2017

The Development of the Treatment Integrity - Efficient Scale for Cognitive Behavioral Treatment for Youth Anxiety (TIES-CBT-YA)

Meghan Smith
Virginia Commonwealth University

THE DEVELOPMENT OF THE TREATMENT INTEGRITY – EFFICIENT SCALE FOR COGNITIVE-BEHAVIORAL TREATMENT FOR YOUTH ANXIETY (TIES-CBT-YA)

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

By: MEGHAN MARIE SMITH
Bachelor of Arts; Psychology, University of Virginia; May, 2010
Master of Science; Virginia Commonwealth University, Richmond, VA; December, 2013

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

Virginia Commonwealth University
Richmond, Virginia
May, 2017

Acknowledgement

I would like to thank the numerous people who have supported me during graduate school. Thank you, Dr. Bryce D. McLeod, for working with me closely to develop and polish this dissertation; I appreciate your guidance and inspiration, not only in this project, but over the last six years within research and clinical domains. Thank you, Dr. Michael Southam-Gerow, for your mentorship and encouragement within the TIMS lab; I feel lucky to have had so many opportunities to collaborate with you. I would also like to thank the coders who spent a year training and coding for this study; thank you for being so dedicated to mastering the coding, handling recording issues, and scheduling precious time each week for this project. Finally, I would like to thank my parents, brothers, and graduate student colleagues; I appreciate your unwavering support, encouragement, and love throughout my training.

Table of Contents

Page

List of Tables

List of Figures

Page

Abstract

THE DEVELOPMENT OF THE TREATMENT INTEGRITY – EFFICIENT SCALE FOR COGNITIVE-BEHAVIORAL TREATMENT FOR YOUTH ANXIETY (TIES-CBT-YA)

By Meghan M. Smith, M.S.

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

Virginia Commonwealth University, 2017

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

Brief, easy to use, psychometrically strong (i.e., pragmatic) instruments are needed to support implementation research; the current study assessed whether it was possible to develop a pragmatic observational treatment integrity instrument and reduce the amount of time coders spend making treatment integrity ratings (while maintaining score validity) of therapists delivering two protocols of individual cognitive-behavioral treatment (ICBT) for youth anxiety in research and practice settings. The 12-item instrument was derived from four observational treatment integrity instruments with promising score reliability and validity that assess adherence, competence, differentiation, and alliance. A sample of 106 youths (*M* age = 10.12, *SD* = 1.81, ages 7-14; 42.50% Female; 69.80% Caucasian) received one of three treatments to address anxiety: standard ICBT in a research setting (*n* = 51) or standard ICBT (*n* = 22), modular

ICBT (*n* = 16), or usual care (UC; *n* = 17) in practice settings. Four coders independently coded five- and 15- minute segments sampled from four sessions from each client (*N* = 756 sessions). Ten percent of sessions were double-coded for reliability purposes. Reliability, sensitivity to change, construct validity, and predictive validity from the two segments were compared to full session treatment integrity scores independently archived in a study assessing the same clients. Across five- and 15-minute segments, the instrument produced promising score reliability and convergent validity evidence for adherence, competence, and alliance items (items intended for inclusion in ICBT for youth anxiety; *M* ICCs = .62, *SD* = .17; *M* *r*s = .58, *SD* = .12) and poor score reliability and validity evidence for differentiation items (items intended for inclusion in other treatment domains; *M* ICCs = .21, *SD* = .28; *M* *r*s = .27, *SD* = .25). This study met its primary aim, to develop an instrument that can be coded in less than 20 minutes while maintaining evidence of score validity. Researchers interested in developing such instruments can use this study design as a roadmap. Future research should investigate whether psychometric findings replicate across samples, why certain items (e.g., client-centered interventions) did not evidence score validity, and how this type of instrument can inform EBT training.

The Development of the Treatment Integrity – Efficient Scale for Cognitive Behavioral Treatment for Youth Anxiety (TIES-CBT-YA)

Researchers and practitioners have suggested that efficient (i.e., brief, easy to administer and use), yet effective (i.e., valid, accurate) instruments are needed to support implementation research (Glasgow & Riley, 2013). Currently, implementation research is hindered by the measurement of key constructs with inefficient (i.e., lengthy, difficult to administer and use) instruments that are not suited for use in typical practice contexts (Garland & Schoenwald, 2013). Thus, though implementation studies pledge to improve the quality of care in community settings, the current instrumentation used does not always contribute to this goal. Implementation researchers could use more efficient instruments to enhance implementation efforts and make quality improvements. Some have therefore argued for the development of *pragmatic instruments,* or instruments that are efficient, have strong validity evidence for use in practice settings (i.e., community clinics where most children receive psychosocial treatment), and are sensitive to change (Glasgow & Riley, 2013).

One key construct in implementation research is treatment integrity (McLeod, Southam-Gerow, Tully, Rodríguez, & Smith, 2013; Proctor et al., 2011; Schoenwald et al., 2011), defined as both the extent to which the interventions were delivered as designed and the quality of that delivery (Bellg et al., 2004; Southam-Gerow & Mcleod, 2013). Treatment integrity has a number of components, each playing an important role in operationalizing treatment delivery, and ultimately, believed to contribute to the process of therapeutic change (Fjermestad, McLeod, Tully, & Liber, 2015). Researchers use treatment integrity to assess how therapists deliver evidence-based treatments [EBTs] in practice settings; for example, some researchers speculate that the extent to which therapists deliver therapeutic interventions that align with a treatment protocol (i.e., a component of treatment integrity referred to as adherence) may influence client outcomes (e.g., symptom reduction, overall functioning; McLeod et al., 2013). Instruments that assess treatment integrity information can be used to facilitate research (e.g., Schoenwald et al., 2011), drive implementation efforts (e.g., Sheidow, Donohue, Hill, Henggeler, & Ford, 2008), and serve as quality control tools (i.e., instruments designed to establish and preserve the integrity of the therapeutic interventions being implemented; Garland & Schoenwald, 2013).

Pragmatic treatment integrity instruments represent an important need for implementation researchers (Schoenwald et al., 2011). Observational instruments are considered the gold standard for treatment integrity assessment (Hogue, Liddle, & Rowe, 1996; Miller, Sorensen, Selzer, & Brigham, 2006; Lerner, McLeod, & Mikami, 2013). However, most observational instruments are costly in terms of the time and resources needed to employ them (Manuel, Hagedorn, & Finney, 2011). Though client- and therapist-report treatment integrity instruments may represent cost-effective alternatives, they often fail to show adequate correspondence with observational instruments; thus the validity evidence of these instruments has been called into question (e.g., Chapman, McCart, Letourneau, & Sheidow, 2013; Hogue, Dauber, & Henderson, 2012; Martino, Ball, Nich, Frankforter, & Carroll, 2009). The development of pragmatic *observational* treatment integrity instruments (i.e., brief, observational treatment integrity instruments that have demonstrated strong validity evidence and sensitivity to change), could fulfill an important need in implementation research (Glasgow & Riley, 2013; Schoenwald et al., 2011; Sheidow et al., 2008). These instruments would possess the advantage of validity without the limitation of needing excessive resources, as is typical of comprehensive observational instruments. This method could help implementation researchers develop or use treatment integrity instruments that optimally balance efficiency and effectiveness in practice settings (Schoenwald et al., 2011).

A pragmatic observational treatment integrity instrument has the potential to quicken the pace of research and serve as an ideal quality control tool. For example, researchers sometimes use such an instrument to guide therapist certification procedures, to ascertain that therapists have acquired the competencies necessary to participate in the study (Carroll et al., 2002; Roth, Pilling, & Turner, 2010). In addition, researchers sometimes instruct supervisors to use such an instrument to give therapists feedback and enhance their performance during EBT training (Sheidow et al., 2008; Smith et al., 2012). Using pragmatic observational instruments would be an efficient way (i.e., reduced time, financial resources, and burden) to conduct these quality control methods. However, the field has not yet identified suitable instruments for these tasks. As a result, research that develops such an instrument would have many potentially important applications.

The goal of the proposed project was to take an initial step toward developing an observational treatment integrity instrument that strikes a balance between effectiveness (i.e., reliability, sensitive to change, convergent validity, discriminant validity, predictive validity) and efficiency (i.e., brief, easy to administer and use; Schoenwald et al., 2011). The current study focused on shortening the coding process while maintaining strong validity evidence for the study's sample; the goal was to develop an instrument that can be field tested in the future, to determine the instrument's clinical utility for implementation research conducted in practice settings. In particular, the study assessed whether it is possible to reduce the amount of time that coders spend making treatment integrity ratings of therapists delivering individual cognitive-behavioral treatment (ICBT) for youth anxiety in both research and practice settings. The efficient observational treatment integrity instrument was derived from four observational treatment integrity instruments (see McLeod, Smith, Southam-Gerow, Weisz, & Kendall, 2015; McLeod et al., 2016; Shelef & Diamond, 2008; Southam-Gerow et al., 2016) and assessed four different treatment integrity components.

However, before I fully describe the purpose of the current study, I will provide a more comprehensive literature review. I will review the need for pragmatic instruments, the importance of treatment integrity in implementation research, the lack of pragmatic treatment integrity instruments in implementation science, and the role that pragmatic observational treatment integrity instruments could play in implementation research.

Literature Review

Researchers and practitioners agree that efficient, effective instruments are needed to support implementation research (Glasgow, Fernald, & Green, 2005; Glasgow & Riley, 2013). Presently, the measurement of constructs critical to implementation science is hampered by impractical instruments that are not suited for use outside research contexts (Garland & Schoenwald, 2013). For example, a number of instruments are too long, difficult to use, or insensitive to change to use in typical practice settings (e.g., Rabin et al., 2012). Although researchers have focused on making clinical trials more feasible to conduct in practice settings (e.g., Chamberlain et al., 2012), much less attention has been devoted to developing instruments for optimal use in practice settings (Berwick, 2005). Efficient instruments with strong validity evidence are needed to promote implementation efforts and serve as quality control instruments (e.g., supervision feedback tools; Glasgow et al., 2005; Glasgow & Riley, 2013; Schoenwald et al., 2011). To increase the number of these instruments, some have advocated for the development of *pragmatic instruments*, defined as instruments that are (a) efficient–brief, easy to administer and use (b) have strong validity evidence for use with underserved populations, and (c) are sensitive to change (Glasgow & Riley, 2013). However, few instruments that meet these criteria exist, so it is unclear if pragmatic instruments can even be developed.

Treatment Integrity Measurement

One important facet of implementation research that calls for more pragmatic instruments is the study of how therapists actually deliver EBTs in practice settings. Researchers that focus on this aspect of implementation often use treatment integrity instruments, or tools that assess the extent to which therapeutic interventions are delivered as designed (Bellg et al., 2004; Southam-Gerow & Mcleod, 2013). Treatment integrity instruments allow researchers to assess how therapists' EBT delivery influences the overall implementation of EBTs in practice settings (Garland & Schoenwald, 2013). In this section, I will review treatment integrity components, the role of treatment integrity in the process of therapeutic change, and the importance of treatment integrity instruments in guiding implementation efforts and serving as quality control tools.

Treatment integrity components. Treatment integrity breaks therapists' EBT delivery into components that together, and in isolation, reflect whether the treatment was delivered as designed. Treatment integrity components were originally designed to assess whether the independent variable in treatment research was manipulated as intended (Waltz, Addis, Koerner, & Jacobson, 1993). Since then, researchers have used treatment integrity for multiple reasons and differed in their views of the components that comprise treatment integrity. Most researchers agree that at least three components (i.e., adherence, competence, and differentiation) contribute to the construct of treatment integrity; these components assess the quantity and quality of therapeutic intervention delivery, and are believed to play a central role in therapeutic change (Bellg et al., 2004; Hagermoser Sanetti, Chafouleas, Christ, & Gritter, 2009; Perepletchikova &

Kazdin, 2005). Some researchers have also hypothesized that relational elements (e.g., the child-therapist relationship) influence therapeutic change and should therefore be included in the construct of treatment integrity (e.g., Allen, Linnan, & Emmons, 2012; Lichstein, Riedel, & Grieve, 1994; McLeod, Southam-Gerow, & Weisz, 2009; Stallard, Myles, & Branson, 2014). These researchers argue that relational elements should be included because their presence is essential for treatment delivery components to be effective. Regardless of the exact components, the literature suggests that treatment integrity components contribute to therapeutic change, either directly or indirectly influencing outcomes.

Treatment integrity as part of the treatment process. Fjermestad, et al. (2015) recently developed a theoretical model of treatment processes (i.e., the components that play a role in treatment) that makes a compelling argument for how adherence, differentiation, competence, and relational elements all work together to produce therapeutic change (see Figure 1); this model can therefore be used to support the argument that relational elements (i.e., alliance, client involvement, client comprehension), alongside adherence, competence, and differentiation, constitute treatment integrity (e.g., Allen et al., 2012). The model demonstrates that each of these treatment integrity components plays a unique, critical role in successful treatment.

The two "treatment delivery" components of the model focus on either the therapeutic interventions that therapists deliver (adherence and differentiation) or the therapist's skill (competence; McLeod et al., 2009). Adherence refers to the extent to which the therapist's delivery of therapeutic interventions corresponds to the treatment's specified protocol (i.e., the manual developed to standardize and specify treatment procedures and steps; Perepletchikova & Kazdin, 2005; Schoenwald et al., 2011). Differentiation, in contrast, refers to the extent to which

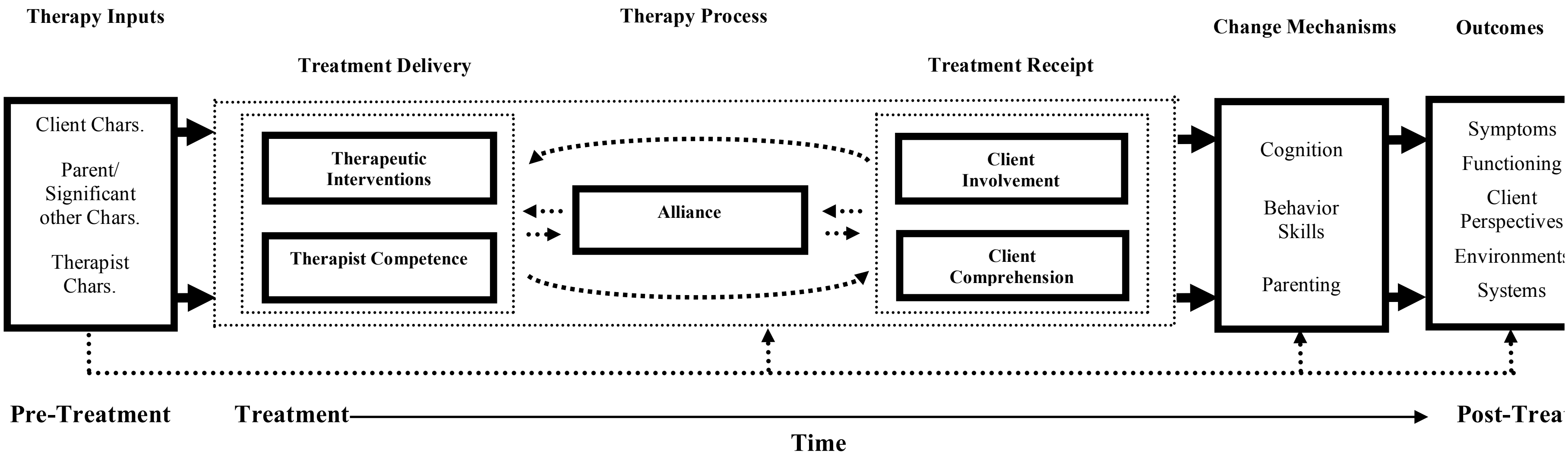

Figure 1. Theoretical Model of Therapeutic Change in Therapy (Fjermestad et al., 2015). Chars. = Characteristics.

the therapist diverges from the specified treatment protocol (e.g., delivering psychodynamic interventions while using an individual cognitive-behavioral treatment [ICBT] protocol; Southam-Gerow & Mcleod, 2013). Combined, these components have an important position in the treatment process; some researchers speculate that the delivery of specific therapeutic interventions could be directly associated with a reduction in client symptoms (e.g., Silverman, Pina, & Viswesvaran, 2008). Finally, competence reflects the therapist's degree of skill and responsiveness when delivering therapeutic interventions; this includes both "common" competencies (i.e., competencies that are considered important across theoretical orientations) and "technical" competencies (i.e., competencies that are specific to the EBT protocol; Bellg et al., 2004; Perepletchikova, Treat, & Kazdin, 2007). A number of researchers hypothesize that this aspect of treatment integrity may enhance the effectiveness of the therapeutic interventions delivered, by solidifying the alliance and increasing client involvement (e.g., Smith, Dishion, Shaw, Wilson, 2013). Together, these factors of treatment integrity operationalize the core information found in treatment protocols, give treatment sessions structure and depth, and strengthen the alliance (Fjermestad et al., 2015).

