency refers to how often a therapist uses an intervention during a session (see Hogue et al., 1996). Both thoroughness and frequency are considered in making a rating; therefore, extensiveness ratings provide quantity, or dosage, information about each intervention.

Scoring manual. Following the adoption of a scoring strategy a full draft of the scoring manual was produced that detailed how to recognize each item, provided exemplars, and described item distinctions. Two coders used the scoring manual to pilot code ICBT and UC therapy sessions to help refine the manual. At the end of piloting, edits were made and a final version of the scoring manual was produced.

Measures for Validity Analyses

Therapy Process Observational Coding System for Child Psychotherapy - Revised Strategies scale (TPOCS-RS; McLeod et al., 2015). The TPOCS-RS (42 items) consists of five subscales: cognitive (four items; e.g., cognitive distortions); behavioral (nine items; e.g., operant interventions); psychodynamic (five items; e.g., interpretation); family (seven items; e.g., parenting intervention); and client-centered (four items; e.g., positive regard). In addition, there are 13 items (e.g., homework, play therapy) that represent therapeutic interventions that play a meaningful role in therapy but are not associated with a specific subscale. Coders rate the extent to which the therapist engages in each item during an entire session using a 7-point Likert-type extensiveness scale with the following anchors: 1 = *not at all*, 3 = *somewhat*, 5 = *considerably*, 7 = *extensively*. The TPOCS-RS item scores have demonstrated promising reliability and validity in past studies (e.g., McLeod & Weisz, 2010; McLeod et al., 2015). The mean interrater reliability, ICC(2,2), for the TPOCS-RS items in this study was .76 (*SD* = .18).

Therapy Process Observational Coding System for Child Psychotherapy-Alliance scale (TPOCS-A; McLeod & Weisz, 2005). The TPOCS-A consists of six items that assess affective aspects of the client–therapist relationship, and three items that assess client participation in therapeutic activities. Coders observe entire sessions and rate items on a 6-point scale ranging from 0 = *not at all* to 5 = *a great deal*. The TPOCS-A item scores have demonstrated mean item interrater reliability ranging from .48 to .80 (*M* ICC = .67), internal consistency ranging from .91 to .95 (*M*

α = .92), convergent validity with self-report alliance measures ranging from .48–.53 (e.g., McLeod & Weisz, 2005), and predictive validity with outcomes (e.g., Liber et al., 2010; McLeod & Weisz, 2005). Interrater reliability, ICC(2,2), for the TPOCS-A scale score in the present sample was .82; internal consistency was α = .81.

Coding and Session Sampling Procedures

Coders. Two doctoral students in clinical psychology with training and experience in ICBT for youth anxiety (one Latina female and one Caucasian male) coded the CBAY-A and two female doctoral students in clinical psychology (one Asian American and one Caucasian) coded the TPOCS-RS and TPOCS-A.

Coder training. Coder training involved three steps. First, coders reviewed the *Coping Cat* treatment manual, received didactic instruction and discussion of the scoring manual, participated in review sessions with the trainers, and engaged in coding exercises designed to test and expand understanding of each item. Next, coders engaged in independent coding and results were discussed in weekly meetings. Finally, coders independently coded 32 recordings. Reliability for each coder was assessed against master codes. For this article, interrater reliability was calculated using intraclass correlation coefficients (ICC; Shrout & Fleiss, 1979; see also e.g., Smith, Vannest, & Davis, 2011). We used model ICC(2,2) based on a two-way random effects model; this approach provides an estimate of the ratio of the true score variance to total variance. Thus, these correlations provide a reliability estimate of the mean scores of all coders considered as a whole, and allow for generalizability of the findings to other samples. As described by Cicchetti (1994), ICCs below .40 reflect "poor" agreement, ICCs from .40 to .59 reflect "fair" agreement, ICCs from .60 to .74 reflect "good" agreement, and ICCs .75 and higher reflect "excellent" agreement. The training period lasted 3 months for the two-coder team. Adequate reliability for both was achieved after the 3-week independent coding period (i.e., the 32 recordings). Once coders met "good" reliability on each item, ICC(2,2) ≥ .60 (Cicchetti, 1994), independent coding commenced.

Coding assignment plan. We sought to code every session except for the first and last as these sessions may contain intake or termination content. In addition, sessions were not rated if (a) missing or damaged, (b) shorter than 15 min, (c) less than 15 min was audible, or (e) less than 75% of the dialogue was in English. Of the 1,428 sessions, 954 (67%) met these criteria and were coded (532 or 66% from ICBT; 212 or 75% from YAS-ICBT; and 210 or 67% from YAS-UC). There were no significant differences across the three groups in the percent of sessions coded. Coding order was randomly determined. Each session was double-coded. Coders were naïve to study hypotheses and differences between data sources.

Data Analytic Strategy

In accord with common psychometric practice for integrity measures (e.g., Carroll et al., 2000; Hogue, Henderson et al., 2008), we targeted four goals: (a) item performance (i.e., descriptive statistics); (b) interrater reliability of the item scores; (c) scale scoring approach; and (d) validity of the score interpretations for the items and scales.

Table 2

Cognitive-Behavioral Therapy Adherence Scale for Youth Anxiety (CBAY-A): Descriptive Data and Reliability Results

Item type	Item	Brief description	Range	*M*	*SD*	ICC	Lower	Upper
Standard	Agenda setting	Therapist outlines a CBT plan or agenda related to child anxiety for the session.	5.5	2.07	1.18	0.62	0.57	0.67
Standard	Homework review	Therapist reviews CBT homework related to child anxiety from a past session.	6	2.67	1.70	0.85	0.83	0.87
Standard	Homework assignment	Therapist assigns CBT homework related to child anxiety to client and/or caregiver.	6	2.44	1.61	0.81	0.79	0.83
Standard	Rapport building	Therapist engages with client to build a relationship to facilitate client participation in CBT activities.	5.5	1.33	0.94	0.80	0.78	0.83
Model	Psychoeducation-anxiety	Therapist presents information about anxiety (e.g., inter-relations of thoughts, feelings, and actions) and its treatment (e.g., rationale behind exposure).	4.5	1.35	0.67	0.49	0.42	0.55
Model	Emotion education	Therapist teaches about feelings (e.g., generating list of feelings, physiological responses associated with feelings), with an emphasis on anxiety, and/or encourages client to identify physical cues of feelings.	6	1.65	1.42	0.92	0.91	0.93
Model	Fear ladder	Therapist works with client to create an ordered list of feared stimuli.	6	1.55	1.07	0.78	0.75	0.81
Model	Relaxation	Therapist teaches about how relaxation can be used to manage anxiety and/or encourages rehearsal.	6	1.43	1.17	0.92	0.91	0.93
Model	Cognitive-anxiety	Therapist teaches about and/or encourages rehearsal of the role of thoughts in creating, maintaining, and reducing anxiety.	6	1.52	1.24	0.86	0.84	0.88
Model	Problem solving	Therapist teaches about and/or encourages rehearsal of a multistep problem-solving model for coping with anxiety.	6	1.17	0.79	0.93	0.92	0.94
Model	Self-reward	Therapist teaches about and/or encourages rehearsal of evaluating and rewarding oneself for efforts to cope with anxiety.	5.5	1.26	0.92	0.91	0.90	0.92
Model	Coping plan	Therapist describes a multistep coping plan that involves the combination of more than one distinct anxiety management skill (e.g., relaxation and cognitive skills).	5.5	2.25	1.48	0.71	0.67	0.75
Model	Exposure preparation	Therapist prepares client for an exposure task.	6	1.84	1.6	0.93	0.92	0.94
Model	Exposure	Therapist encourages client participation in one or more exposure tasks.	6	1.58	1.34	0.91	0.90	0.92
Model	Exposure debrief	Therapist debriefs with client after exposure task(s) (e.g., praises client, discusses outcome).	6	1.48	1.06	0.86	0.84	0.88
Model	Maintenance	Therapist describes how the skills learned in CBT for child anxiety can be applied to future challenges related to coping with anxiety once therapy has completed.	3.5	1.05	0.29	0.52	0.46	0.58
Delivery	Didactic teaching	Therapist teaches CBT material related to child anxiety in a didactic and verbal manner.	5.5	2.13	1.21	0.73	0.69	0.76
Delivery	Collaborative teaching	Therapist teaches CBT material related to child anxiety in a collaborative manner.	6	2.96	1.67	0.69	0.65	0.73
Delivery	Modeling	Therapist teaches specific CBT skills related to coping with child anxiety using observational learning methods.	6	1.78	1.21	0.74	0.71	0.77
Delivery	Rehearsal	Therapist encourages client to participate in behavioral enactments in order to practice a CBT skill related to coping with child anxiety.	6	3.23	2.06	0.89	0.88	0.90
Delivery	Coaching	Therapist directs or provides feedback to a client who is practicing a CBT skill related to coping with child anxiety.	3.5	1.08	0.33	0.43	0.36	0.50
Delivery	Self-disclosure	Therapist shares information about his/her personal life, feelings, and/or experiences in order to teach an element of the CBT for youth anxiety model.	6	1.32	0.74	0.71	0.67	0.74
Model scale	Skills phase		5.5	3.46	1.95	N/A	N/A	N/A
Model scale	Exposure phase		6	2.75	1.90	N/A	N/A	N/A
Model scale	Total model		6	3.86	2.05	N/A	N/A	N/A

Note. CBT = cognitive–behavioral therapy.

Interrater reliability. Initial steps involved examining descriptive statistics of the items to ensure that the items functioned as designed (e.g., displayed adequate range) and evaluating interrater reliability for item scores. We hypothesized that scores for each item from the CBAY-A would demonstrate at least good interrater reliability, ICC(2,2) ≥ .60 (Cicchetti, 1994). Interrater reliability was calculated using ICCs, as described earlier.

Preliminary scale development. Because model items represent core technical elements of ICBT for youth anxiety and have been the focus of past efforts to measure adherence, we focused scale development efforts on those items. Guided by the structure of ICBT programs, we developed three scale scores from the model items, one total model scale, and two scales: skills phase and exposure phase. These latter scales represent the two phases of the *Coping Cat* program (Kendall & Hedtke, 2006) and are common phases in many ICBT approaches for anxiety. For the skills phase scale, the following items were included: psychoeducation, emotion education, fear ladder, relaxation, cognitive-anxiety, problem solving, self-reward, and coping plan. For the exposure phase scale, we included: coping plan, exposure prep, exposure, and exposure debrief. Coping plan was included in both scales as it is prescribed in both treatment phases; the maintenance item was not included in either scale.