The "alliance" component of the treatment process model is a relational element (McLeod et al., 2009). Though there are many definitions of alliance in the literature, Bordin's (1979) definition of alliance incorporates the perspective of a number of theoretical orientations and is endorsed by multiple researchers (e.g., Fjermestad et al., 2015; McLeod & Weisz, 2005; Shelef & Diamond, 2008; Shirk & Saiz, 1992). Bordin (1979) proposes that the alliance represents the extent to which (a) affective aspects of the client-therapist relationship are present, and the extent to which therapists and clients agree on (b) tasks and (c) goals in the session. For example, the alliance may include the extent to which clients feel supported by their therapists or clients and therapists share a similar perception of the client's problems (Shelef & Diamond, 2008). Though different definitions of alliance exist, a number of researchers have agreed that the alliance is essential to successful psychosocial treatment, either directly influencing outcome (e.g., Norcross & Lambert, 2001), or influencing outcome by increasing client involvement (e.g., Kendall & Ollendick, 2004).

Finally, the "treatment receipt" component of the model refers to the aspects of treatment that involve clients' behaviors within the treatment session; this component includes both client involvement and client comprehension. Client involvement depicts the extent to which the youth participates and psychologically engages in therapeutic activities (Braswell, Kendall, Braith, Carey, & Vye, 1985; Chu & Kendall, 2004; Eugster & Wampold, 1996). Client comprehension, in comparison, focuses on the extent to which the youth portrays an understanding of the therapeutic material (Bellg et al., 2004; Hulleman & Cordray, 2009). Though there is a dearth of research in the area of treatment receipt, a past study found that client involvement was associated with symptom reduction (Chu & Kendall, 2004).

In sum, treatment integrity enables implementation researchers to operationalize EBT implementation by separating it into components that operate in a larger treatment process framework. The "treatment delivery" components of the model constitute the primary ingredients of treatment protocols, whereas the "alliance" and "treatment receipt" components constitute the aspects of implementation that facilitate clients' openness to learning core EBT content (Fjermestad et al., 2015). Therapists can skillfully deliver elements from the EBT protocol, but they may only produce positive outcomes if the youth has a good relationship with the therapist and is engaged in treatment (Fjermestad et al., 2015). Combined, these treatment integrity components have a substantial position in the overall process of therapeutic change (McLeod, Islam, & Wheat, 2013).

Figure 1 depicts the central role treatment integrity components play in psychosocial treatment (Fjermestad et al., 2015). Pre-treatment factors such as therapist, client, parent, and setting characteristics (labeled “therapy inputs” in the model) influence the extensiveness and quality of therapeutic intervention delivery (all labeled “treatment delivery” in the model). These components in turn influence the alliance, and the alliance in turn influences client involvement and comprehension (labeled “treatment receipt” in the model). Then, treatment delivery, alliance, and client involvement become intertwined, working in a bidirectional way to influence client cognition, behavior skills, or parenting (labeled “change mechanisms” in the model; Doss, 2004; Kazdin, 2000; Marker, Comer, Abramova, & Kendall, 2013; McLeod et al., 2015). Finally, these change mechanisms lead to alterations in post-treatment variables such as symptoms, functioning, client perspectives, environments, and systems (labeled “outcomes” in the model). The model depicts the function of treatment integrity and highlights its importance in youth psychosocial treatment.

The importance of treatment integrity. Given treatment integrity’s role in successful treatment, it is especially valuable to implementation science (Proctor et al., 2011). Implementation researchers can use treatment integrity components in a variety of ways to suit their unique goals. For example, some implementation researchers may investigate how unique aspects of practice settings (e.g., therapist attitudes) influence treatment delivery (e.g., less adherence to the treatment protocol), which in turn may influence alliance (e.g., client feels less understood) and client involvement (e.g., client does not participate fully in session). This information may help them improve implementation efforts, for instance by altering treatment protocols to accommodate therapists of more diverse backgrounds (Aarons, Hurlburt, & Horwitz, 2011). Other implementation researchers may request that supervisors directly and regularly assess therapist's adherence, competence, and alliance; in these instances, the treatment integrity instruments are quality control tools, helping supervisors provide therapists feedback about their EBT performance and better monitor the implementation process (Schoenwald et al., 2011; Sheidow et al., 2008). In general, treatment integrity tools have a number of applications in implementation science; they can facilitate research and, more practically, serve as a tool to determine and maintain the treatment integrity of implementing interventions (i.e., as a quality control instrument; Garland & Schoenwald, 2013).

Need For Pragmatic Treatment Integrity Instruments

Treatment integrity components play a key role in the process of therapeutic change and implementation efforts, but impractical instruments hinder their informational value. In this section, I will describe why inefficient treatment integrity instruments present a barrier to implementation science. Then, I will delineate how the development of pragmatic treatment integrity instruments could help researchers overcome a number of EBT implementation barriers, and ultimately advance current implementation efforts.

Impractical treatment integrity instruments. Developing treatment integrity instruments that balance both validity evidence and efficiency is an important need for implementation researchers (Garland & Schoenwald, 2013; Schoenwald et al., 2011). To date, treatment integrity measurement has been hindered in implementation research by impractical instruments that have established little validity evidence for use in practice settings (Berwick, 2005; Glasgow, 2013). For example, a number of researchers use ad-hoc instruments that have only been developed and tested in research settings (e.g., Kahler et al., 2008; Litt, Kadden, &

Kabela-Cormier, 2009). In addition, Perepletchikova et al. (2007) found that only 3.5% of 147 child and adult psychosocial treatment studies published between 2000 and 2004 adequately measured treatment integrity; implementation researchers may use inappropriate instruments or fail to assess key treatment integrity constructs in practice settings. Moreover, methods that implementation researchers currently use for treatment integrity measurement, observational instruments and therapist- and client-report instruments, have notable disadvantages in practice settings.

Observational instruments (i.e., the process in which data are gathered through in-vivo observation, videotape, audiotape, or digital files; Schoenwald et al., 2011) are considered the gold standard for treatment integrity assessment (Hogue et al., 1996; Miller et al., 2006). This measurement method has a number of advantages for implementation researchers: it fits a diverse array of contexts, can include instruments of multiple domains in one observation, and may be less biased than self-report methods (DiMatteo, 2004; Noell et al., 2005). Implementation researchers who use this method often hire trained coders (i.e., employed by the research project) who are blind to study purposes, to watch therapists' sessions and rate their behaviors (e.g., on a dimensional scale or using presence/absence checklists; e.g., Godley, Garner, Smith, Meyers, & Godley, 2011). Some researchers who use this method request that expertly trained supervisors (i.e., employed by the research project) watch therapists' sessions and rate therapist behaviors (e.g., Stein, Herman, & Anderson, 2009). Further, some request that local supervisors (i.e., employed by the practice setting) watch therapists' sessions and rate therapist behaviors (e.g., Ball et al., 2009; Martino et al., 2011). However, all of these methods are costly in terms of the time and money needed to employ them (Manuel et al., 2011). Hiring blind coders or requesting that supervisors watch and rate therapists' *entire* treatment sessions is often not affordable or sustainable as a long term strategy in practice settings (Bellg et al., 2004; Hogue, Ozechowski, Robbins, & Waldron, 2013; McLeod et al., 2009; Shelef & Diamond, 2008). In addition, practice setting therapists and supervisors, due to their diverse training background and values, may not always be open to being observed or videotaped for research purposes; for example, they may perceive this method as intrusive or anxiety-provoking (Garland, Bickman, & Chorpita, 2010; Miller et al., 2006). Implementation researchers considering this method must therefore weigh advantages against factors such as feasibility, budgeting, and staff opinions.

Client- and therapist-report instruments, in comparison, are inventories (e.g., dimensional ratings, presence/absence checklists, or diaries) that are completed independently by either therapists or clients (Hogue et al., 1996). Therapists and/or clients are often administered these instruments weekly, immediately after the session (e.g., Hogue et al., 2008). Client- and therapist-report instruments are cost-effective and require minimal resources and time, optimal for use in practice settings (e.g., Orimoto, 2012; Weisz et al., 2012). However, they often fail to show adequate correspondence with observational instruments (e.g., Chapman et al., 2013; Hogue et al., 2012; Martino et al., 2009). For example, Martino and colleagues (2009) found poor agreement between observers and therapists in regards to the extent to which they delivered specific interventions, with therapists reporting greater use of EBT interventions than observers. Thus, although client- and therapist-report instruments are usually the most efficient measurement method in practice settings, some question whether these instruments have sufficient validity evidence for use with underserved populations in practice settings.

Pragmatic treatment integrity instruments. Given the limitations of observational and client- and therapist-self report methods, there is reason to believe that a better way to balance effective versus efficient treatment integrity measurement is needed. Implementation researchers must develop or use treatment integrity instruments that can be administered quickly, scored accurately, and used for multiple quality control purposes (e.g., therapist training and supervision); they therefore require pragmatic treatment integrity instruments, tools that are brief (e.g., take 10 minutes to code rather than the current average of 50 minutes; McLeod et al., 2015), have strong validity evidence for use in practice settings, and are sensitive to change in intervention delivery across settings. Implementation researchers could turn to pragmatic treatment integrity tools to improve the balance of efficiency versus effectiveness measurement in practice settings.

Need for Pragmatic *Observational* Treatment Integrity Instruments

Pragmatic *observational* treatment integrity instruments could be leveraged to improve implementation research and practice. In this section, I will review how pragmatic observational treatment integrity instruments have the potential to enhance key implementation efforts (e.g., assessment of internal and external validity, funding, treatment protocol development) and serve as key quality control tools (e.g., used to guide therapist training, supervision; Glasgow & Riley, 2013; Schoenwald et al., 2011).

Streamlining the assessment of external and internal validity. Pragmatic observational treatment integrity instruments could enhance implementation efforts in a number of ways. For one, researchers could use them in place of current, lengthy observational instruments to quickly ascertain that an EBT or specific independent variable was manipulated as purported (i.e., rule out confounding variables and establish internal validity) and that outcomes generalize to similar populations (i.e., establish external validity; Kazdin, 2003). Researchers can take over a year just to produce observational ratings of sessions, not including the time needed to analyze findings (e.g., McLeod et al., 2015; Shelef & Diamond, 2008); pragmatic observational treatment integrity instruments could shorten this timeframe. In addition, client- and therapist-report instruments do not always produce strong validity evidence in practice settings (Chapman et al., 2013; Hogue et al., 2012; Martino et al., 2009); this method could therefore be used to improve the accuracy of internal and external validity conclusions. Overall, this method would allow researchers to more rapidly interpret study findings, publish papers, and potentially disseminate important clinical information.

Increasing implementation resources. A pragmatic observational instrument could also increase the resources available for implementation efforts. Stakeholders often seek treatment integrity information to confirm that the program they have funded is actually being implemented as designed in practice settings (McLeod et al., 2013; Schoenwald et al., 2011); sending stakeholders observational treatment integrity information shortly after study completion could encourage them to bolster their investments in EBT implementation efforts. In addition, both researchers and stakeholders could benefit from the resources gained from reduced coding time (e.g., allocation of important funds elsewhere; Miller et al., 2006). For example, clinics could be allotted extra funds to do things such as enhance their video recording technology or reward therapists for their EBT delivery efforts (Aarons et al., 2011). Pragmatic observational treatment integrity instruments could increase stakeholder investments and optimize implementation funding.

Facilitating protocol development. In addition, pragmatic observational treatment integrity instruments could promote treatment protocol development. Researchers could use the instruments to quickly and accurately decipher which interventions are effective in practice settings and thus warrant incorporation into treatment protocols (i.e., practice-based evidence;

Garland, Hurlburt, & Hawley, 2006). For instance, if therapists diverge from the treatment protocol (e.g., deliver family interventions while using an ICBT protocol), and this enhances client outcomes (e.g., increases symptom reduction), researchers may decide to incorporate certain family interventions into the ICBT protocol. Similarly, if therapists cannot understand content in a protocol during therapist training and this consistently influences treatment integrity scores or client outcomes, researchers may consider altering language or interventions recommended in a protocol (Aarons et al., 2011). Pragmatic observational instruments could help researchers make important protocol changes more quickly than they could with lengthy observational instruments.

Improving therapist training. Pragmatic observational treatment integrity instruments could also be used as quality control tools, with applications such as guiding therapist training. Researchers could use these instruments to establish more efficient, yet effective therapist certification procedures (i.e., the process in which therapists demonstrate that they have acquired the competencies necessary to participate in the study; Carroll et al., 2002). For example, project experts could watch a segment of therapists' sessions and require that therapists meet certain competency ratings on the pragmatic observational integrity instruments (e.g., 4 out of 7) before they can be assigned clients (e.g., Godley et al., 2011). This could save implementation researchers valuable time training therapists so that training could focus more intensely on methods known to enhance therapist EBT performance (i.e., feedback, modeling, role-playing; Beidas & Kendall, 2010; Miller, Yahne, Moyers, Martinez, & Pirritano, 2004). Moreover, this method may be more effective than therapist- and client-report instruments. For example, therapists may be more likely to exaggerate their use of EBTs on treatment integrity self-report instruments if the data are used for evaluative purposes, rather than kept anonymous for study analyses (Martino et al., 2009). This instrument could help staff provide therapists immediate, accurate feedback about their EBT performance, and could help researchers better manage the EBT training process.

Enhancing supervision. Finally, supervisors could use these quality control tools to enhance their feedback to therapists throughout EBT implementation (Sheidow et al., 2008; Smith et al., 2012). These efficient instruments could be used *during* supervision, feasible for supervisors who do not have time outside of meetings to evaluate EBT performance (Hogue et al., 2013). For example, supervisors could watch a 10-minute segment of therapists' sessions and give them feedback about their adherence, competence, differentiation, and alliance (Ball et al., 2009; Sheidow et al., 2008). Further, these efficient tools could promote the sustainability of the EBT program; local supervisors (i.e., supervisors within the funding and organizational structure of the practice settings) could use them long after the contracted EBT experts withdraw resources or consultation infrastructure from the clinic (Dorsey et al., 2013; Hogue et al., 2013). Overall, these instruments could help researchers and supervisors monitor EBT implementation more effectively and efficiently.

A pragmatic observational treatment integrity instrument would have potentially important implications. Researchers, stakeholders, and practitioners would all benefit from the applications of these instruments, including reduced resources needed to invest in EBT studies and enhance quality control practices. However, research to date has not identified suitable pragmatic observational treatment integrity instruments for these tasks. Thus, it is an open question – can efficient, yet effective observational treatment integrity tools even be developed?

Overview of Current Study

The goal of the current study was to take an *initial* step toward answering that question. I aimed to develop an observational treatment integrity instrument that strikes a balance between efficiency (i.e., brief, easy to administer and use) and validity evidence (e.g., Glasgow & Riley, 2013; Schoenwald et al., 2011). The instrument tested whether it is possible to reduce the time needed to code treatment sessions, evaluating the *feasibility* of such a treatment integrity measurement method. If treatment integrity scores produced by coders in shorter time frames evidenced strong validity evidence, then results would suggest that it is worthwhile for future researchers to conduct similar field tests of pragmatic observational treatment integrity instruments in efficacy and effectiveness studies. Moreover, the methodology for developing and validating this instrument could serve as a guide for future research. In this section, I will describe why I selected the study's coding process, treatment setting, treatment type, treatment design, treatment integrity components, and treatment integrity instruments. Then, I will describe why I assessed specific validity dimensions (i.e., reliability, sensitivity to change, convergent validity, discriminant validity, predictive validity). Finally, I will describe the specific aims, hypotheses, and research questions.

Coding process. To test the feasibility of developing an implementation and quality control tool, the present study aimed to reduce the time that coders spend making treatment integrity ratings while maintaining the instrument's validity evidence that approximates ratings from the full session. To date, most studies that have attempted to shorten the observational coding process have selected portions of treatment sessions to code, using conceptual rather than empirical rationale (e.g., Carpenter et al., 2012); thus, it has been unclear whether their treatment integrity findings accurately depict treatment delivery of the entire sessions. Pierson and colleagues (2007) were among the first researchers to attempt to *empirically* shorten the observational treatment integrity coding process, by coding adherence and competence in 10 out of the full 20 minutes of a role-played motivational interview session and statistically comparing those segments with the full 20-minute treatment integrity scores. They found that the 10-minute session segment scores yielded similar inter-rater reliability and treatment integrity results as the full scores, and thus could be a more efficient way to assess treatment integrity.

Weck, Grikscheit, Höfling, and Stangier (2014) also empirically tested whether an observational treatment integrity instrument could be effective with a portion of a treatment session; however, their study had a number of advantages, including assessing the instrument's use with *real* clients, rather than role-plays (Decker, Carroll, Nich, Canning-Ball, & Martino, 2013). Coders rated adherence and competence, in *the middle third* of a treatment session (average of 18 minutes in duration) and authors evaluated whether it adequately represented independent coder ratings from the entire treatment session. They found that (a) adherence and competence ratings were strongly correlated across observation length groups, (b) inter-rater reliability did not differ across groups, and (c) ratings similarly predicted client outcomes across groups; they therefore concluded that coding the middle third of sessions might be an efficient, cost-effective way to use observational treatment integrity instruments and improve intervention studies.

The current study builds on past studies by evaluating whether shortening the coding process can increase the instrument's efficiency while preserving its effectiveness. Similar to Weck et al. (2014), the current study compared middle session segments to entire session ratings, but the current study also manipulated coding time (i.e., coding the middle five- and 15-minute portions of a session). Rather than just dividing the session into thirds and coding the middle portion, the current study assessed which time frame produced the strongest reliability and validity evidence for the sample (i.e., comparable to the sample's entire session ratings) while remaining efficient (i.e., shortest coding time). The middle portion of treatment was sampled because this portion is considered the "working stage" of treatment, and should therefore include the most relevant EBT content (McLeod & Weisz, 2010). Five-minute segments were chosen because studies have found that judges can make relatively accurate ratings of clinician's nonverbal behavior in less than five minutes (Slepian, Bogart, & Ambady, 2014); thus, it is an open question if judges can rate both verbal and nonverbal behavior in that timeframe. Fifteen-minute segments were chosen because similar timeframes have been found to map onto full session scores in the past (e.g., Weck et al., 2014), and thus this study could replicate past findings. Overall, the current study evaluated whether validity evidence can be maintained when the coding process is shortened.

Treatment setting, type, and design. Next, the treatment setting, treatment type, and treatment design were chosen. In addition to having a brief assessment period, an ideal implementation and quality control tool would accurately represent the intended theoretical criterion (e.g., establish convergent and discriminant validity with the full set instruments) in both research and practice settings; thus, the current study aimed to test whether it was possible for the instrument to establish distinct treatment integrity ratings of delivery in different settings, with different treatments, and different treatment designs. The study evaluated therapists delivering (a) standardized ICBT for youth anxiety in a research setting, and therapists delivering (b) standard ICBT, (c) modular ICBT, and (d) usual clinical care (UC) for youth anxiety in practice settings. The study builds on the Weck et al. (2014) study design, by assessing how empirically reducing coding time may differentially influence the efficient instrument's ability to maintain reliability and validity in reference to the full set scores based on treatment setting, type, and design.

Treatment setting. The current study used data from treatment sessions conducted in both research and practice settings. Weck and colleagues (2014) only tested whether coding time could be empirically reduced by using data from research setting sessions. The current study aimed to expand upon Weck and colleagues' findings, by assessing if findings can be replicated in both research and practice settings. This treatment setting choice also has important implications for implementation research. By comparing the delivery of the same EBT across settings, the current study assessed whether the effectiveness of the efficient observational treatment integrity tool has the potential to generalize across settings. For example, if an efficient observational instrument produces strong validity evidence in research settings, but not practice settings, findings would raise concerns that the instruments may have little utility in practice contexts and vice versa. The present study aimed to pinpoint differences in the instrument's effectiveness across the two settings that can be explored in future studies.

Treatment type. The current study also drew from clients receiving two different types of treatments, ICBT and UC. ICBT for youth anxiety was chosen because it is an ideal EBT for study purposes. For one, ICBT for youth anxiety is a structured EBT program that lends itself well to systematic measurement (Kendall, Hudson, Gosch, Flannery-Schroeder, & Suveg, 2008). ICBT for youth anxiety also has received substantial empirical support in efficacy trials, and is considered "well established" (Chambless & Hollen, 1998; Walkup et al., 2008). It is therefore worthwhile to develop an instrument that can be further developed and used in numerous studies involving ICBT (Silverman et al., 2008). Finally, ICBT for youth anxiety is well-suited for implementation research, as it has not always performed as expected when delivered in practice settings (Barrington, Prior, Richardson, & Allen, 2005; Southam-Gerow et al., 2010); thus, there may be differences in the way that therapists deliver the same ICBT interventions across settings, providing ample data for analyzing the convergent and discriminant validity evidence for the instrument. The UC treatment group was also chosen for discriminant validity purposes. The current study aimed to assess if the efficient observational treatment integrity instrument could accurately discriminate between interventions delivered by therapists who were instructed to deliver ICBT for youth anxiety and interventions delivered by those who were instructed to deliver interventions they regularly used and believed to be effective in practice settings. Using the efficient instrument with different treatment types could have value for future implementation research. If this instrument minimally distinguishes between UC and ICBT, then results would indicate that the instrument cannot accurately differentiate between the typical care in practice settings and EBTs, and may not warrant future tests of its usefulness as a manipulation check or quality control tool; for instance, therapists will not be able to receive helpful feedback during EBT training if a tool does not sufficiently portray their progress in EBT rather than UC skills. The current study aimed to assess the extent to which the instrument distinguished between treatments, to contribute to findings about whether it is worthwhile to test as an implementation and quality control tool in future studies.

Treatment design. The current study sampled from participants receiving two different ICBT designs: standard manualized treatment (SMT) and modular manualized treatment (MMT). The therapists in the SMT condition were instructed to deliver the *Coping Cat* (CC) protocol, a structured ICBT program aimed at treating youth diagnosed with anxiety disorders in 16-20 sessions (Kendall & Hedtke, 2006; Kendall, Kane, Howard, & Siqueland, 1990). Therapists in modular manualized treatment (MMT), in comparison, were instructed to deliver

Modular Approach to Therapy for Children with Anxiety, Depression, and Conduct Problems (MATCH; Chorpita & Weisz, 2005), a flexible ICBT program that was designed to cater to clients' individual needs and includes 31 modules aimed at treating youth diagnosed with anxiety, depression, or conduct problems. Sampling from clients receiving different treatment designs allowed the study to more thoroughly assess the instrument's convergent validity. This sample tested if the efficient observational treatment integrity instrument could comparably assess therapists delivering the same general treatment, but different protocols. This comparison could guide future implementation researchers who aim to field-test the efficient instrument (a) with either of these two treatment designs, or (b) for further assessment of how well the instrument generalizes across treatment designs.

Treatment integrity components. The present study aimed to design an instrument that could eventually be used in implementation studies to promote successful therapeutic change. Thus, the efficient instrument systematically assessed a range of treatment processes theorized to play an important role in therapeutic change (Fjermestad et al., 2015). Though researchers do not agree on the exact components of treatment integrity, this study takes the stance of some researchers (e.g., Allen et al., 2012) who propose that treatment integrity includes alliance (e.g., due to the evidence behind its contribution to therapeutic change; Karver, Handlesman, Fields, & Bickman, 2006; McLeod, 2011). The current instrument thus assessed four treatment integrity components: adherence, competence, differentiation, and alliance. Taken together, this instrument aimed to contribute to implementation science, by assessing the feasibility of combining treatment integrity components into one rich, yet efficient, implementation and quality control tool.

Treatment integrity instruments. Existing treatment integrity instruments were carefully selected for the study's efficient instrument development. To meet the needs of the current study, an instrument needed to (a) assess adherence, competence, differentiation, or alliance with youth receiving ICBT for youth anxiety, (b) be an observational instrument, and (c) have strong reliability and validity evidence based on ratings of entire treatment sessions. Four observational treatment integrity instruments developed or tested in the Treatment Integrity Measurement Study (TIMS; NIMH RO1 MH086529) fulfilled this need. These four instruments separately assess both technical and relational treatment integrity components, including adherence (Southam-Gerow et al., 2016), competence (McLeod et al., 2016), differentiation (McLeod et al., 2015), and alliance (Shelef & Diamond, 2008). In addition, all of these instruments have demonstrated reliability and convergent validity. Lastly, all the instruments have similar terminology and include parallel interventions across scales (e.g., the same core interventions in adherence and competence instruments); thus, they are poised for combination into one, efficient observational instrument.

Reliability. An instrument can be considered reliable if it performs in predictable, repeatable ways; though reliability is not a sufficient condition to measure validity, an instrument must be reliable in order to be considered valid (Devellis, 2012). Thus, reliability analyses were an important component of the current study. By assessing inter-rater agreement, I could make conclusions about how well the instrument can be expected to perform again, when rated by different coders.

Sensitivity to change. Given my aim to develop an instrument that can eventually be used as an ongoing quality control tool, I examined how sensitive the instrument was to change in therapist behavior over time (Glasgow & Riley, 2013). Researchers could use a pragmatic observational treatment integrity instrument multiple times (e.g., biweekly) over the course of EBT training to assess therapists' ability to learn and improve their delivery of therapeutic interventions; thus, the current study's efficient instrument should be able to detect changes in therapists' treatment delivery (e.g., Godley et al., 2011). In addition, some implementation researchers propose that the type and dosage of interventions delivered may change even after EBT training has ended, throughout EBT trials (e.g., Boswell et al., 2013; Henggeler et al., 2008); thus, staff and/or local supervisors could use an efficient instrument to identify alterations in therapists' behaviors and provide feedback throughout EBT implementation. The current study assessed the instrument's sensitivity to change to gauge its potential to serve as an ongoing quality control tool in future implementation studies.

Convergent validity. As aforementioned, the current study also evaluated the convergent validity of the instrument, or the extent to which scores on the efficient instrument were related to scores on the full instruments applied to the full session (Kazdin, 2003). This evaluation contributed to the study's goal of developing an instrument that is *effective* in implementation studies (Schoenwald et al., 2011) by assessing whether the efficient instrument scores accurately map onto the full set scores they intended to represent. For example, if the efficient scores were not associated with the full set scores, the findings would indicate that the instrument does not assess the intended criterion. These analyses would help inform the instrument's potential to be as effective as full set instruments for implementation research.

Discriminant validity. In addition, the current study established the instrument's discriminant validity, or the extent to which scores on the efficient instrument demonstrated little relationship with scores that were *not* expected to be related (Kazdin, 2003). For similar reasons as convergent validity, this validity dimension would help assess the extent to which the abbreviated instrument appropriately represents the treatment integrity criterion it purports to represent. Analyses were conducted to ascertain if the efficient instrument's treatment integrity scales can adequately differentiate between treatments (i.e., CBT vs. UC). These analyses would contribute to information about the instrument's ability to be used by future implementation researchers to assess EBT delivery above and beyond the delivery typically found in practice settings.

Predictive validity. Finally, the current study assessed the extent to which the efficient treatment integrity scores predicted client outcomes post-treatment, and the extent that those predictions align with the full set treatment integrity-client outcome findings. These analyses assessed if the instrument can predict client outcomes, as some researchers theorize. They also assessed if the instrument can predict outcomes as well as the comprehensive observational treatment integrity instruments. Overall, this validity dimension would contribute to the study's goal of assessing the instrument's effectiveness (Schoenwald et al., 2011) for use in future field tests in practice settings.

Specific study aims. Study goals were realized via three aims. First, the current study relied on expert consultation to select a finite number of items from adherence, competence, differentiation, and alliance instruments that represent core therapeutic interventions and have established validity evidence (Aim 1). Then, the new, efficient integrity instrument was used to code two pools of recorded sessions: sessions from an efficacy trial (Kendall et al., 2008) and sessions from an effectiveness trial (Weisz et al., 2012). The resulting data were used to establish reliability and validity (Aim 2) as well as ascertain the optimal observational length (if possible; Aim 3) of the new integrity instrument.

Hypothesis and research questions. Given the lack of previous research in this area, I developed three hypotheses and a number of exploratory research questions. I hypothesized that items identified for each treatment integrity scale (adherence, competence, differentiation, alliance) would (a) capture full variability in treatment integrity scores (e.g., ranging from 1-7), (b) produce scores that are strongly related to scores from the full set of items, and (c) discriminate treatment integrity across treatment settings (research, practice) and treatment types (CBT, UC). Together, these results would provide evidence to confirm or refute my hypothesis that a small set of items can be used as an implementation and quality control instrument for use in practice settings**.** I explored the following questions: What is the shortest TIES-CBT-YA observation segment that best approximates the (a) item distribution and (b) reliability evidence of the full session treatment integrity scores independently archived for TIMS? What is the shortest observation length that represents (c) sensitivity to change, (d) convergent validity, and (e) discriminant validity based upon comparisons with the treatment integrity scores independently archived for the two trials? What is the shortest observation length that (f) predicts improvements in symptoms at follow-up across the treatment integrity scores independently archived for the two trials (i.e., predictive validity)?

Method

Participants

Participants included youth diagnosed with a primary anxiety disorder drawn from two separate randomized controlled trials. The first study compared the efficacy of Individual Cognitive-Behavioral Treatment (ICBT), Family Cognitive Behavioral Treatment (FCBT), and Family-based Education/Support/Attention (FESA), at Temple University (Kendall et al., 2008). The second study, also referred to as the Child STEPS trial, investigated the influence of treatment design on the effectiveness of youth evidence-based treatment outcomes and procedures in Hawaii and Massachusetts (Weisz et al., 2012). This trial compared the effectiveness of modular manualized treatment (MMT) vs. standard manualized treatment (SMT) vs. usual care (UC) in 10 outpatient community mental health clinics, and included youth who were primarily diagnosed with anxiety, depression, and conduct problems (Weisz et al., 2012).

In the current study, participants were drawn from the ICBT condition of the efficacy trial (i.e., Kendall et al. study; *N* = 55) and all three conditions in the effectiveness study (i.e., Weisz et al., study); however, in the Weisz et al. study, only those participants with primary anxiety disorders were included (SMT; *N* = 25, MMT; *N* = 19, Usual Care (UC); *N* = 24). See Table 1 for descriptive information for each sample selected for the current study.

Table 1

Current Sample Demographic and Descriptive Information by Study

Variable	Kendall et al. ICBT (*n* = 51)	Weisz et al. (*n* = 55)
Youth age in years *(SD)*	10.36 *(1.90)*	9.89 *(1.71)*
Percent of female youths	39.20	45.50
Percent of Caucasian youths	86.30	54.50
Percent of family income		
Up to $60,000	35.30	52.70
Above $60,000	56.90	43.60
Missing	7.80	3.60
Percent of female therapists	88.20	80.00
Therapist age in years *(SD)*	Missing	40.91 *(9.62)*
Therapist years of experience *(SD)*	Missing	5.71 *(7.10)*

Kendall et al. study. Youth participants in the Kendall et al. (2008) study were recruited through multiple sources (e.g., clinics and practitioners, flyers, media). Their parents were given a brief phone screen after they contacted the Child and Adolescent Anxiety Disorders Clinic (CAADC) at Temple University (where the study was conducted). An intake interview was scheduled if the child was between ages seven and 14 and the parent was interested in participating in the study. At the intake, a semi-structured interview, the Anxiety Disorders Interview Schedule for DSM-IV: Child and Parent Versions (ADIS-C/P), was administered to both the youth and parents. The youth's diagnoses were determined using a composite approach (using the "or" rule) recommended by the ADIS-C/P protocol (Silverman & Albano, 1996). If the youth met diagnostic criteria for an anxiety disorder, the clinician assigned a clinician severity rating (CSR) that ranged from 1 to 8 (with 4 or greater indicating the presence of an anxiety disorder). Parents provided informed consent and the youth provided assent. Exclusion criteria consisted of: a disabling medical condition, an intellectual disability, psychotic symptoms, both parents being non-fluent in the English language, the child's taking medication for anxiety or depression, or the child's participation in concurrent psychosocial treatment. The institutional review board approved of all of these procedures, and compensation was provided for participation in the study.

Youth participants. In the Kendall et al. study, the parent sample consisted of 161 youths and their parents who were referred between 2000 and 2006 and were diagnosed with a primary anxiety disorder (generalized anxiety disorder [GAD], separation anxiety disorder [SAD], social phobia [SOP]). The current study drew from the 55 participants in the Kendall et al. study who were assigned to receive ICBT (attrition of five youths prior to the data collection from the 55 participants). Youth participants were excluded from the current sample (*n* = 4) if they received treatment from multiple therapists and/or if they had data from fewer than two sessions (as it is necessary to have at least two time points to assess treatment integrity over time). Thus, 51 youth remained in the final sample.

Of the 51 youth participants, ages seven to 14 (*M* = 10.36, *SD* = 1.90), 39.20% were females. The majority of children identified as Caucasian (86.30%); 9.80% identified as African American, 2.00% Latino, and 2.00% other. ADIS diagnostic information indicated that approximately 43.10% of participants were diagnosed with a comorbid anxiety disorder (not "co-principal"); only 47.10% were diagnosed with a comorbid non-anxiety disorder, and 9.80% had no comorbid diagnoses, indicating a high occurrence of comorbidity. Youth had an average of 3.02 (*SD* = 1.45) diagnoses. Approximately 35.30% of families reported family income below $60,000 a year and 56.90% of families reported family income above $60,000 a year.