Scale scores were generated as follows. For each recording, the item with the highest extensiveness score from each scale was used as the scale score. For example, the skill phase scale score for each recording was produced by taking the highest scoring item from the items included in skill phase scale. The same process was used for the exposure phase and total model scales. We retained the highest score for each recording because in each prescribed *Coping Cat* (and many other ICBT programs) session, there is a focus on one model item (e.g., exposure, cognitive-anxiety).

Construct validity: CBAY-A item scores. We evaluated the discriminant validity of the CBAY-A item scores. The discriminant validity of the CBAY-A item scores was assessed by examining the magnitude of the correlations among the scores on the model items. Because the model items were designed to measure distinct aspects of ICBT for youth anxiety, we hypothesized that the correlations among the item scores would be small to medium in strength (Cohen, 1992; cf. Hogue, Dauber, et al., 2008). Correlations were interpreted following Cohen's (1992) guidelines: r is a "small" effect if 0.10–0.23; "medium" if 0.24–0.36; and "large" if > 0.36.

We assessed the convergent validity of the CBAY-A item scores with scores on an observational measure that measured cognitive and behavioral interventions (TPOCS-RS). Because these correlations were between scores on items that were designed to measure the same therapeutic content we hypothesized that the correlations would be large.

Construct validity: CBAY-A scale scores. We evaluated the convergent and discriminant validity of the CBAY-A scale scores. Analyses assessed the magnitude of the correlations between the CBAY-A scale scores and subscale scores on an observational measure of therapeutic interventions (TPOCS-RS). As the CBAY-A is designed to assess adherence to an ICBT program we hypothesized that the scale scores would evince (a) strong correlations with treatment approaches prescribed by ICBT represented by the TPOCS-RS cognitive and behavioral subscale scores, (b) zero or negative correlations with treatment approaches proscribed by ICBT represented by the TPOCS-RS family and psychodynamic subscale scores, and (c) zero to small correlations with the TPOCS-RS client-centered subscale scores.

We also assessed the discriminant validity of the CBAY-A scale scores by evaluating the magnitude of the correlations between the scale scores and scores on an observational alliance measure (TPOCS-A). Given that the CBAY-A scale scores and TPOCS-A scores are intended to represent discrete but related therapy processes, we hypothesized that the correlations would be small to medium (Carroll et al., 2000; Hogue, Dauber et al., 2008).

Discriminant validity: CBAY-A item and scale scores. We examined discriminant validity by evaluating whether the CBAY-A item and scale scores could detect expected differences between ICBT and UC. We computed adjusted least square means (LSMs) scores using SAS/STAT Software 9.4 to account for the nested design of these data (cf. Barber, Foltz, Crits-Christoph, & Chittams, 2004). The LSMs are scores that account for the influence of other variables. To produce the LSMs, we used a mixed model with restricted maximum likelihood estimation for the following random factors: (a) study group (i.e., ICBT, YAS-ICBT, and YAS-UC); (b) therapist (nested within study group); (c) client (nested within study group, therapist); (d) time (nested within client, therapist, study group); and (e) coder. Each factor represents a possible source of variation in CBAY-A item and scale scores (Barber et al., 2004). The term *study group* reflects the influence of the three groups (ICBT, YAS-ICBT, YAS-UC) on each CBAY-A item and scale score; the term *therapist* represents systematic differences across therapists on each CBAY-A item/scale score; the term *client* reflects systematic differences in CBAY-A item/scale scores across each client; the term *time* reflects the effect time in treatment (measured in weeks since the intake) has on each CBAY-A item and scale score variance; the term *coder* reflects systematic differences in coder ratings (tendency to score high or low) on a given CBAY-A item/scale. Because our primary interest was to conduct group comparisons among the treatment groups on the item and scale scores, we examined the overall F test for *study group*. All significant effects were followed up with pairwise comparisons of the adjusted means for each item or scale score. Because *Coping Cat* is an ICBT program, we hypothesized that the ICBT groups (ICBT, YAS-ICBT) would have higher scores than YAS-UC on the CBAY-A item and scale scores. We also anticipated that ICBT would be higher than the YAS-ICBT because the ICBT therapists were supervised by the program developer in an efficacy trial.

Results

Reliability Analyses: Interrater Reliability

We inspected CBAY-A items with the expectation that the range of each item would come close to the full possible range (i.e., 6 points or 1 to 7). Table 2 reflects the descriptive statistics for all items, along with the skills phase, exposure phase, and total model scales. Only two items had a range below 4 points, maintenance and coaching. Next, we examined interrater reliability for the CBAY-A item scores. We also inspected the item and scale distributions. As anticipated, items were positively skewed and those with the highest skew were the items with the lowest mean scores, smallest ranges, and in two of the three, the lowest reli-

ability coefficients (i.e., problem solving, maintenance, and coaching). The three CBAY-A scale scores, however, were not skewed and appeared to be normally distributed.

Table 2 also summarizes the ICC results, along with 95% confidence intervals, for all items. On the whole, the ICCs suggest good to excellent reliability for most items; tight confidence intervals strengthen this conclusion. Fully 19 of the 22 items were in the "good" or better range using the standards described by Cicchetti (1994), with 13 of the 22 in the "excellent" range. Only three items, psychoeducation-anxiety, maintenance, and coaching, had ICC values below 0.60 and none were below 0.50. Two of the items with lower ICC values (maintenance and coaching) also exhibited limited range.