Therapist participants. In the Kendall et al. parent study, 18 masters- and doctoral-level therapists delivered ICBT at the CAADC. Two therapists were excluded from the current study's ICBT condition due to exclusion criteria (e.g., cases involving multiple therapists). Data regarding therapists' age and overall years of clinical experience were not gathered, but the therapists (*N* = 16) were 75.00% female and 75.00% Caucasian. All therapists had 2-3 years of experience at the CAADC and were supervised by doctoral-level psychologists with 6-7 years of experience in the community. Doctoral candidates in clinical psychology also conducted the structured diagnostic interviews and assessments.

Weisz et al. study. Youth participants in the Weisz et al. parent study consisted of families that sought outpatient care (i.e., they were not recruited). An intake interview was scheduled if (a) the child was between ages seven and fourteen, (b) anxiety, depression, or conduct symptoms were reported, and (c) the parent was interested in participating in the study.

At the intake, a structured interview, the Children's Interview for Psychiatric Syndromes (CHIPS; Weller, Weller, Rooney, & Fristad, 1999a; 1999b), was administered to both the youth and parents. In the current study, youth were included if (a) they met diagnostic criteria for a primary anxiety disorder (GAD, SAD, SOP, Specific Phobia [SP], Obsessive Compulsive Disorder [OCD], Posttraumatic Stress Disorder [PTSD], and Panic Disorder without agoraphobia [PD]) according to the *Diagnostic and Statistical Manual of Mental Disorders,* 4th ed. (American Psychiatric Association, 2000) during the CHIPS interview, or (b) reported clinical elevations of anxiety (i.e., *T* scores of 65 or higher) on either the Child Behavior Checklist or Youth Self-Report (Achenbach & Rescorla, 2001). Parents provided informed consent and the youth provided assent. Exclusion criteria consisted of the youth having the following characteristics: an intellectual disability, psychotic symptoms, bipolar disorder, or a primary problem of inattention and/or hyperactivity. The institutional review board approved of all of these procedures, and compensation was provided for participation in the study.

Youth participants. In the Weisz et al. study, the parent sample consisted of 174 youths and their parents who sought outpatient treatment between 2005 and 2009. The current study drew from the 68 participants in the Weisz et al. study who were diagnosed with primary anxiety disorders. Youth participants were excluded from the current sample (*n* = 13) if they received treatment from multiple therapists and/or if they had data from fewer than two sessions. Therefore, 55 youth remained in the final sample.

Of the 55 youth participants, (aged 7 to 13 years; *M* = 9.89, *SD* = 1.71), 45.50% were females. Approximately 54.50% of participants identified as Caucasian, 30.90% multiethnic, 5.50% African American, 3.60% Latino, 1.80% Asian American/Pacific Islander, and 3.60% other. Youth had an average of 2.95 (*SD* = 2.01) diagnoses. Approximately 52.70% of parents reported family income below $60,000 a year and 43.60% reported family income above $60,000 a year.

Therapist participants. In the Weisz et al parent study, 40 therapists delivered the three treatments. Five therapists were excluded from the current study's ICBT condition due to exclusion criteria (e.g., cases involving multiple therapists). The remaining therapists (*n* = 35) were mostly female (80.00%). Over half identified as Caucasian (56.00%), 23.00% identified as Asian American, 6.00% identified as African American, and 6.0% identified as Pacific Islander. Therapists averaged 40.0 years old (*SD* = 9.6 years; *Range* = 25-59 years) and varied in years of clinical experience, ranging from 1.5 to 30.0 years (*M* = 5.75 years; *SD* = 7.2 years). Most had a Masters degree in social work (40.00%) or the arts (31.40%), some were licensed clinical social workers (14.30%), and some had doctorates in psychology (14.30%).

Treatment Methods

Standard Manualized Treatment (SMT). Therapists in both Kendall et al. (i.e., two conditions in a research setting) and Weisz et al. (three conditions in a practice setting) parent studies were trained to use standard treatment protocols, therapist manuals that instruct them to deliver specific interventions in a session-by-session format. Given the current study focus on anxiety disorders, the present study drew only from those clients assigned to receive the *Coping Cat* (CC) protocol individually; in other words, one SMT condition was selected from each parent study (ICBT *n* = 51 youth; SMT *n* = 22 youth). CC, an ICBT program aimed at treating youth diagnosed with anxiety disorders, includes 16-20 sessions conducted individually with the youth participant (Kendall & Hedtke, 2006; Kendall et al., 1990). The program aims to reduce anxiety through skill-building (e.g., relaxation, cognitive restructuring), graduated exposure to feared stimuli, and regularly assigned homework. Parents are included periodically to gather information about client's functioning, teach them about youth anxiety, and update them on their child's treatment progress.

Modular Manualized Treatment (MMT). In one condition in the Weisz et al. study, therapists delivered modular manualized treatment (MMT), a flexible treatment that was designed to cater to clients' individual needs. In comparison to the standard treatment, this treatment is much less structured and does not always follow a particular intervention sequence. Therapists used *Modular Approach to Therapy for Children with Anxiety, Depression, and Conduct Problems* (MATCH; Chorpita & Weisz, 2005), a protocol that consists of 31 modules that aim to address three problem areas: anxiety, depression, and conduct problems. The interventions correspond with those delivered in standard protocols, such as *Coping Cat*, *Primary and Secondary Control Enhancement Training* (*PASCET)*, and *Defiant Children*. Therapists were instructed to focus on the client's primary problem area based on findings from standardized and idiographic measures. If a crisis or new problem hindered the use of the current sequence of modules in MATCH, therapists methodically altered the sequence using other modules. For instance, if anxiety was the target problem but depression became an issue (e.g., youth began having thoughts about death with no self-harm plan), the therapist would deliver appropriate depression modules to address the disruption (e.g., activity selection), returning to anxiety-based modules after the problem was resolved. The current study will only focus on anxiety-based interventions (*n* = 16 youths).

Usual care. The Weisz et al. study included an UC condition. Therapists in this condition agreed to use the therapeutic interventions they regularly delivered and believed to be effective in their routine clinical practice. Treatment consisted of a variety of interventions from a multitude of theoretical orientations (McLeod & Weisz, 2010).

Therapist Training

Kendall et al. study. Therapists who were randomly assigned to both ICBT and FCBT conditions of the Kendall et al. parent study studied the Coping Cat protocol and participated in training (two 3-hour workshops) before beginning supervised trial experience. Workshops consisted of didactic lectures, role rehearsal, videotape review, trainee demonstration, and discussion groups. All therapists participated in weekly two-hour supervision groups throughout the trial.

Weisz et al. study. Therapists who were randomly assigned to both SMT and MMT groups of the Weisz et al. Child Steps study (CS-SMT and CS- MMT, respectively) were trained together in the respective treatment protocols by postdoctoral project consultants. The consultants initially received training from experts in the protocols. Therapists participated in six days of training workshops (two days on each problem area) before beginning supervised trial experience. All therapists received weekly supervision with consultants, and consultants subsequently had weekly discussions with the experts. Supervision included review of measurement feedback on client progress and therapists' treatment delivery records (Chorpita & Daleiden, 2009). Therapists assigned to the UC group (CS-UC) received local supervision, and did not meet with project personnel for the entirety of the study.

Diagnostic and Symptom Instruments

Kendall et al. study. ***Anxiety Disorders Interview Schedule for DSM-IV: Child and Parent Versions*** (ADIS-C/P; Silverman & Albano, 1996). The ADIS-C/P is a semi-structured interview used in the Kendall et al. study to assess the youth's symptom correspondence to DSM-IV disorders. Doctoral clinical psychology students independently evaluated youth DSM-IV disorders based on separate interviews with the caregivers and the youth, and made ratings on the ADIS-C/P Clinician's Severity. Ratings four and above indicated symptoms in the clinical range. Evaluators were trained to follow ADIS-C/P protocol and achieve and maintain an inter-rater diagnostic reliability of 0.85 (Cohen's K; Silverman & Albano, 1996). Evaluators were blind to the condition of each family (ICBT, FCBT, FESA), and conducted diagnostic interviews before, immediately after, and one year following the intervention. The ADIS-C/P has established reliability (Silverman, Saavedra, & Pina, 2001) and convergent validity with other measures (Wood, Piacentini, Bergman, McCracken, & Barrios, 2002). See Silverman and Albano (1996) for specifics of the ADIS-C/P interviewing procedures and psychometric properties.

Weisz et al. study. ***Children's Interview for Psychiatric Syndromes – Child and Parent Versions*** (CHIPS-C/P; Weller, Weller, Rooney, Fristad, 1999a, b). The CHIPS is a structured interview used in the Weisz et al. study to evaluate the youth's symptoms as they relate to DSM-IV disorders. Evaluators interviewed youth and their parents separately, and combined diagnoses were generated using the Silverman-Nelles procedure (Silverman & Nelles, 1988; cf. Weisz et al. 2012); all diagnoses produced by the youth's report were accepted if they were internalizing (i.e., anxiety and depression) and all diagnoses developed by the parent's report were accepted if they were externalizing (i.e., noncompliance). Evaluators were blind to youths' treatment condition (SMT, MMT, UC), and conducted diagnostic interviews before and immediately after the trial. The CHIPS-C/P has demonstrated reliability and validity in both outpatient and inpatient samples (Weller et al., 1999a, b). See the eAppendix (www.archgenpsychiatry.com) for interviewing procedures, training, and diagnostic reliability.

Weisz et al. study. ***Top Problem Assessment*** (TPA; Weisz et al., 2011). The TPA is a brief youth- and parent-report instrument that assesses the severity of the youth's top three problems independently deemed most important. Ratings of problem areas are made on a scale of 0 = *not a problem at all* to 10 = *very, very much*. Overall, the TPA has demonstrated reliability, validity, and sensitivity to change (Weisz et al., 2011).

Weisz et al. study. ***Brief Problem Checklist*** (BPC; Chorpita et al., 2010). The BPC is a 12-item instrument that assesses a youth's internalizing, externalizing and total problems, developed through item response theory and factor analysis with data from the Youth Self Report and Child Behavior Checklist (see below; Achenbach, 1991). This instrument was administered over the phone to both youths and caregivers. The BPC has established reliability and validity, and has predicted youth's symptom change during treatment (Chorpita et al., 2010).

Child Behavior Checklist for Ages 6-18 (CBCL/6-18; Achenbach, 1991). The CBCL/6-18 is a 113-item checklist designed to assess a wide range of emotional and behavioral problems and was collected in both Kendall et al. and Weisz et al. studies. For this instrument, parents report the extent to which their child displays various behaviors in the past six months by circling 0 (*not true*), 1 (*somewhat/sometimes*), or 2 (*very/often true*). The instrument generates *T* scores that represent a youth's symptom level relative to others of the same gender and age. Respondents with a *T* score of 65 or above on the broadband scales may be in need of treatment.

This instrument includes eight narrowband subscales (Anxious/Depressed, Withdrawn/Depressed, Somatic Complaints, Social Problem, Thought Problems, Attention Problems, Rule-Breaking Behavior, and Aggressive Behavior) and three broadband scales (Total, Internalizing, and Externalizing). Overall, the CBCL/6-18 has evidenced validity, internal consistency, and test-retest reliability, and is widely administered in youth mental health treatment research (Achenbach & Rescorla, 2001). Given the focus on youth anxiety disorders, the current study will use the narrowband Anxious/Depressed subscale and broadband

Internalizing and Externalizing subscales to describe the level of symptomatology across the samples.

Assessment Procedures

Kendall et al. study. If a youth's family agreed to participate in the Kendall et al. parent study, the parent and youth were randomly assigned to a condition (ICBT, FCBT, FESA) using a random-number generated schedule. The schedule used a restricted randomization approach to ensure a similar representation of participants across conditions. Cases were randomly assigned to therapists to control for potential therapist variables (i.e., both therapists and clients were randomly assigned to condition), and therapists only delivered one of the treatments in the study. The ADIS-C/P and CBCL were administered pre-treatment, post-treatment, and at a one-year follow up, and coordinators ensured that post-treatment evaluators were blind to condition.

Weisz et al. study. If a youth's family agreed to participate in the Weisz et al. parent study, the parent and youth were randomly assigned to a condition (MMT, SMT, UC) using block randomization stratified by therapist's education level (i.e., doctoral versus master degree). Therapists only delivered one of the treatments in the study. The TPA and BPC were administered at seven time points: baseline, 3 months, 6 months, 9 months, 12 months, 18 months, and 24 months following study enrollment. The CHIPS and CBCL were administered at baseline and at post-treatment.

Coding Instruments

Cognitive-Behavioral Therapy Adherence Scale for Youth Anxiety (CBAY-A; Southam-Gerow et al., 2016) is a 22-item instrument that assesses the extent that therapists adhere to the ICBT protocol for youth anxiety. There are four different coding dimensions: (a) *Standard*, four items that represent standard ICBT interventions (e.g., *Homework Review*), (b)

Model, 12 items with ICBT model-specific interventions (e.g., *Exposure*), and (c) *Delivery,* six items that capture how model items are delivered (e.g., *Modeling*). Coders watch or listen to entire sessions and rate each item on a seven-point extensiveness scale (Hogue et al., 1996): 1 = *not at all*, 3 = *somewhat*, 5 = *considerably*, 7 = *extensively*. Coders assess two qualities to make the ratings: frequency and thoroughness. Frequency refers to how often each therapeutic intervention is delivered over the course of a session. Thoroughness, on the other hand, refers to the depth that the therapeutic intervention is delivered. For example, an intervention would be considered thorough if a therapist delivered the intervention throughout a session and did not get distracted. Coders considered both components in tandem to decide on an extensiveness rating. For this measure, Southam-Gerow and colleagues (2016) found that ICC inter-rater reliability averaged 0.77 (*SD* = 0.15) at the item level and item scores demonstrated evidence of convergent and discriminant validity; item scores also discriminated between therapists delivering ICBT across research and practice settings from therapists delivering usual care.

Cognitive-Behavioral Therapy Competence Scale for Youth Anxiety (CBAY-C; McLeod et al., 2016) is a 23-item instrument with the same structure as the CBAY-A (i.e., *Standard*, *Model*, *Delivery* dimensions). Coders watch entire audio or video sessions and rate items on a seven-point competence scale (Carroll et al., 2000; Hogue et al., 2008): 1 = *very poor*; 3 = *acceptable*; 5 = *good*; 7 = *excellent*. For example, an intervention would be considered excellent if the therapist delivers an intervention skillfully and with high responsiveness to the client's individual needs. ICC inter-rater reliability for the item scores averaged 0.68 (*SD* = 0.11). Commensurate with the CBAY-A, the CBAY-C item scores evidenced convergent and discriminant validity. More importantly, the CBAY-C item scores showed moderate overlap with the CBAY-A items scores (Mean *r* = .43; *SD* = .20) indicating that the measures capture different aspects of treatment integrity (see McLeod et al., 2016).

Therapy Process Observational Coding System for Child Psychotherapy – Revised Strategies Scale (TPOCS-RS; McLeod et al., 2015). The TPOCS-RS is a 42-item measure that includes five subscales from different theoretical domains (Cognitive, Behavioral, Psychodynamic, Family, Client-centered) and eight items that reflect interventions that are considered integral to typical treatment, but are not related to a particular theoretical domain (e.g., *Rehearsal*). Similar to the CBAY-A, coders observe entire sessions and give therapeutic intervention ratings based on two qualities: frequency and thoroughness (McLeod & Weisz, 2010). Frequency represents how often each intervention is delivered during a session. Thoroughness, in contrast, depicts the complexity or diligence with which the intervention is delivered. Coders consider both components together to decide on a seven-point extensiveness rating with the anchors: 0 = *not at all,* 4 = *considerably*, 7 = *extensively*. For example, if an intervention was delivered at a low frequency, but with great diligence, the coder might code that intervention item as a four or five in extensiveness. The TPOCS-RS and variants thereof have established item inter-rater reliability ranging from .71 to .86 (*M* ICC = .81) in previous studies, and the item and subscale scores provide evidence of construct validity across research and practice settings (McLeod et al., 2015; McLeod & Weisz, 2010; Southam-Gerow et al., 2010; Weisz et al., 2009; Wood, Piacentini, Southam-Gerow, Chu, & Sigman, 2006).

Vanderbilt Therapeutic Alliance Scale-Revised-Short Form (VTAS-R-SF; Shelef & Diamond, 2008). The VTAS-R-SF consists of five items that assess affective aspects of the client–therapist relationship and the extent to which they agree on the client's problem and goals. Coders observe entire sessions and rate items on a six-point scale ranging from 0 (*not at all*) to 5 (*a great deal*). For example, one item focuses on the extent that the client acts in a mistrustful or defensive way towards the therapist, and may be rated high if the client refuses to speak or says something such as “I don’t want to talk about that.” Average ICC inter-rater reliability of items ranged from .72 to .87 (*M* ICC = .80), internal consistency was .90, and convergent validity with the full VTAS form was $r = .94$, $p < .001$ (Shelef & Diamond, 2008). Moreover, the instrument demonstrated predictive validity with outcomes (e.g., higher therapist-adolescent alliance predicted lower cannabis use at three month follow-up; Shelef & Diamond, 2008).

Summary of the Two Parent Studies

Kendall et al. study. Therapists and youth participants were randomized to ICBT, FCBT, or FESA in a university lab. Therapists in both ICBT and FCBT conditions used the *Coping Cat* protocol for youth with anxiety (Kendall & Hedtke, 2006); however, therapists in the FCBT condition also were trained to use techniques aimed to alter parental beliefs, distress, and communication skills. Parents attended two sessions in ICBT and were considered “collaborators;” in contrast, parents in FCBT attended all but two sessions and were considered “co-clients” (Kendall et al., 2008). Therapists in the FESA condition were trained to use a protocol for family education, support, and attention for youth with anxiety; both parents and youths attended all sessions. Findings indicated that principal diagnoses were no longer present post-treatment for 64.00% of clients in ICBT, 64.00% of clients in FCBT, and 42.00% of clients in FESA. ICBT and FCBT also resulted in better post-treatment symptom reduction outcomes than FESA, and these outcomes persisted at the one-year follow up. Overall, the ICBT and FCBT conditions produced similar outcomes, indicating that greater parental involvement in CBT for anxiety may not increase positive outcomes. This study also suggests that ICBT for youth anxiety outperforms FESA and comparable education/support programs.

Weisz et al. study. Therapists and youth were assigned to SMT, MMT, or UC in community clinics. Therapists in SMT used *Coping Cat, PASCET*, or *Defiant Children* protocols, depending on the youth's presenting problem (anxiety, depression, and noncompliant behaviors, respectively). In comparison, therapists in MMT used the *MATCH* protocol regardless of the youth's presenting problem. As described above, the main difference between SMT and MMT was that the MMT protocol allowed for more flexibility; therapists could tailor the content and sequence of interventions to meet client's unique presenting issues. Therapists assigned to UC used the interventions they regularly used and believed to be effective. Due to the broad spectrum of disorders included in this study, diagnostic differences were compared using the number of diagnoses at post-treatment. Results indicated that youths in MMT (M = 1.23, SD = 1.01) had significantly fewer diagnoses than did youths in UC (M = 1.86, SD = 1.52); no significant differences were found between SMT (M = 1.54, SD = 1.30) and UC or SMT and MMT. In addition, MMT produced significantly steeper trajectories of improvement than SMT and UC on weekly symptom measures, whereas SMT did not differ from UC in symptom improvement. Overall, MMT outperformed SMT and UC on a multitude of client outcome measures, suggesting that a modular approach, through its flexible design, may optimize the delivery of evidence-based treatments to youth receiving care in practice settings.

The Treatment Integrity – Efficient Scale for CBT for Youth Anxiety (TIES-CBT-YA)

Instrument development. The instrument in this project, the Treatment Integrity – Efficient Scale for CBT for Youth Anxiety (TIES-CBT-YA), was developed with the assistance of experts in ICBT for youth anxiety (Drs. Michael A. Southam-Gerow and Bryce D. McLeod), using the items and subscales from the previously mentioned instruments that were developed for full session observations (i.e., the parent instruments). The aim was to include items that aligned with Schoenwald's (2011) characteristics of an efficient instrument. Items were retained based on clinical utility (i.e., helpfulness in clinical contexts) and ease of administration and use (Schoenwald et al., 2011). The final TIES-CBT-YA instrument included 12 items extracted from the four parent instruments (list instruments), using the same scaling system as the parent items. See the Appendix for the TIES-CBT-YA code sheet.

First, I selected items from the CBAY-A (Southam-Gerow et al., 2016) and the CBAY-C (McLeod et al., 2016) to depict adherence and competence in ICBT for youth anxiety delivery. One challenge in this stage of instrument development was that there are a number of skill-based ICBT interventions (e.g., *Relaxation*) that are only used in one or two sessions. Thus, to shorten the instrument without eliminating valuable interventions, adherence and competence scales were reduced into two items each, assessing the adherence and competence of both skill-building and exposure (*Skill-building - Adherence, Exposure - Adherence, Skill-building - Competence, and Exposure - Competence*). These items met the conceptual selection criteria of clinical utility, as they could be used across multiple sessions of ICBT (i.e., rather than in just a few sessions that focus on specific skill-building techniques). They also had strong empirical evidence as adherence and competence subscale scores on the full session instrument had strong inter-rater reliability (e.g., ICCs = .77, .68, respectively), and the TIES-CBT-YA scores could feasibly be compared to these subscales. Moreover, they were easy to use, given that these two items represented the two main phases in which therapists receive intensive training and were two subscales from the parent instruments.

Next, I determined how to efficiently represent differentiation items, or items that would depict the extent that other (i.e., non-ICBT specific) interventions were delivered in the session. The TPOCS-RS was well-suited for this purpose because it contains general focus items for each of the five theoretical orientations (Cognitive, Behavioral, Psychodynamic, Family, and Client-Centered). These items require coders to consider the extent that all interventions within those domains are employed when they are making an overall rating (thus, they are similar to subscale scores, but do not require further calculations). I extracted the TPOCS-RS interventions not captured by ICBT adherence and competence scores (i.e., Psychodynamic, Family, and Client-centered general focus items) since the Cognitive and Behavioral general focus items were already represented by the above items. Thus, three differentiation items were included in TIES-CBT-YA (*General Psychodynamic Focus*, *General Family Focus*, and *General Client-Centered Focus)*.

Finally, I selected items from the Vanderbilt Therapeutic Alliance Scale-Revised Short Form (VTAS-R-SF) that would represent the relational bond between the child and therapist, or the alliance. Unlike the more technical items of adherence, competence, and differentiation, these items captured intangible, affective elements of the session, such as nonverbal cues, tone, and rapport building tasks. Given that the VTAS-R-SF instrument was a shortened version of a longer scale, it was appropriate to include all five, unique items. Moreover, the VTAS-R-SF has strong psychometric properties (e.g., *M* ICC = .80; Shelef & Diamond, 2008). Thus, the entire five items were included in TIES-CBT-YA, and a subscale could be calculated by taking the average of all items (with one reverse score on VTAS-R-SF item 3). The following items were included: (a) *She indicates the therapist as understanding and supporting her* (understood), (b) *Seem to identify with the therapist's method of working, so that he assumed part of the therapeutic task himself* (assumes task), (c*) Act in a mistrustful or defensive manner toward the therapist* (mistrust), (d) *Share a common viewpoint about the definition, possible causes, and potential alleviation of the adolescent's problems* (agree on problems), and (e) *Agree upon the goals and tasks for the session* (agree on goals). For the current study, the subscale score for these items was calculated in the same way that was on the original VTAS-R-SF, and was the only item included in the analyses that represents alliance (labeled *VTAS-R-SF*).

Manual development. Once items were chosen for the instrument, the TIES-CBT-YA Scoring Manual was developed using the parent instrument coding manuals as guides. Item scales, definitions, and examples of each code were included. In addition, coders were trained to avoid common coder biases (e.g., coding something when you think it is about to happen, but it has not happened yet) and external influences (e.g., refraining from multitasking). Coders were instructed to watch the exact middle five- or 15-minute portions of the sessions; their beginning and end times were pre-calculated based on the start time and duration of the session, to avoid coder calculation or observation errors (e.g., watching the wrong time segment). Coders were asked to take notes on each code every two and a half minutes to help them make extensiveness (adherence and differentiation) and competence ratings.

Coding and Session Sampling Procedures

Coders. Given that I aimed to assess the extent that efficient instrument scores compared to full set instrument scores, I used coding procedures that matched those of the parent studies, to minimize coder effects. Thus, coders were four clinical psychology graduate students who were blind to condition and coded sessions independently. Coders were assigned recordings in a randomly assigned order (i.e., not in order of session number) and randomly assigned duration (e.g., not all five-minute segments followed by fifteen-minute segments). To further control for coder effects, 10% of all recordings (*N* = 76) were double coded (all four coders overlapped). To eliminate practice effects, coders did not code both the five- and 15-minute segments of a given session.

Coder training. Coders trained on the TIES-CBT-YA Scoring Manual. Training entailed a thorough reading of the scoring protocol, coding practice sessions, and attending regular training meetings. I further developed the scoring manual with the input of the three doctorate student coders and myself (the fourth coder) during training. Coders trained on the scoring manual for one to two hours weekly for approximately six months, from August 2015 to January 2016. Certification coding was required before coders could begin coding sessions for the current data. I aimed to have coders reach a reliability level that is considered adequate for each item (ICC (2,2) of at least 0.59 (Cicchetti, 1994) on a set of 40 full session criterion recordings that were consensus coded by the author and previously coded by two experts in the field. All coders were certified on all items except *General Psychodynamic Focus*, which maintained agreement in the "poor range," throughout training and final certification (ICC = .15) due to the limited variability and low detection frequency. Given the exploratory nature of this study, I determined that this item should be maintained in the analyses to ascertain whether the inter-rater reliability improved with more observations; I moved forward with the understanding that this item may need to be removed from future analyses and studies. After coders were certified for coding, recordings were randomly assigned.

Sampling of treatment sessions. Four sessions were coded from each client in all four conditions. Two sessions that included skill-building content or were conducted early in treatment (i.e., sessions 1-7) were coded and two that included exposure content or were conducted late in treatment (i.e., sessions 8 and over) were coded. If clients had fewer than four sessions, all available sessions were coded. Four sessions were sampled because these allowed for analyses of two time points within two distinct phases of ICBT (i.e., two sessions in skill-building and two sessions in exposure); thus, there would be sufficient data for assessing sensitivity to change in both phases. Including four sessions from both phases of treatment in the analyses allowed me to assess the extent that the efficient instrument mean and slope scores converged with full set mean and slope scores. In addition, by sampling similar session numbers for each client, I controlled for session order/treatment content effects. Two different observation lengths from each session (i.e., five- and 15-minutes) were sampled. Coding two observation lengths for each client (rather than observation lengths of different clients) allowed me to control for coder effects. Sessions were excluded from analyses if the therapist was out of the room for less than 80% of the coded session (e.g., less than twelve-minutes of a 15-minute session and one-minute of a five-minute session). Sessions were randomly assigned to coders using incomplete block design; this design was selected because it accounts for the heterogeneity of the four coders assigned across two time segments (Fleiss, 1981). I did not include parent only sessions (as indicated by the full session data) in my sample, as these may not have included core ICBT content. Although parent-only sessions were excluded from the full session sampling, if there were only parents and a therapist present during the middle five- and 15-minute segments, the session was still included. For 13 five-minute segments and 9 15-minute segments, only parents and a therapist happened to be present in the segment coders observed. These sessions were included to fully characterize the middle portion of ICBT (especially differentiation scores; e.g., the extent to which family interventions were delivered).

Study coding. Coders aimed to code for two hours per week from January 2016 to August 2016. Coders engaged in rating less than five hours a day and took frequent (e.g., every two hours) breaks to minimize fatigue. During coding, regular meetings were held and inter-rater reliability was regularly assessed to prevent coder drift. Coding lasted approximately seven months and coders each attempted to code a total of 228 or 229 session observations (approximately half five-minute segments and half 15-minute segments). A total of 756 observations (377 sessions with both observations) were coded and included in the primary analyses (76 observations were double coded for reliability purposes and 80 observations could not be coded due to the disk freezing [10 observations] or the therapist being out of the room for longer than 80% of the session [70 observations]). A main coder was identified and TIES-CBT-YA item scores were derived from her scores from each session.

Data Analysis Plan

Preliminary analyses. First, preliminary analyses were conducted to ascertain that the sample coded accurately represents the parent samples, to ensure that findings could generalize to that sample (CS vs. CS-parent sample; Kendall vs. Kendall-parent sample). These analyses focused on demographic, clinical (i.e., pre-treatment symptom scores), youth, and therapist characteristics. Of note, post-treatment symptoms scores were unavailable for youth not included in this study so post-treatment symptoms were not able to be compared across parent and current samples. Unpaired t- and Pearson chi-square tests were conducted for continuous and categorical variables, respectively. I also assessed whether there were any significant youth demographic, therapist demographic (sex, ethnicity, theoretical orientation), youth baseline and post-treatment clinical, and session differences across treatment groups, as a randomization check. One-way ANOVAs were conducted for continuous measurements and Pearson chi-square tests were conducted for categorical measurements.

Treatment integrity scoring approach. For the main analyses, I assessed how the reliability and validity of the newly developed TIES-CBT-YA instrument scores compared to scores on the full TIMS instruments. Full session treatment integrity subscale scores of both the (a) four sessions per client selected for this study (hereafter referred to as the full session, four set scores) and the (b) entire sessions per client (e.g., 16-23 sessions) coded in TIMS (hereafter referred to as the full session, entire set scores) were calculated. These two sets of data were used to assess which TIES-CBT-YA scores were most similar to the full-length scores of the *study sample* and which TIES-CBT-YA scores best represent the full length scores of the *TIMS sample*, respectively; both were included to enrich the information available for future researchers or clinicians interested in different aspects of the instruments (e.g., those interested in sensitivity to change for supervisory feedback purposes). The full session, four set scores were considered the "gold standard" comparator from which the TIES-CBT-YA performance was interpreted because these sessions best captured the full session scores for the study sample and focused the results on the intended independent variable, session length. For example, if I had used the full session, entire set scores as the gold standard comparator, this would make interpreting findings difficult (e.g., are the scores discrepant because of the length of the session or because of the number of sessions?). Importantly, the same clients were used across full and brief session scoring strategies (i.e., no clients were excluded from the full sample in the development of TIES-CBT-YA). Scores were kept at the session level for all analyses except the predictive validity analyses (due to the nature of the TIES-CBT-YA scores being the dependent variable at the session level and CBCL scores being the dependent variable at the client level). For the ICBT instrument items (i.e., adherence and competence items), full instrument subscales were scored by taking the maximum score within each subscale. This strategy was adopted because it is expected that therapists would only deliver one or two interventions in the middle five- or 15-minutes of the session; taking the average of items in each subscale of the full session would likely misrepresent the core one or two interventions delivered in the middle of a session (i.e., the recommended intervention for ICBT treatment). Of note, when competence was not coded, the competence items were considered missing (rather than scored as a zero, per the scale), to appropriately depict the variance in this instrument. For the TPOCS-RS (differentiation) items, the full session *General Focus* items were used. These items had already been scored as independent items on the TPOCS-RS, by coders who considered the general extensiveness of all items within a subscale while making a rating (thus no subscale calculations needed to be made and they could be directly compared to the TIES-CBT-YA items). This strategy was used to capture any potential non-ICBT interventions used within five- and 15-minute time segments. Finally, the average of the items was used for the VTAS-R-SF (alliance) scale.

Overview of primary data analytic approach. Data analyses were conducted in stages. First, I determined which observational length (i.e., five- or 15-minute segments) had an (a) item distribution and (b) inter-rater reliability that best approximated both sets of full session scores using descriptive and ICC analyses, respectively. I also assessed which observation was the best combination of (c) sensitivity to change and (d) congruence with both full set scores using hierarchical linear modeling (HLM) analyses and Pearson correlation coefficients, respectively. Finally, I assessed which observational length (e) discriminated between treatments and (6) predicted improvements in client symptom outcomes that were consistent with both full set scores using HLM. One session recording was found to be a duplicate of another session after all analyses were completed. This session was retained in all sample analyses because the change in means for all TIES-CBT-YA items, for both five- and 15-minute segments, was less than .01 when the session was removed. For all the validity analyses, three treatment integrity (CBAY-A, CBAY-C, TPOCS-RS), one alliance (VTAS), and one outcome (CBCL [internalizing, externalizing, and anxious/depressed subscales]) instrument were used.

Observation length performance. I assessed the performance of each observation length by comparing distribution patterns, Pearson Correlations, and/or statistical significance of TIES-CBT-YA scores to both full session set scores across the different reliability and validity analyses. Cohen's *d* effect sizes were included in the discriminant validity analyses due to research suggesting that statistical significance (i.e., *p* values) data are important, but not sufficient when interpreting group comparison results (APA, 2001; Cumming & Finch, 2005). More specific guidelines for assessing the five- versus 15-minute treatment integrity score performance depended on the type of analysis and are described at the introduction of each separate analysis below. Full session, entire set scores were included in the comparison analyses to provide a more comprehensive picture of how TIES-CBT-YA scores *did* or *did not* approximate scores from the full course of treatment coded in TIMS.