Construct Validity: CBAY-A Item Scores

We first examined correlations among the model item scores. As can be seen in Table 3, there were positive relationships among most item scores; though the items are designed to represent distinct aspects of the ICBT model content is presented in an integrated manner. The strongest correlations were observed among scores for the items related to the exposure phase (exposure prep, exposure, and exposure debrief). These coefficients ranged from 0.71 to 0.83, suggesting the need to consider redundancy among these items. However, these three items reflect distinct parts of the process of conducting an exposure. An argument for retaining these three as separate items despite the high intercorrelations can be captured in the following scenario. A therapist may extensively prepare a client for an exposure situation (i.e., high score on exposure prep item) and then not follow through on the exposure because the client balks (i.e., low score on exposure item). Although the correlation results do not suggest this was a common occurrence, having separate scores for the three could help in the therapist training (e.g., emphasizing the importance of following through on all elements of exposure delivery). Aside from the exposure items, the mean of the absolute value of the correlations among the item scores was small ($M = .12$, $SD = 0.10$; range 0.00 to 0.44), suggesting the item scores capture distinct aspects of ICBT for youth anxiety.

We also examined correlations between the CBAY-A item scores and the corresponding TPOCS-RS cognitive and behavioral item scores. Our specific hypothesized relationships are summarized in Table 4, with correlations shown. The mean of the absolute value of the correlations was large and positive ($M = .57$, $SD = 0.21$; range 0.21 to 0.88), with only one (for psychoeducation) correlation lower than .24. Thus, CBAY-A.[2] item scores across the standard, model, and delivery categories demonstrated "large" correlations with similar item scores on the TPOCS-RS, providing convergent validity evidence for these items.

Construct Validity: CBAY-A Scale Scores

As can be seen in Table 5, the correlations between the CBAY-A scale scores and the TPOCS-RS subscale scores provided general support for our hypotheses. The mean of the absolute value of the correlations between the scores on the CBAY-A scales and the TPOCS-RS subscales followed the predicted pattern. For the cognitive and behavioral subscales, the correlations were large and positive ($M = .68$, $SD = .13$) whereas for the psychodynamic

and family subscales the correlations were medium and negative ($M = .39$, $SD = .08$), and for the TPOCS-RS client-centered subscale ($M = .03$, $SD = .02$), the correlations were near zero. Taken together, these findings support the convergent and discriminant validity of the CBAY-A scale scores.

We also assessed the magnitude of the correlations between the CBAY-A scale scores and scores on an observational alliance measure (TPOCS-A). As can be seen in Table 5, the mean scores of the absolute value of the correlations between the CBAY-A scale scores and the TPOCS-A scores ($M = .36$, $SD = .07$; range .26 to .45) were medium and positive, further supporting the discriminant validity of the CBAY-A scale scores.

Discriminant Validity: CBAY-A Item and Scale Scores

As can be seen in Table 6, ICBT and YAS-ICBT had significantly higher least square mean scores than YAS-UC on almost every item and scale. Effect sizes (Cohen's d) for the ICBT versus YAS-UC item and scale score differences were all greater than 0.30 (except for maintenance) and averaged 0.96 (range 0.10 to 2.31), suggesting large differences. The three scale score effect sizes were all higher than 1.50, ranging from 1.76 to 2.31, reflecting very large differences. The YAS-ICBT versus YAS-UC differences in extensiveness scores were also in the expected direction, though smaller in magnitude, with an average effect size of 0.55 (range 0.08 to 1.52); differences between YAS-ICBT and YAS-UC for the scale scores were all large and ranged from 0.78 to 1.52. Effect sizes of the extensiveness scores differences between the ICBT and YAS-ICBT group generally favored the ICBT group, as expected ($M = 0.41$; range −0.14 to 0.99). Scale score differences were larger, ranging from 0.63 to 0.99.

Discussion

We developed an adherence measure for ICBT for youth anxiety, the CBAY-A, and reported initial psychometric properties. The instrument gauged standard items, treatment elements common across multiple CBT approaches, model items specific to ICBT for youth anxiety, and delivery items. We examined the psychometric performance of these items along with three scale scores: skills phase, exposure phase, and total model. Results were largely supportive of the reliability and validity of the item and scale scores. For instance, independent coders reliably rated extensiveness of delivery of a variety of ICBT interventions. The item and scale scores also demonstrated convergent validity, with medium to large correlations with similar measures. A similar pattern was found with measures of distinct constructs, supporting the discriminant validity of the item and scale scores. Finally, the findings suggest that scores on the model items and scales differentiate therapists providing ICBT for youth anxiety from those not doing so, supporting discriminant validity of those item and scale scores.

The results support the CBAY-A items on a critical indicator for an observational scale, interrater reliability. Across all three types of items rated, ICCs almost all exceeded 0.60, generally consistent

---

[2] Note that one standard (i.e., rapport building) and two delivery (i.e., didactic teaching, collaborative teaching) items did not have similar items on the TPOCS-RS and thus were not evaluated.

Table 3

Correlations Among Model Items of the Cognitive–Behavioral Therapy for Youth Anxiety-Adherence Scale (CBAY-A)

	1	2	3	4	5	6	7	8	9	10	11
1. Psychoeducation-anxiety[a]											
2. Emotion education[a]	**0.44**										
3. Fear ladder[a]	*0.23*	*0.27*									
4. Relaxation[a]	0.07	0.08	−0.03								
5. Cognitive-anxiety[a]	0.06	−0.01	−0.05	<u>0.16</u>							
6. Problem solving[a]	−0.06	−0.04	−0.02	0.09	0.22						
7. Self-reward[a]	−0.01	−0.09	0.22	0.05	0.01	0.07					
8. Coping plan[a,b]	0.06	<u>−0.16</u>	0.18	0.01	0.16	*0.27*	*0.35*				
9. Exposure preparation[b]	−0.13	<u>−0.21</u>	0.03	<u>−0.14</u>	−0.10	−0.10	<u>−0.11</u>	*0.36*			
10. Exposure[b]	−0.12	<u>−0.18</u>	0.00	<u>−0.11</u>	−0.06	−0.07	<u>−0.10</u>	<u>0.25</u>	**0.71**		
11. Exposure debrief[b]	−0.09	<u>−0.18</u>	0.02	−0.12	−0.08	−0.08	<u>−0.10</u>	*0.33*	**0.83**	**0.75**	
12. Maintenance	0.00	−0.06	−0.06	−0.04	−0.04	−0.03	−0.04	0.08	0.09	<u>0.15</u>	<u>0.12</u>

Note. Bolded correlations represents values in the "large" range, using Cohen's (1992) standards; *italicized* correlations are in the "medium" range; underlined correlations were in the "small" range.
[a] These items comprise the skills phase scale. [b] These items comprise the exposure phase scale.

with findings from other observational adherence measures (e.g., Barber, Liese, & Abrams, 2003; Hogue, Dauber, et al., 2008). Further, nearly all of the items exhibited the expected full range of scores, with all but three items showing a range of at least 5.5 points (maximum range was 6). Overall, the items can be coded reliably by trained coders and capture a range of adherence-related therapist behaviors.

Three of the items demonstrated ICCs below 0.60 and more restricted ranges (3.5 to 4.5 points). Two of these were model

Table 4

Convergent Validity Correlations of the Cognitive–Behavioral Therapy for Youth Anxiety-Adherence Scale (CBAY-A) Items and Scales

CBAY-A item type	CBAY-A item	TPOCS-RS items/scales	*r*
Standard	Agenda setting	Session goals item	**0.38**
Standard	HW review	HW item	**0.76**
Standard	HW assignment	HW item	**0.72**
Model	Psychoeducation-anxiety	Cognitive education item	**0.37**
Model	Psychoeducation-anxiety	Psychoeducation item	*0.21*
Model	Emotion education	Cognitive education item	**0.54**
Model	Relaxation	Relaxation item	**0.88**
Model	Relaxation	Behavioral scale	**0.32**
Model	Cognitive-anxiety	Cognitive education item	**0.41**
Model	Cognitive-anxiety	Cognitive distortion item	**0.56**
Model	Cognitive-anxiety	Cognitive scale	**0.50**
Model	Problem solving	Coping skills item	*0.36*
Model	Self-reward	Operant item	**0.57**
Model	Coping plan	Coping skills item	**0.75**
Model	Exposure preparation	Respondent item	**0.83**
Model	Exposure	Respondent item	**0.72**
Model	Exposure debrief	Respondent item	**0.75**
Delivery	Modeling	Modeling item	**0.68**
Delivery	Rehearsal	Rehearsal item	**0.86**
Delivery	Coaching	Coaching item	*0.26*
Delivery	Self-disclosure	Self-disclosure item	**0.49**

Note. Bolded correlations represents values in the "large" range, using Cohen's (1992) standards; *italicized* correlations are in the "medium" range. TPOCS-RS = Therapy Process Observational Coding System for Child Psychotherapy—Revised Scale.

items (i.e., items that gauge specifically prescribed ingredients of the treatment program): psychoeducation (e.g., therapist provides information about anxiety) and maintenance (e.g., therapist reflects on most useful aspects of the treatment); and one was a delivery item (i.e., items that gauge how a therapist delivers a specific model item): coaching (e.g., therapist provides feedback to the client related to her/his practice of a specific skill). For all three items, lower reliability may be due to low variation in the scores, as these three items are not found in the *Coping Cat* program (Kendall & Hedtke, 2006). Our long-term aim was to develop a generic adherence measure for ICBT for youth anxiety; thus, we included items not key in *Coping Cat*. Ultimately, however, the ICCs for psychoeducation and maintenance items, though below 0.60, were not below 0.50, suggesting they could be refined in future work.

With regard to convergent and discriminant validity, standard, model, and delivery item scores all correlated highly with related TPOCS-RS item and subscale scores, results on par with past integrity studies (e.g., Carroll et al., 2000). The model item, psychoeducation, performed least well; correlations with similar TPOCS-RS items ranged from 0.21–0.37, highlighting the need to evaluate that item in future work. Further, the model items and scales demonstrated discriminant validity from subscales on the TPOCS-RS capturing incompatible/unrelated treatment approaches (e.g., psychodynamic, client-centered). Similar discriminant validity evidence was found for the CBAY-A item and scale scores with a measure of alliance.