Item distribution. Once the instrument was developed and all sessions were coded, I assessed whether each dependent variable (all TIES-CBT-YA items) met assumptions of normal distribution for full session scores and brief segment scores (See Table 2 for item distribution data across observation time points); then, I compared these findings in a systematic manner. Researchers have posited that values of skewness/kurtosis that are less than -1.5 and greater than 1.5 should be considered outliers, and those that are less than 2.25 and greater than 2.25 have a high degree of heterogeneity (Blanca, Arnaue, Lopez-Montiel, Bono, & Bendayan, 2013). Thus, I assessed if (a) the item fell within a standard deviation of the mean for the full session, four set score, (b) had the same range of the full session, four set score, (c) fell within the standard error of skewness and kurtosis of the full session, four set score, and (d) was not considered skewed or kurtotic by statistical researchers (i.e., less than -1.5 and greater than 1.5).

Reliability. A total of 9.79% of sessions were double coded for reliability analyses (five-minute segments = 10.07% of sessions, 15-minute segments = 9.81% of sessions). One way random single-measure Intraclass Correlation Coefficients, ICC(1, 1), were calculated for the double-coded (a) full session, entire set scores, (b) full session, same set scores (i.e., the same sessions selected for the ICC analyses for the brief segments; 39 sessions each), (c) five-minute scores, and (d) 15-minute scores (See Table 3). Following Cicchetti's guidelines (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. Each observation segment was assessed on the reliability of its scores. I note (a) if the item fell within the confidence interval of the ICC for the full session, same set score (i.e., to assess similarity to full session score) and (b) the category of the agreement strength (i.e., to depict the reliability within the instrument for that segment).

Sensitivity to change. To assess sensitivity to change (i.e., how therapist behavior changes over the course of treatment), I used multilevel modeling (Raudenbush & Bryk, 2002) with HLM 7.01 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011) to account for the nesting of (a) sessions within clients and (b) clients within therapists. Next, I determined the level of analysis for each of the treatment integrity subscale scores based on the variance found at each level in the model: sessions (level 1) nested in clients (level 2), and clients nested in therapists (level 3). Then, time (e.g., weeks in treatment, weeks in treatment squared) was entered in level 1 of each model. Unconditional three-level growth models were fit to each TIES-CBT-YA item to examine whether change was linear or quadratic. Over 98 percent of the variance in all item quadratic slope scores occurred at the client level for all unconditional means models. Thus, all effects were set to random at the client level and fixed at the therapist level for quadratic slopes. Time was centered at the earliest session to assess differences at the beginning of treatment.

Scores were calculated for each treatment integrity item at both observation segments to determine which segment was most sensitive to change. The -2 Log Likelihood (-2LL) statistic was referenced to determine how much unexplained variance was left after each model was fit to the data; if the difference between two models (i.e., unconditional and linear, linear and quadratic) had a significant reduction in -2LL (i.e., χ^2 $p < .05$), adding time was considered an improvement in model fit. Even if the model did not warrant adding time as a predictor (due to χ^2 not being significant at the 0.05 significance level), all statistics were reported at the base quadratic model to better compare statistical significance across items (See Table 4). When comparing the brief segments to the full session, four set scores, I describe if the TIES-CBT-YA items (a) fit the same growth level (i.e., unconditional, linear, quadratic models) as the full session, four set scores according to -2LL statistics, (b) had the same significance (i.e., significant or not significant) on linear and quadratic slopes and (c) had similar directions of slope change.

Convergent validity. I assessed how well the subscales assessed the criterion they intend to assess (i.e., criterion validity), by examining the degree to which the scales of the efficient instrument were associated with scales in which they should be theoretically similar (i.e., convergent validity; Kazdin, 2003). Pearson correlation coefficient *r* analyses were run with each TIES-CBT-YA item and each full session, four set item (See Table 3); when interpreting *r* correlation coefficients, .10-.23 is considered "small," .24-.36 "medium," and greater than .36 "large" (Rosenthal & Rosnow, 1984). I assessed each TIES-CBT-YA item across observation segments based on how well they fit into these categories to ascertain how closely each TIES-CBT-YA item fell into categories that paralleled those within the full session scores.

Discriminant validity. The above HLM analyses were used to assess discriminant validity, or analyses that investigated whether five- or 15-minute TIES-CBT-YA segments can differentiate between treatments in a pattern that mirrored full session scores; this assessed whether there was an absence of association between two unrelated constructs (i.e., CBT vs. UC treatments; DeVillis, 2012). HLM models were run for each brief observation segment and treatment integrity scale, including the full session, entire set and the full session, four set scores. Treatment terms were entered at the client level in all unconditional means models (See Table 4 for all results); thus there were six two-group comparisons for each item. Given that therapists in CS-UC were not instructed or trained to deliver skill-building and exposure interventions, one prior expectation for group comparison analyses was that therapists in the manualized treatment groups would deliver significantly higher doses of skill-building and exposure interventions than therapists in CS-UC. For analogous reasons, therapists in the manualized treatment groups were expected to deliver skill-building and exposure interventions in a more skillful and responsive manner than therapists in CS-UC. Cohen's *d* effect sizes were run to determine the magnitude of the group differences; Cohen's *d* effect sizes were calculated by dividing the parameter estimate by the raw data standard deviations for each group comparison (Feingold, 2009). Cohen (1988) offered interpretations of *d* = 20 as "small" effect size, *d* = .50 as "medium" effect size, and *d* = .80 as "large" effect size. Effect sizes further informed how the brief segments compared to the full session scores; for example, I could confirm that the brief segment findings that were significant had at least small to medium effect sizes and that non-significant findings had low effect sizes (i.e., avoiding Type I error and Type II error, respectively). *Exposure-Competence* could not be run in HLM due to the very low rate of scores in the CS-UC group across observation segments; this item was analyzed differently and reviewed in the item-differences section. After presenting findings for the full session sets, I characterize each TIES-CBT-YA segment by comparing how the statistical significance and effect sizes for each item correspond to the full session scores. Only statistically significant results are reported in the text to enhance clarity.

Predictive validity. Predictive validity analyses were intended to assess if TIES-CBT-YA scores predicted client outcomes and if they predicted outcomes as well as the full-scale scores. Symptom variables consisted of CBCL Internalizing, CBCL Externalizing, and CBCL Anxious-Depressed subscale scores collected at baseline and post-treatment. First, I determined if there was sufficient pre- to post-treatment change in symptoms to predict by calculating d-type effect sizes for each CBCL subscale (Cohen, 1988). Clients' reduction in CBCL internalizing symptoms from pre- to post-treatment fell within the "large" range ($d = .99$) and their reduction in externalizing and anxious/depressed symptomology fell within the "medium" range ($d = .52$ for both). Thus, it was worth analyzing how each TIES-CBT-YA item contributed to the reduction in client symptoms. Using HLM, I computed a trajectory for each dependent variable. Each efficient treatment integrity item and both sets of full session items were entered into separate models as predictors of each dependent variable. Baseline symptom variables were entered as control variables. Over 98.0% of the variance in all item scale scores occurred at the session level. Thus, only two-level HLM models were run for predictive validity analyses, of aggregated sessions nested within clients. Next, item distributions were evaluated at the client level to determine if the data met assumptions of normality. One client (0.01%) had missing CBCL pre- and post-treatment data and 10 clients (9.43%) had missing CBCL post-treatment data. Clients also had missing treatment integrity data; in the full session, entire set data, two items (*Skill-building-Competence* and *Exposure-Competence)* had missing data (6.60% and 26.42% of clients had missing data, respectively; 4.72% and 22.64% of clients had missing treatment integrity data, but not missing CBCL data, respectively). The amount of missing treatment integrity data also differed across studies (with clients from Child Steps having more missing data than clients from Kendall et al.). The same pattern of missing data across items held for full session, four set scores and both brief segment scores. Past researchers have found that having more than 10.0% missing data may bias results (Bennett, 2001). Missing data were not imputed due to the possibility of the data being missing not at random (Little's MCAR test Chi Square = 78.52, *df* = 33, *p* < .001; e.g., the possibility that parents did not report their child's post-treatment CBCL symptoms due to no symptom change or poor engagement; Finch, 2010). On account of the significant missing data across both independent and dependent variables, predictive validity analyses could not be run in this study.

Results

The goal of this pilot study was to take a step toward developing an observational treatment integrity instrument that strikes a balance between effectiveness (i.e., reliability, sensitivity to change, convergent validity, discriminant validity, predictive validity) and efficiency (i.e., brief, easy to administer and use; Schoenwald et al., 2011). I focused on determining if reliability and validity evidence could be maintained when the coding process was shortened, so that the instrument can be field tested in the future. Specifically, the current study assessed whether it is possible to reduce the amount of time that coders spend making treatment integrity ratings of therapists delivering individual cognitive-behavioral treatment (ICBT) for youth anxiety in both research and practice settings. The efficient observational treatment integrity instrument was derived from four observational treatment integrity instruments for youth anxiety (see McLeod et al., 2015; McLeod et al., 2016; Shelef & Diamond, 2008; Southam-Gerow et al., 2016) and assessed four different treatment integrity components (adherence, competence, differentiation, and alliance).

Hypotheses

A primary aim was to develop an instrument that represents four treatment integrity components and has specific characteristics. For one, I hypothesized that the final efficient instrument would have the full range in treatment integrity scores (e.g., range of 1-7). Second, I expected that the efficient instrument would produce scores (for both observation segments) that were strongly associated with scores from the full session treatment integrity items. Third, I hypothesized that each efficient treatment integrity item would discriminate treatment integrity across treatment settings and treatment types (ICBT, UC). No hypotheses were made about which session observation segment (five- or 15-minutes) would perform best across the various psychometric properties. Rather, a series of research questions were posed for each set of analyses, to help characterize findings for researchers and clinicians with disparate interests in field-testing this kind of instrument in the future.

Preliminary Data Analyses

Sample comparison. The current sample drawn from the Child Steps study (n = 55) significantly differed from the Child Steps parent sample (n = 68) on family income (χ^2 = 5.01, p = 0.025) and therapist sex (χ^2 = 28.12, p < .001). Ten of the 12 participants excluded from the current study had household incomes less than 60,000 and eight of the 12 had missing therapist sex information. The current sample drawn from the Kendall sample did not significantly differ from the parent sample on any variables. Importantly, all of the same clients and therapists included in the studies that coded the full treatment sessions and produced full-scale treatment integrity scores (e.g., McLeod et al., 2016; Southam-Gerow et al., 2016) were present in the current study (i.e., both included the same 106 participants).

Treatment group comparison. One-way ANOVAs and Pearson chi-square tests indicated that the four treatment groups differed on child ethnicity, level of externalizing symptomatology at baseline, level of anxious and depressed symptomatology at baseline, treatment duration, number of sessions coded, and number of sessions held (See Table 2 for more details). All other variables did not significantly differ across groups.

Table 2

Youth Demographic and Clinical Information by Treatment Group, M (SD) or %

Variable	K-SMT	CS-SMT	CS-MMT	CS-UC
	(*n* = 51)	(*n* = 22)	(*n* = 16)	(*n* =17)
Age	10.36 (1.90)	9.77 (1.51)	9.93 (1.87)	10.00 (1.87)
Sex – Male	60.80	50.0	56.30	58.80
Race				
Caucasian	83.30[c]	72.70	43.80	41.20
African-American	9.80	0.00	12.50	5.90
Latino	2.00	4.50	0.00	5.90
Mixed/Other	2.00[b]	25.80	43.80	47.10
Total baseline diagnoses	3.02 (1.45)	3.14 (2.17)	2.63 (2.06)	3.76 (0.66)
CBCL				
Total	63.10 (9.31)	65.27 (7.49)	63.63 (10.39)	66.35 (5.23)
Internalizing	67.40 (8.37)	70.00 (6.72)	69.56 (9.33)	68.82 (5.68)

Externalizing	52.96 (10.08)[b]	59.00 (11.27)	55.06 (11.64)	60.18 (8.76)
Anxious-Depressed	63.10 (9.31)	70.09 (10.05)	69.50 (10.07)	69.47 (8.70)
Family Income - Up to 60 k	35.30	54.50	31.30	70.60
Number of Sessions Held	15.92 (1.43)	21.91 (11.17)	20.69 (6.15)	20.87 (11.95)[a]
Number of Sessions Coded	3.24 (0.81)[b]	3.86 (0.35)	4.00 (0.00)	3.76 (0.66)
Treatment Duration	19.52 (3.97)[b]	34.94 (14.17)	33.13 (9.27)	45.38 (33.67)

Note. K-SMT = ICBT condition in the Kendall et al., study, CS-SMT = Child Steps - Standard Manualized Treatment Sample, CS-MMT- Child Steps - Modular Manualized Treatment Sample, CS-UC = Child-Steps Usual Care. a = CS-SMT > K-SMT. b = K-SMT < CS-SMT, CS-MMT, CS-UC. c = K-SMT > CS-MMT, CS-UC.

Primary Data Analyses

Research question one – item distribution. What is the shortest TIES-CBT-YA observation segment that best approximates the item distribution of the full session treatment integrity scores independently archived for TIMS?

Full session scores. For the full session, four set scores (i.e., the "gold standard" comparator), all items met assumptions of normality (i.e., full range, normal distribution) except *General Psychodynamic Focus* (skewness statistic = 5.42, kurtosis statistic = 36.69, range = 1-3.5) and *General Family Focus* (skewness statistic = 2.27, kurtosis statistic = 5.35, range = 1-7), which both fell outside the -1.5 and 1.5 range for skewness and kurtosis. In addition, it is important to note that *Skill-building-Competence*, *Exposure-Competence*, *General Psychodynamic Focus*, *General Client-Centered Focus*, and *VTAS-R-SF* did not have a full range for the full session, four set scores. These same findings of restricted range were consistent across full session analyses. The full session, entire set scores also had the same pattern of skewed/kurtotic distributions across items. For the entire set scores, *General Psychodynamic Focus* (skewness statistic = 3.95, kurtosis statistic = 18.56, range = 1-3.50) and *General Family*

Focus (skewness statistic = 1.59, kurtosis statistic = 1.57) fell outside the 1.5 skewness/kurtosis range. Thus, *General Psychodynamic Focus* and *General Family Focus* appeared to have poor distributions regardless of the number of sessions coded and phase in treatment. In sum, only two TIES-CBT-YA items did not meet assumptions of normality across full session data sets.

Five-minute segments. Each item was assessed using the aforementioned criteria. *Skill-building-Adherence* met assumptions of normality and fell within the standard deviation, range, and skewness of the full session, four set scores; however, kurtosis was above the standard error of the full session, four set skewness scores (skewness statistic = -0.14, kurtosis statistic = -1.54, range = 1-7). *Exposure-Adherence* fell within the standard deviation and range of the full session, four set scores; however, this item was outside the 1.5 skewness/kurtosis range (skewness statistic = 2.67, kurtosis statistic = 5.81, range = 1-7) and did not fall within the skewness/kurtosis standard error range of the full session, four set score. In contrast, *Skill-building-Competence* met assumptions of normality and fell within the standard deviation, range, and skewness/kurtosis range of the full session, four set scores (skewness statistic = -0.33, kurtosis statistic = -0.55, range = 1-7); thus it was similar to full session, four set scores. *Exposure-Competence* comparably met assumptions of normality and fell within the standard deviation of the full session scores; however, it had more range (skewness statistic = -0.31, kurtosis statistic = 0.32, range = 1-7). *General Psychodynamic Focus* was the least similar to full session scores of all the items; although this item fell within the standard deviation of the full session, four set scores, it did not meet assumptions of normality, had more range, and was outside the 1.5 skewness/kurtosis range (skewness statistic = 11.81, kurtosis statistic = 163.84, range = 1-4). *General Family Focus* and *General Client-Centered Focus* were also somewhat discrepant from full session scores. *General Family Focus* fell within the standard deviation and range of the full session, four set scores, but it did not meet assumptions of normality and was outside the -1.5/1.5 skewness/kurtosis range (skewness statistic = 4.41, kurtosis statistic = 20.63, range = 1-7). *General Client-Centered Focus* fell within the standard deviation of the full session scores and met assumptions of normality for skewness/kurtosis; however, the five-minute segments had slightly more range and discrepant skewness/kurtosis statistics (skewness statistic = 0.53, kurtosis statistic = -0.86, range = 1-7). Lastly, *VTAS-R-SF* (alliance) met assumptions of normality and fell within the standard deviation and skewness/kurtosis range of the full session, four set scores (skewness statistic = -0.88, kurtosis statistic = 0.87, range = 0-5); the only difference between the full session and five-minute scores was that the five-minute scores had slightly more range. Overall, the five-minute scores had three items that did not meet assumptions of normality (*Exposure-Adherence, General Psychodynamic Focus, General Family Focus*); all other items fell within the 1.5 skewness/kurtosis range and had means that fell within the standard deviations of the full session scores.