Analyses yielded group differences between therapists delivering ICBT and those not. Differences were found between two groups of therapists delivering ICBT, one group in an efficacy and the other in an effectiveness trial. Although similar validity evidence has been reported (e.g., Barber et al., 2004), ours is the first to show differences in adherence level between therapists in efficacy and effectiveness trials using the same treatment program. These results support the discriminant validity of the CBAY-A item and scale scores.

The performance of the three model scales was encouraging, as these represent potential composite scores to use in future work (e.g., adherence-outcome relationships; McLeod, Southam-Gerow,

Table 5

Validity-Related Correlations for Cognitive–Behavioral Therapy for Youth Anxiety-Adherence Scale (CBAY-A) Scales

	1	2	3	4	5	6	7	8	9
1. CBAY-A skills		0.22	**0.82**	**0.73**	**0.43**	**−0.38**	**−0.42**	−0.05	**0.40**
2. CBAY-A exposure			**0.68**	**0.77**	**0.70**	*−0.29*	*−0.29*	0.02	*0.30*
3. CBAY-A total model				**0.78**	**0.65**	−.043	**−0.50**	−0.03	**0.45**
4. TPOCS-RS Cognitive					**0.59**	*−0.32*	**−0.43**	*0.34*	**0.41**
5. TPOCS-RS behavioral						*−0.29*	*−0.33*	*0.29*	*0.31*
6. TPOCS-RS psychodynamic							0.19	0.02	**−0.43**
7. TPOCS-RS family								−0.04	*−0.29*
8. TPOCS-RS client-centered									*0.26*
9. TPOCS-A alliance									

Note. Bolded correlations represents values in the "large" range, using Cohen's (1992) standards; *italicized* correlations are in the "medium" range; underlined correlations were in the "small" range. TPOCS-RS = Therapy Process Observational Coding System for Child Psychotherapy—Revised Scale; TPOCS-A = Therapy Process Observational Coding System for Child Psychotherapy—Alliance Scale.

Tully, Rodriguez, & Smith, 2013). Our findings suggest a separation of between .94 to 1.55 adherence points (out of a maximum of 6 possible) between the efficacy and effectiveness study ICBT therapists. Further, our data suggest that within the effectiveness study, ICBT therapists were between 1.17 to 2.32 adherence points higher than the UC therapists. Whether the increase from the effectiveness dose to the efficacy dose is related to improvement in outcomes represents an important future direction.

These findings were consistent with predictions and suggest that the CBAY-A has promise as an observational measure of adherence. As such, the study is an initial step in establishing the reliability and representational validity of the CBAY-A item and scale scores (Foster & Cone, 1995). Specifically, preliminary evidence supports representational validity by indicating that the CBAY-A item and scale scores assess ICBT for youth anxiety (i.e., what the measure "is") and can discriminate between ICBT and other forms of therapy (i.e., what the measure is not). These initial steps provide important preliminary data. Additional psychometric work is needed to establish the elaborative validity (Foster & Cone, 1995) of the CBAY-A item and scale scores (e.g., scores can be used to predict outcomes, monitor treatment adherence).

As noted, adherence measurement is a critical, often overlooked aspect of clinical research with numerous important applications

Table 6

Discriminant Validity Results for Cognitive–Behavioral Therapy for Youth Anxiety-Adherence Scale (CBAY-A)

Item type	Item	Adj. Mean			Cohen's *d*		
		ICBT	*YAS-ICBT*	*YAS-UC*	*ICBT vs. YAS-UC*	*YAS-ICBT vs. YAS-UC*	*ICBT vs. YAS-ICBT*
Standard	Agenda setting	2.51	1.93	1.10	1.19	.71	.49
	HW review	3.32	2.69	1.02	1.35	.98	.37
	HW assignment	3.08	2.16	1.09	1.24	.66	.57
	Rapport building	1.42	1.39	1.04	.40	.37	.03
Model	Psychoed-anxiety	1.50	1.27	1.07	.64	.31	.33
	Emotion education	1.83	1.79	1.06	.54	.51	.03
	Fear ladder	1.83	1.35	1.01	.77	.32	.45
	Relaxation	1.59	1.40	1.05	.46	.30	.16
	Cognitive	1.71	1.53	1.05	.53	.39	.14
	Problem solving	1.27	1.06	1.00	.34	.08	.26
	Self-reward	1.36	1.27	1.01	.39	.29	.10
	Coping plan	2.81	2.06	1.01	1.22	.71	.51
	Exposure prep	2.41	1.22	1.00	.88	.14	.74
	Exposure	1.97	1.17	1.00	.72	.13	.59
	Exposure debrief	1.81	1.13	1.00	.76	.12	.64
	Maintenance	1.05	1.09	1.02	.10	.24	−.14
Delivery	Didactic teaching	2.57	2.02	1.12	1.20	.74	.45
	Collaborative teaching	3.72	2.89	1.11	1.56	1.06	.50
	Modeling	2.10	1.70	1.04	.88	.54	.33
	Rehearsal	4.28	2.73	1.06	1.56	.81	.75
	Coaching	1.13	1.06	1.00	.39	.17	.21
	Self-disclosure	1.43	1.37	1.01	.57	.48	.09
Scale	Skills	4.37	3.43	1.18	2.15	1.52	.63
	Exposure	3.66	2.18	1.01	1.76	.78	.99
	Total	5.06	3.51	1.19	2.31	1.39	.93

Note. ICBT = individual cognitive–behavioral therapy delivered in Kendall et al. (2008) study; *YAS-ICBT* = ICBT delivered in YAS; *YAS-UC* = usual care delivered in YAS. All tests had *df* = 1,899.