Fifteen-minute segments. In comparison to five-minute scores, *Skill-building Adherence*, *Exposure-Competence, and General Client-Centered Focus* had similar findings for the 15-minute scores; *Skill-building-Adherence* met assumptions of normality and fell within the standard deviation and range of the full session, four set scores; however, skewness/kurtosis in the 15-minute segments was outside the standard error of the full session, four set skewness/kurtosis scores (skewness statistic = -0.41, kurtosis statistic = -1.24, range = 1-7). *Exposure Competence* (skewness statistic = -0.42, kurtosis statistic = -0.26, range = 1-7) and *General Client-Centered Focus* (skewness Statistic = 0.17, kurtosis statistic = -0.97, range = 1-7) met assumptions of normality and fell within the standard deviation of the full session scores; however, they had full range. Unlike in the five-minute segments, even though *Skill-building-*

Competence met assumptions of normality and fell within the standard deviation of the full session, four set scores, it had a full range (skewness statistic = -0.44, kurtosis statistic = -0.23, range = 1-7). In addition, *Exposure-Adherence* (Skewness Statistic = 2.36, Kurtosis Statistic = 4.13, Range = 1-7), *General Psychodynamic Focus* (skewness statistic = 7.10, kurtosis statistic = 53.58, range = 1-3), and *General Family Focus* scores (skewness Statistic = 3.37, kurtosis Statistic = 11.03, range = 1-7) all did not meet assumptions of normality and fell outside the standard error of skewness/kurtosis for full session scores. In general, the same three items did not meet assumptions of normality for 15- and five-minute segments.

Outliers. All items were further examined to ascertain if cases needed to be removed or modified for the remaining preliminary analyses. On account of the data nesting, outliers (greater than 3 standard deviations above the mean) were assessed at the client level. Conservative criteria were used for identifying outliers; clients were considered outliers if they were equal to or above 2.5 standard deviations away from the mean, a cut off point that a number of researchers suggest is appropriate for smaller sample sizes (e.g., Van Selst & Jolicoer, 1994). For the five-minute segments, there was one outlier on *Exposure - Competence* and one on *General Client-Centered Focus*. For the 15-minute segments, there were three outliers on *Exposure-Adherence* and one on *General Client-Centered Focus*. These outliers were retained in the sample due to the exploratory nature of the study, as well as the equal dispersion across observations**.**

Summary-observation segment differences. Overall, both observation segments had normal distributions for the same five items in the full session, four set scores. However, five-minute scores had two fewer client-level outliers on *General Client-Centered Focus* than the 15-minute segment scores. Thus, five-minute segments had a slightly more normal distribution for

General Client-Centered Focus; all other items had the same pattern of distributions across observation segments.

Summary-item differences. Across TIES-CBT-YA observation segments, adherence, competence, client-centered, and alliance items had normal distributions, with relatively full ranges and little heterogeneity of the data. Two treatment differentiation items, *General Psychodynamic Focus* and *General Family Focus* did not meet assumptions of normality, exhibiting highly heterogeneous data across observation segments. The full scale range was captured with *General Family Focus*; in contrast, a limited scale range (1-3;1-4) was captured with *General Psychodynamic Focus*. This pattern of item differences across observation segments paralleled the full session, four set scores; thus, although both observation segments only moderately approximated full scores numerically, they closely approximated full session, four set scores in their patterns of item distributions.

Table 3

TIES-CBT-YA Item Distribution across Observation Samples

Item	*M* (*SD*)	Range	Skewness (*SE*)	Kurtosis (*SE*)
Skill-building - Adherence				
Full session, entire set	3.59 (1.82)	1-7	0.17 (0.07)	-1.20 (0.14)
Full session, four set	4.08 (1.83)	1-7	-0.24 (0.12)	-0.16 (0.24)
Five minute sessions	3.70 (2.10)	1-7	-0.14 (0.13)	-1.54 (0.25)
Fifteen minute sessions	4.16 (2.07)	1-7	-0.41 (0.13)	-1.24 (0.25)
Exposure - Adherence				
Full session, entire set	2.53 (1.82)	1-7	0.74 (0.07)	-0.90 (0.14)
Full session, four set	2.12 (1.81)	1-7	1.25 (0.12)	.008 (0.24)

Five minute sessions	1.50 (1.39)	1-7	2.67 (0.13)	5.80 (0.25)
Fifteen minute sessions	1.57 (1.42)	1-7	2.36 (0.13)	4.13 (0.25)
Skill-building - Competence				
Full session, entire set	4.58 (1.18)	1.5-7	-0.14 (0.09)	-0.82 (0.18)
Full session, four set	4.78 (1.12)	2-7	-0.31 (0.14)	-0.77 (0.29)
Five minute sessions	4.78 (1.20)	2-7	-0.33 (0.15)	-0.55 (0.30)
Fifteen minute sessions	4.80 (1.30)	1-7	-0.44 (0.14)	-0.23 (0.28)
Exposure - Competence				
Full session, entire set	4.63 (1.27)	2-7	-0.41 (0.12)	-0.87 (0.21)
Full session, four set	5.10 (1.16)	2-7	-0.91 (0.23)	0.04 (0.46)
Five minute sessions	4.65 (1.28)	1-7	-0.31 (0.34)	0.32 (0.67)
Fifteen minute sessions	4.14 (1.54)	1-7	-0.42 (0.30)	-0.26 (0.59)
General Psychodynamic Focus				
Full session, entire set	1.09 (0.29)	1-3.5	3.95 (0.07)	18.56 (0.13)
Full session, four set	1.06 (0.25)	1-3.5	5.43 (0.12)	36.69 (0.24)
Five minute sessions	1.02 (0.19)	1-4	11.81 (0.13)	163.84 (0.25)
Fifteen minute sessions	1.03 (0.22)	1-3	7.10 (0.13)	53.58 (0.25)
General Family Focus				
Full session, entire set	1.86 (1.38)	1-7	1.59 (0.07)	1.57 (0.14)
Full session, four set	1.55 (1.04)	1-7	2.27 (0.12)	5.35 (0.24)
Five minute sessions	1.23 (0.88)	1-7	4.41 (0.13)	20.63 (0.25)
Fifteen minute sessions	1.34 (1.04)	1-7	3.37 (0.13)	11.03 (0.25)
General Client-centered Focus				

Full session, entire set	2.90 (0.97)	1-6.5	0.67 (0.07)	0.35 (0.13)
Full session, four set	2.88 (0.95)	1-6.5	0.69 (0.12)	0.58 (0.24)
Five minute sessions	3.12 (1.82)	1-7	0.53 (0.13)	-0.86 (0.25)
Fifteen minute sessions	3.52 (1.67)	1-7	0.17 (0.13)	-0.97 (0.25)
VTAS-R-SF				
Full session, entire set	2.65 (0.75)	0.1-4.4	-0.66 (0.85)	0.07 (0.17)
Full session, four set	2.70 (0.65)	0.7-4	-0.74 (0.15)	0.56 (0.30)
Five minute sessions	3.18 (0.80)	0.6-4.8	-0.59 (0.13)	0.36 (0.26)
Fifteen minute sessions	3.34 (0.87)	0-5	-0.88 (0.13)	0.87 (0.26)

Note. SD = Standard Deviation *SE* = Standard Error

Research question two - reliability. What is the shortest TIES-CBT-YA observation segment that best approximates the reliability evidence of the full session treatment integrity scores independently archived for TIMS?

Full session scores. For the full session, entire set scores, all ICCs fell in the “fair” to “excellent” agreement range (ICC range =.42 - .85). When only the sessions included in the TIES-CBT-YA ICC analyses were analyzed, the full-scale scores did not exhibit the same level of agreement. Although *General Client-Centered Focus* fell in the “fair” range (five-minute set, ICC = .49; 15-minute set, ICC = .51) across sets, *Skill-building-Competence* dropped from the “good” range (ICC = .60) into the “fair” range (ICC = .45) when only the five-minute set of sessions were analyzed. Moreover, three items fell within the “poor” agreement range when the full session, ICC sets were analyzed: *Exposure-Competence* for the five-minute ICC set [(ICC (1, 1) = 0)], *General Psychodynamic Focus* for both the five-minute set (ICC unable to be scored due to no variance) and 15-minute set (ICC = .38), and *VTAS-R-SF* for the five-minute set (ICC

= .39). All other items were consistent across full session sets. Overall, the five-minute ICC set appears to have somewhat less reliable full session scores than the 15-minute ICC set.

Five-minute segments. For the five-minute segment ICCs, *Skill-building-Adherence* (ICC = 0.82) and *Skill-building-Competence* (ICC = .79) fell within the “excellent” agreement range, and both fell within the confidence intervals of the corresponding full session, same set scores; however, *Skill-building-Competence* was two levels of agreement higher in the five-minute segments than in the corresponding full session scores. *Exposure-Adherence* (ICC = .53) fell within the “fair” agreement range and fell below the lower limit of the full session, same set 95% confidence interval. In contrast, *Exposure-Competence* fell within the “good” agreement range and fell within the confidence interval of the full session ICC. It is important to note that *Exposure-Competence* had better inter-rater agreement when coded for the five-minute segment (ICC = .61) than coded for the full segment of the same set of sessions (ICC = 0). *General Psychodynamic Focus* fell within the “poor” range of agreement (ICC = 0), but could not be compared to the full session ICC due to no variance in the full session set. *General Family Focus* fell within the “fair” range of agreement (ICC = .42) and was below the lower limit of the confidence interval of the full session score. Finally, both *General Client-Centered Focus* and *VTAS-R-SF* fell within the “poor” agreement range (ICC = .16 and ICC = .24, respectively). However, *General Client-Centered Focus* ICC fell below the lower limit of the confidence interval of the full session ICC and *VTAS-R-SF* was within the confidence interval of the full session ICC. In sum, five-minute segments had two items that were two agreement categories below the full session, same set scores (and below the lower limit of the confidence intervals), two items that were two agreement categories above full scores (and one above the upper limit of the confidence interval), one that was one agreement category below full scores (and below the lower limit of the confidence interval), and two that fell in the same agreement category.

Fifteen-minute segments. Fifteen-minute segments had three ICC scores that fell within the same agreement categories as the five-minute scores: *Skill-building-Adherence*, *Exposure-Competence, and General Client-Centered Focus. Exposure-Adherence*, *Skill-building Competence*, *General Psychodynamic Focus*, *General Family Focus*, and *VTAS-R-SF* items all differed across five- and 15-minute segments. *Exposure-Adherence* fell within the “good” agreement range (ICC = .63) and was below the lower limit of the confidence interval for the full session score. *Skill-building-Competence* fell within the “good” agreement range and within the ICC confidence interval of the full session score. *General Psychodynamic Focus* fell within the “poor” agreement range (ICC = 0), and fell below the lower limit of the confidence interval for the full session score. In contrast, *General Family Focus* fell within the “good” agreement range and within the confidence interval for the full session score. Lastly, *VTAS-R-SF* fell within the “fair” agreement range and the confidence interval for the full session score. In general, 15-minute segments had one item that fell above the ICC agreement category for the full session, same set scores (but within the confidence interval), four items that fell one agreement category below the full session scores (but within the confidence intervals), two items that fell two agreement categories below the full scores (and below the lower limit of the confidence interval), and one that fell within the same agreement category.

Summary-observation segment differences. Taken together, 15-minute segments had six items that fell within one agreement category of full session, same set scores (and ICCs that fell within the confidence intervals) whereas the five-minute segments only had two items that met this criterion. In comparison to five-minute segment scores, 15-minute segment scores performed slightly better on *VTAS-R-SF* (i.e., one agreement category higher) and significantly better on *General Family Focus* and *Exposure-Adherence* (one agreement category higher and within confidence intervals of the full session scores). Five-minute scores performed slightly better on *Skill-building-Competence* (one agreement category higher). *General Psychodynamic Focus* could not be compared across segments due to no variance in the five-minute ICC set. Overall, the 15-minute segments had more items that better approximated the full session agreement categories than the five-minute segments.

Summary-item differences. Across TIES-CBT-YA observation segments, *Adherence-Skill-building*, *Competence-Skill-building*, and *Competence-Exposure* had ICCs that fell within the "good" to "excellent" agreement range (i.e., ICCs [1,1] = .61-.82). *Exposure-Adherence* and *General Family Focus* had ICCs that fell within the "fair" to "good" agreement range (i.e., ICCs = .42-. 67). *VTAS-R-SF* had ICCs that fell within the "poor" to "fair" agreement range (ICCs = .24-.49). Lastly, *General Client-Centered Focus* and *General Psychodynamic Focus* both had ICCs that fell within the "poor" range. This pattern of ICC differences across observation segments generally paralleled the full session, four set scores; the major differences were that *General Client-Centered Focus* and *General Family Focus* were an agreement level higher for the full session scores.

Table 4

TIES-CBT-YA ICCs across Observations and Correlations with Full Session, Four Set Scores

Item	ICC [95% CI]	*r* [95% CI]
Skill-building - Adherence		
Full session, entire set	.73 [.70,.75]	-
Five minute sessions	.82 [.69,.90]	.65 [.65,.71]***

Full session, same set	.81 [.66,.89]	-
Fifteen minute sessions	.79 [.63,.89]	.72 [.62,.76]***
Full session, same set	.67 [.46,.81]	-
Exposure - Adherence		
Full session, entire set	.79 [.77,.81]	-
Five minute sessions	.53 [.26,.73]	.61 [.58,.76]***
Full session, same set	.90 [.81,.95]	-
Fifteen minute sessions	.63 [.39,.79]	.76 [.76,.91]***
Full session, same set	.83 [.70,.91]	-
Skill-building - Competence		
Full session, entire set	.60 [.46,.70]	-
Five minute sessions	.79 [.63.89]	.40 [.27,.52]***
Full session, same set	.45 [.02,.90]	-
Fifteen minute sessions	.69 [.47,.83]	.50 [.33,.55]***
Full session, same set	.60 [.13,.92]	-
Exposure - Competence		
Full session, entire set	.62 [.54,.67]	-
Five minute sessions	.61 [.36,.79]	.50 [.19,.67]***
Full session, same set	.00 [.00-.87]	-
Fifteen minute sessions	.59 [.34,.77]	.43 [.11,.43]**
Full session, same set	.79 [.23,.96]	-
General Psychodynamic Focus		
Full session, entire set	.46 [.41,.50]	-

Five minute sessions	.00 [.00,.32]	.08 [.00,.17]
Full session, same set	*	-
Fifteen minute sessions	.00 [.00,.32]	.12 [.02,.21]*
Full session, same set	.38 [.07,.62]	-
General Family Focus		
Entire set, full sessions	.85 [.84,.87]	-
Five minute sessions	.42 [.12,.65]	.55 [.46,.62]***
Full session, same set	.83 [.71,.91]	-
Fifteen minute sessions	.67 [.45,.82]	.62 [.53,.69]***
Full session, same set	.76 [.59,.87]	
General Client-centered Focus		
Full session, entire set	.42 [.37-.46]	
Five minute sessions	.16 [.00,.45]	.12 [.02,.22]*
Full session, same set	.49 [.21-.69]	
Fifteen minute sessions	.00 [.00-.27]	.13 [.03,.23]**
Full session, same set	.51 [.24-.71]	
VTAS-R-SF		
Full session, entire set	.70 [.66,.73]	
Five minute sessions	.24 [.00,.52]	.59 [.48,.69]***
Full session, same set	.39 [.06,.64]	
Fifteen minute sessions	.49 [.19,.70]	.68 [.60,.80]***
Full session, same set	.62 [.34,.79]	

Note. ICC = Intraclass Correlation Coefficient, *r* = Pearson Correlation Coefficient, CI = Confidence Interval, * $p < .05$; ** $p < .01$; *** $p < .001$

Research question three – sensitivity to change. What is the shortest observation length that represents sensitivity to change based upon comparisons with the treatment integrity scores independently archived for the two trials?

Full session, entire set scores. When the full session, entire set of session scores were entered into the HLM models, all items except the *General Client-Centered Focus* item changed over time (i.e., the unconditional model was the best fitting model for *General Client-Centered Focus*). The *Skill-building-Adherence* subscale (Linear Slope y_{100} = -0.13; $p < .001$ and Quadratic Slope y_{200} = .0003; $p < .001$; Quadratic Deviance Difference χ^2 = 25.16, $p < .001$), *Exposure-Adherence* subscale (Linear Slope y_{100} = 0.26; $p < .001$ and Quadratic Slope y_{200} = -.0006; $p < .001$; Quadratic Deviance Difference χ^2= 108.82, $p < .001$), *Exposure-Competence* subscale (Linear Slope y_{100} = .05; $p = .002$ and Quadratic Slope y_{200} = -.002; $p < .001$; Quadratic Deviance Difference χ^2 = 19.82, $p < .001$), *General Psychodynamic Focus* item (Linear Slope y_{100} = .0004; $p = .12$ and Quadratic Slope y_{200} = .00 $p = .60$; Quadratic Deviance Difference χ^2 = 33.16, $p < .001$), *General Family Focus* item (Linear Slope y_{100} = -.04; $p < .001$ and Quadratic Slope y_{200} = .00; $p < .001$; Quadratic Deviance Difference X^2 = 12.30, $p < .05$), and *VTAS-R-SF* (Linear Slope y_{100} = -.04 $p = .39$ and Quadratic Slope y_{200} = .00 ; $p = .174$; Quadratic Deviance Difference χ^2 = 13.95 $p < .05$) subscale all best fit quadratic models, with therapists either delivering increasing doses/skill in those interventions toward the middle of treatment, and decreasing doses/skill towards the end of treatment or vice versa. The *Skill-Building-Competence* scale best fit a linear model, indicating that therapists across groups delivered less competent skill-building interventions over time (Linear Slope y_{100} = -.03; $p = .005$ and Quadratic Slope y_{200} = .00; $p = .74$; Linear Deviance Difference χ^2 = 58.42, $p < .001$ and Quadratic Deviance Difference χ^2 = 4.85, $p > .05$). In sum, when the full session, entire set of sessions were analyzed, all items demonstrated sensitivity to change (either in dosage or skill level) except *General Client-Centered Focus*.