(Perepletchikova et al., 2007). Our report involves preliminary data on the initial development of such a measure. Thus, the applications discussed next may be best considered after further measure development. A potential application for an adherence measure is to gauge the extent to which the independent variable in an RCT was delivered as intended. For this purpose, the CBAY-A would differ from most past adherence measures insofar as the item and scale scores reflect the relative dose of specific aspects of ICBT rather than a percentage of content delivered "adequately." As one example, our results suggest that in an efficacy trial, the average total model scale scores were just over 5 points (out of 7). Our data suggested that this score was significantly lower in the effectiveness trial. Relatedly, the new measure could be used to establish benchmarks for implementation studies, with research clinic adherence scores serving as a possible goal for effectiveness studies.

Another application would be to examine relations between adherence and client outcomes. This is already a focus of some studies (e.g., Schoenwald, Carter, Chapman, & Sheidow, 2008) and represents a critical validity component to the portfolio of the item and scale scores of an adherence instrument. From an application perspective, understanding adherence-outcome relations could inform therapist training (e.g., training to a criterion adherence level that produces a desired outcome). Further, if adherence is related to outcome, adherence scores could be used to gauge service quality in mental health agencies or systems, consistent with calls for increasing accountability in health service delivery (Garland et al., 2013; Pincus, Spaeth-Rublee, & Watkins, 2011), and with the advent of initiatives like pay for performance (e.g., Campbell, Reeves, Kontopantelis, Sibbald, & Roland, 2009). Demonstrating the representational validity of the item and scale scores lays the groundwork for these various applications of treatment integrity measurement, though the measure needs to be more fully developed before such applications are considered.

Potential limitations merit attention. First, although we coded every available session for all clients participating in the two trials, there were some recordings we were unable to code. As a result, judging therapist adherence for each client is limited by missing data. Related to this, ICBT for youth anxiety involves exposure tasks, some of which occur outside of the therapy room. As a result, it is possible that the recordings we coded may have undersampled delivery of exposure and thus may underestimate of the extensiveness of delivery of exposure. Another limitation is the lack of an alternative method of measuring adherence. We relied solely on trained observers' ratings. Other studies have included client and therapist ratings of adherence (e.g., Schoenwald et al., 2008), both of which are more efficient methods (Schoenwald et al., 2011). This study involved two groups of therapists delivering the same program. Because the measure was designed to capture ICBT for youth anxiety broadly, examining its performance with therapists delivering different ICBT for youth anxiety programs represents a next step. The current study had a relatively limited sample of coders and therapists, reducing our ability to gauge the effects of these facets on adherence coding. Future studies should include a broader array of clients, therapists, and coders. Finally, additional research is needed to further examine the psychometric properties of the measure.

Despite the limitations, the findings provide preliminary data supporting the CBAY-A as an observational measure of treatment adherence to ICBT for youth anxiety. The CBAT-A items can be coded reliably across three categories of items (standard, model, and delivery). Further the items performed as expected (i.e., have expected range values), suggesting that adherence can be measured on an interval (vs. nominal) scale. In addition, the model items and to a lesser extent the standard and delivery items demonstrated strong preliminary representational validity (e.g., convergent, discriminant; Foster & Cone, 1995). Finally, the three model item scales we developed possess a similarly positive psychometric profile.

References

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,* 29–41. http://dx.doi.org/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,* 205–221. http://dx.doi.org/10.1093/ptr/kpg019

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,* 1–30. http://dx.doi.org/10.1111/j.1468-2850.2009.01187.x

Campbell, S. M., Reeves, D., Kontopantelis, E., Sibbald, B., & Roland, M. (2009). Effects of pay for performance on the quality of primary care in England. *The New England Journal of Medicine, 361,* 368–378. http://dx.doi.org/10.1056/NEJMsa0807651

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,* 225–238. http://dx.doi.org/10.1016/S0376-8716(99)00049-6

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,* 566–579. http://dx.doi.org/10.1037/a0014565

Chorpita, B. F., Daleiden, E. L., Ebesutani, C., Young, J., Becker, K. D., Nakamura, B. J., . . . Starace, N. (2011). Evidence-based treatments for children and adolescents: An updated review of indicators of efficacy and effectiveness. *Clinical Psychology: Science and Practice, 18,* 154–172. http://dx.doi.org/10.1111/j.1468-2850.2011.01247.x

Chorpita, B. F., & Weisz, J. R. (2009). *Modular Approach to Therapy for Children with Anxiety, Depression, Traumatic Stress, and Conduct Problems (MATCH-ADTC).* Satellite Beach, FL: PracticeWise.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155

Comer, J., & Kendall, P. C. (2013). Methodology, design, and evaluation in psychotherapy research. In M. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change,* 6th ed. (pp. 21–48). New York, NY: Wiley.

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

Garland, A. F., Haine-Schlagel, R., Brookman-Frazee, L., Baker-Ericzen, M., Trask, E., & Fawley-King, K. (2013). Improving community-based mental health care for children: Translating knowledge into action. *Administration and Policy in Mental Health and Mental Health Services Research, 40,* 6–22. http://dx.doi.org/10.1007/s10488-012-0450-8

Garland, A. F., Hurlburt, M. S., Brookman-Frazee, L., Taylor, R. M., & Accurso, E. C. (2010). Methodological challenges of characterizing usual care psychotherapeutic practice. *Administration and Policy in Mental Health and Mental Health Services Research, 37,* 208–220. http://dx.doi.org/10.1007/s10488-009-0237-8

Gresham, F. M. (2009). Response to intervention in the identification of specific learning disabilities. In A. Akin-Little & S. Little (Eds.), *Handbook of behavioral interventions in schools* (pp. 205–220). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/11886-013

Hagermoser Sanetti, L. M., Fallon, L. M., & Collier-Meeka, M. A. (2011). Treatment integrity assessment and intervention by school-based personnel: Practical applications based on a preliminary study. *School Psychology Forum, 5,* 87–102.

Hagermoser Sanetti, L. M., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. *School Psychology Review, 38,* 445–459.

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,* 137–147. http://dx.doi.org/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,* 544–555. http://dx.doi.org/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,* 332–345. http://dx.doi.org/10.1037/0033-3204.33.2.332

Hogue, A., Rowe, C., Liddle, H., & Turner, R. (1994). *Scoring manual for the Therapist Behavior Rating Scale (TBRS).* Unpublished manuscript, Center for Research on Adolescent Drug Abuse, Temple University, Philadelphia, PA.

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

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

Kendall, P. C., Hudson, J. L., Gosch, E., Flannery-Schroeder, E., & Suveg, C. (2008). Cognitive-behavioral therapy for anxiety disordered youth: A randomized clinical trial evaluating child and family modalities. *Journal of Consulting and Clinical Psychology, 76,* 282–297. http://dx.doi.org/10.1037/0022-006X.76.2.282

Liber, J. M., McLeod, B. D., Van Widenfelt, B. M., Goedhart, A. W., van der Leeden, A. J. M., Utens, E. M. W. J., & Treffers, P. D. A. (2010). Examining the relation between the therapeutic alliance, treatment adherence, and outcome of cognitive behavioral therapy for children with anxiety disorders. *Behavior Therapy, 41,* 172–186. http://dx.doi.org/10.1016/j.beth.2009.02.003

McLeod, B. D., Islam, N. Y., & Wheat, E. (2013). Designing, conducting, and evaluating therapy process research. In J. Comer & P. Kendall (Eds.), *The Oxford handbook of research strategies for clinical psychology* (pp. 142–164). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/oxfordhb/9780199793549.013.0009

McLeod, B. D., Smith, M. M., Southam-Gerow, M. A., Weisz, J. R., & Kendall, P. C. (2015). Measuring differentiation for implementation research: The Therapy Process Observational Coding System for Child Psychotherapy Revised Strategies scale. *Psychological Assessment, 27,* 314–325. http://dx.doi.org/10.1037/pas0000037

McLeod, B. D., Southam-Gerow, M. A., Tully, C. B., Rodríguez, A., & Smith, M. M. (2013). Making a case for treatment integrity as a psychosocial treatment quality indicator for youth mental health care. *Clinical Psychology: Science and Practice, 20,* 14–32. http://dx.doi.org/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,* 323–333. http://dx.doi.org/10.1037/0022-006X.73.2.323

McLeod, B. D., & Weisz, J. R. (2010). The therapy process observational coding system for child psychotherapy-strategies scale. *Journal of Clinical Child and Adolescent Psychology, 39,* 436–443. http://dx.doi.org/10.1080/15374411003691750

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,* 829–841. http://dx.doi.org/10.1037/0022-006X.75.6.829

Pincus, H. A., Spaeth-Rublee, B., & Watkins, K. E. (2011). Analysis & commentary: The case for measuring quality in mental health and substance abuse care. *Health Affairs, 30,* 730–736. http://dx.doi.org/10.1377/hlthaff.2011.0268

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,* 379–394. http://dx.doi.org/10.1007/s10488-008-0181-z

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,* 32–43. http://dx.doi.org/10.1007/s10488-010-0321-0

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428. http://dx.doi.org/10.1037/0033-2909.86.2.420

Sifry, R. L., Nuro, K. F., Ball, S., Corvino, J., Bisighini, R. M., & Carroll, K. M. (1994). *Rater's manual for Yale psychotherapy development center treatment rating scale.* Unpublished manuscript, Yale University, New Haven, Connecticut.

Smith, S. L., Vannest, K. J., & Davis, J. L. (2011). Seven reliability indices for high-stakes decision making: Description, selection, and simple calculation. *Psychology in the Schools, 48,* 1064–1075. http://dx.doi.org/10.1002/pits.20610

Southam-Gerow, M. A., Weisz, J. R., Chu, B. C., McLeod, B. D., Gordis, E. B., & Connor-Smith, J. K. (2010). Does CBT for youth anxiety outperform usual care in community clinics?: An initial effectiveness test. *Journal of the American Academy of Child and Adolescent Psychiatry, 49,* 1043–1052. http://dx.doi.org/10.1016/j.jaac.2010.06.009