Full session, four set scores. In comparison, when the full session, four set scores were entered into the HLM model, only *Skill-building-Adherence*, *Exposure-Adherence*, and *VTAS-R-SF* exhibited change over time. *Skill-building-Adherence* fit a linear model (Linear Slope y_{100} = -.17; $p < .001$ and Quadratic Slope y_{200} = .001; $p = .024$; Quadratic Deviance Difference χ^2 = 8.52, $p > .05$), indicating that skill-building scores slightly decreased over time. The *Exposure-Adherence* (Linear Slope y_{100} = 0.24; $p < .001$ and Quadratic Slope y_{200} = -0.01; $p < .001$; Quadratic Deviance Difference χ^2 = 20.99, $p < .001$) and *VTAS-R-SF* (Linear Slope y_{100} = 0.03; $p = .198$ and Quadratic Slope y_{200} = 0.00; $p = 0.098$; Quadratic Deviance Difference χ^2 = 11.83, $p < .001$) subscales best fit quadratic models, suggesting that exposure dosage and alliance increased slightly in the middle of treatment and dipped towards the end of treatment. Taken together, the full session, four set scores only paralleled the full session, entire set score trajectories on two items: *Exposure –Adherence* and *VTAS-R-SF.* The significance levels and direction of change for linear and quadratic slopes were consistent across full session scores on five out of eight items: *Skill-building Adherence, Exposure-Adherence*, *General Psychodynamic Focus*, *General Client-Centered Focus*, and *VTAS-R-SF*. Thus, compared to the entire set scores, selecting four sessions for this study reduced the number of items that demonstrated sensitivity to change.

Five-minute segment. For the five-minute segments, three TIES-CBT-YA items exhibited change over time (two of which were different from the items that changed over time in the full session, four set scores): *Exposure-Adherence*, *Exposure-Competence*, and *General Family Focus*. *Exposure-Adherence* was relatively consistent with the full session scores; this item fit a quadratic model (same as the full session, four set score), had the same pattern of significant time effects, and a similar direction of change (Linear Slope $y_{100} = 0.15$, $p < .001$ and Quadratic Slope $y_{200} = 0.00$, $p = .002$; Quadratic Deviance Difference $\chi^2 = 12.86$, $p < .05$). *Exposure-Competence* and *General Family Focus* were slightly different across full and five-minute scores. In the five-minute scores, both of these items fit a linear model whereas the full session, four set scores fit unconditional models; however, the linear and quadratic slopes were both not significant for base quadratic models (*Exposure-Competence* Linear Slope $y_{100} = -.20$, $p = .413$ and Quadratic Slope $y_{200} = 0.01$, $p = .408$; Quadratic Deviance Difference $\chi^2 = 0.63$, $p > .05$. *General Family Focus* Linear Slope $y_{100} = 0.01$; $p = 0.631$ and Quadratic Slope $y_{200} = 0.00$; $p = .317$; Quadratic Deviance Difference $\chi^2 = 27.94$, $p < .001$). The *Skill-building-Adherence* item was also relatively discrepant from the full session, four set scores; it fit an unconditional model (unlike the full session, four set score, which fit a linear model), did not have significant time effects, and did not have a significant -2LL when time was entered into the model (Linear Slope $y_{100} = -0.06$; $p = .333$ and Quadratic Slope $y_{200} = 0.00$; $p = .890$; Quadratic Deviance Difference $\chi^2 = -869.25$, $p > .05$). Finally, *Skill-building Competence* (Linear Slope $y_{100} = 0.01$; $p = .824$ and Quadratic Slope $y_{200} = 0.00$; $p = .939$; Quadratic Deviance Difference $\chi^2 = 1.01$, $p > .05$), *General Psychodynamic Focus* (Linear Slope $y_{100} = 0.01$; $p = 0.217$ and Quadratic Slope $y_{200} = 0.00$; $p = .452$; Quadratic Deviance Difference $\chi^2 = 5.27$ $p > .05$), *General Client-Centered Focus* (Linear Slope $y_{100} = 0.03$; $p = .514$ and Quadratic Slope $y_{200} = -0.002$; $p = .312$; Quadratic Deviance Difference $\chi^2 = 18.41$, $p < .001$) and *VTAS-R-SF* (Linear Slope $y_{100} = 0.01$; $p = .676$ and Quadratic Slope $y_{200} = 0.00$; $p = .624$; Quadratic Deviance Difference $\chi^2 = -267.40$, $p > .05$) all fit unconditional means models and mirrored the full session, four set scores in growth model fit, statistical significance of slopes, and slope direction. Taken together, three items exhibited sensitivity to change and five items fit the same trajectories of change (i.e., same growth model, slopes, and slope direction) in the five-minute segments as the full session, four set scores.

Fifteen-minute segments. For the 15-minute segments, three TIES-CBT-YA items also exhibited sensitivity to change (two of which were the same as the full session, full set scores): *Exposure-Adherence*, *Skill-building-Competence*, and *VTAS-R-SF*. *Exposure-Adherence* was very similar to the full session scores; this item fit a quadratic model (same as the full session, four set score), had similarly significant time effects, and a exhibited a parallel direction of change (*Exposure-Adherence* Linear Slope $y_{100} = 0.12$, $p < .001$ and Quadratic Slope $y_{200} = 0.00$, $p = .003$; Quadratic Deviance Difference $X^2 = 10.43$, $p < .05$). *Skill-building Competence* was slightly different from the full session, four set scores; this item had similar linear and quadratic slopes (i.e., no change over time), however it better fit a linear model (rather than an unconditional model) due to the reduced -2LL across models (Linear Slope $y_{100} = -0.05$, $p = .123$ and Quadratic Slope $y_{200} = 0.00$, $p = .843$; Linear Deviance Difference $\chi^2 = 12.25$, $p < .01$). Finally, *VTAS-R-SF* was similar to the full session, four set scores in both linear and quadratic slope significance and direction, except that it better fit a linear rather than quadratic model (*VTAS-R-SF* Linear Slope $y_{100} = 0.12$, $p < .001$ and Quadratic Slope $y_{200} = 0.00$, $p = .003$; Quadratic Deviance Difference $\chi^{2\,2} = 4.47$, $p > .05$). All other items paralleled the five-minute versus full session comparisons. Thus, three items exhibited sensitivity to change and six items fit the same trajectories of change in the 15-minute segments as the full session, four set scores.

Summary-observation segment differences. In general, three out of eight TIES-CBT-YA items in both observation segments were sensitive to change and the majority of items closely mirrored the full session scores in trajectories, linear and quadratic slope, and slope direction. The 15-minute scores had one more item than the five-minute scores that better resembled the full session, four set scores in growth model, slope, and slope direction. Thus, the 15-minute segment appeared to *slightly* better approximate the full session scores on sensitivity to change than the five-minute segment.

Summary-item differences. Across TIES-CBT-YA observation segments, *Adherence-Exposure* exhibited the most sensitivity to change, fitting a quadratic model across both brief and full session segments. In contrast, *General Psychodynamic Focus* and *General Client-Centered Focus* exhibited the least sensitivity to change across both brief and full session segments, both fitting unconditional means models. The rest of the items fit different trajectories depending on the observation length. Overall, the 15-minute segment scores somewhat better approximated the pattern of item differences of the full session, four set scores than the five-minute segment scores.

Table 5

Multilevel Models Comparing Sensitivity to Change Scores across Observation Segments

Base Quadratic Models	Coefficient	S.E.	Deviance	*n* Parameters in Model
Skill-building - Adherence				
Full session, entire set			4609.43	11
Intercept (first session value), γ_{000}	4.78***	0.19		
Slope (change over time in weeks), γ_{100}	-0.13**	0.02		
Slope (change over time in weeks squared), γ_{200}	0.00***	0.00		
Full session, four set			1008.57	11
Intercept (first session value), γ_{000}	4.91***	0.31		
Slope (change over time in weeks), γ_{100}	-0.17***	0.05		
Slope (change over time in weeks squared), γ_{200}	0.00*	0.00		
Five minute sessions				
Intercept (first session value), γ_{000}	4.05***	0.33	1528.95	11
Slope (change over time in weeks), γ_{100}	-0.06	0.06		
Slope (change over time in weeks squared), γ_{200}	0.00	0.00		
Fifteen minute sessions				
Intercept (first session value), γ_{000}	4.29***	0.34	1489.95	11
Slope (change over time in weeks), γ_{100}	-0.05	0.05		
Slope (change over time in weeks squared), γ_{200}	0.00	0.00		
Exposure - Adherence				

Full session, entire set			4447.03	11
Intercept (first session value), γ_{000}	0.64***	0.14		
Slope (change over time in weeks), γ_{100}	0.26***	0.02		
Slope (change over time in weeks squared), γ_{200}	-0.01***	0.00		
Full session, four set			867.58	11
Intercept (first session value), γ_{000}	0.38	0.19		
Slope (change over time in weeks), γ_{100}	0.24***	0.04		
Slope (change over time in weeks squared), γ_{200}	-0.01***	0.00		
Five minute sessions			391.12	11
Intercept (first session value), γ_{000}	0.57*	0.28		
Slope (change over time in weeks), γ_{100}	0.15***	0.03		
Slope (change over time in weeks squared),γ_{200}	0.00**	0.00		
Fifteen minute sessions			807.53	11
Intercept (first session value), γ_{000}	0.71***	0.17		
Slope (change over time in weeks), γ_{100}	0.12***	0.03		
Slope (change over time in weeks squared), γ_{200}	0.00**	0.00		
Skill-building - Competence				
Full session, entire set			1977.31	11
Intercept (first session value), γ_{000}	4.58***	0.14		
Slope (change over time in weeks), γ_{100}	-0.03**	0.01		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
Full session, four set			704.21	11
Intercept (first session value), γ_{000}	4.61***	0.19		
Slope (change over time in weeks), γ_{100}	-0.02	0.03		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
Five minute sessions			801.76	11
Intercept (first session value), γ_{000}	4.51***	0.22		
Slope (change over time in weeks), γ_{100}	0.01	0.03		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
Fifteen minute sessions			943.12	11
Intercept (first session value), γ_{000}	5.04***	0.21		
Slope (change over time in weeks), γ_{100}	-0.05	0.03		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
Exposure - Competence				
Full session, entire set			1322.36	11
Intercept (first session value), γ_{000}	4.09***	0.21		
Slope (change over time in weeks), γ_{100}	0.05**	0.02		
Slope (change over time in weeks squared),γ_{200}	0.00***	0.00		
Full session, four set			224.99	11
Intercept (first session value), γ_{000}	1.98	0.89		
Slope (change over time in weeks), γ_{100}	0.28	0.15		
Slope (change over time in weeks squared),γ_{200}	-0.01	0.01		
Five minute sessions			148.88	11
Intercept (first session value), γ_{000}	5.78***	1.69		

Slope (change over time in weeks), γ_{100}	-0.20	0.24		
Slope (change over time in weeks squared),γ_{200}	0.01	0.01		
Fifteen minute sessions			225.99	8
Intercept (first session value), γ_{000}	1.93	1.01		
Slope (change over time in weeks), γ_{100}	0.3	0.17		
Slope (change over time in weeks squared),γ_{200}	-0.01	0.01		
General Psychodynamic Focus				
Full session, entire set			290.19	11
Intercept (first session value), γ_{000}	1.04***	0.02		
Slope (change over time in weeks), γ_{100}	0.00	0.00		
Slope (change over time in weeks squared),γ_{200}				
Full session, four set			3.90	11
Intercept (first session value), γ_{000}	1.04***	0.04		
Slope (change over time in weeks), γ_{100}	0.01	0.01		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
Five minute sessions			179.96	8
Intercept (first session value), γ_{000}	0.98***	0.03		
Slope (change over time in weeks), γ_{100}	0.01	0.01		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
Fifteen minute sessions			65.71	11
Intercept (first session value), γ_{000}	1.06***	0.03		
Slope (change over time in weeks), γ_{100}	0.00	0.01		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
General Family Focus				
Full session, entire set			4436.54	11
Intercept (first session value), γ_{000}	2.23***	0.13		
Slope (change over time in weeks), γ_{100}	-0.04***	0.01		
Slope (change over time in weeks squared),γ_{200}	0.00***	0.00		
Full session, four set			1079.76	11
Intercept (first session value), γ_{000}	1.66***	0.17		
Slope (change over time in weeks), γ_{100}	0.02	0.02		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
Five minute sessions			927.54	11
Intercept (first session value), γ_{000}	1.24***	0.14		
Slope (change over time in weeks), γ_{100}	0.01	0.02		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
Fifteen minute sessions			1054.66	11
Intercept (first session value), γ_{000}	1.39***	0.18		
Slope (change over time in weeks), γ_{100}	0.01	0.02		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
General Client-Centered Focus				
Full session, entire set			1502.22	11
Intercept (first session value), γ_{000}	2.93***	0.10		
Slope (change over time in weeks), γ_{100}	0.00	0.01		
Slope (change over time in weeks squared),γ_{200}	0.00***	0.00		
Full session, four set			1437.98	11

Intercept (first session value), γ_{000}	2.81***	0.15		
Slope (change over time in weeks), γ_{100}	0.04	0.20		
Slope (change over time in weeks squared),γ_{200}	0.00*	0.00		
Five minute sessions			1502.22	11
Intercept (first session value), γ_{000}	3.02***	0.29		
Slope (change over time in weeks), γ_{100}	003	0.05		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
Fifteen minute sessions			1437.98	11
Intercept (first session value), γ_{000}	2.96***	0.25		
Slope (change over time in weeks), γ_{100}	0.08	0.04		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
VTAS-R-SF				
Full session, entire set			1034.65	11
Intercept (first session value), γ_{000}	2.69***	0.08		
Slope (change over time in weeks), γ_{100}	0.00	0.01		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
Full session, four set			353.95	11
Intercept (first session value), γ_{000}	2.80***	0.10		
Slope (change over time in weeks), γ_{100}	0.03	0.03		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
Five minute sessions			828.95	11
Intercept (first session value), γ_{000}	3.12***	0.12		
Slope (change over time in weeks), γ_{100}	0.01	0.02		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		
Fifteen minute sessions			607.62	11
Intercept (first session value), γ_{000}	3.43***	0.13		
Slope (change over time in weeks), γ_{100}	0.00	0.02		
Slope (change over time in weeks squared),γ_{200}	0.00	0.00		

* $p < .05$; ** $p < .01$; *** $p < .001$

Research question four – convergent validity. What is the shortest observation segment that represents convergent validity based upon comparisons with the treatment integrity scores independently archived for the two trials?

Five-minute segments. When five-minute scores were compared to full session, four set items scores, all items had large associations with full session scores (*Skill-building-Adherence r* = .65, *p* < .001; *Exposure-Adherence r* = .76, *p* < .00; *Skill-building-Competence r* = .40, *p* = .001; *Exposure-Competence r* = .50, *p* = .001; *General Family Focus r* = .55, *p* < .001; *VTAS-R-SF r* = .59, *p* < .001) except *General Client-Centered Focus* (*r* = .13, *p* = .021) and *General*

Psychodynamic Focus (*r* = .08, *p* = .140). Specifically, *General Client-Centered Focus* fell within the "small" association range and *General Psychodynamic Focus* did not have any association with the full set scores. Thus, for five-minute segments, five TIES-CBT-YA items appeared to be related to the items intended to be theoretically similar on the full-scale instruments; two TIES-CBT-YA items had no to little association with the full scale items.

Fifteen-minute segments. When 15-minute item scores were compared to full session item scores, the same five items had "large" associations with full set scores (*Skill-building-Adherence r* = .72, *p* < .001; *Exposure-Adherence r* = .61, *p* < .001, *Skill-building-Competence r* = .46, *p* < .001; *Exposure-Competence r* = .43, *p* = .001; *General Family Focus r* = .62, *p* < .001; *VTAS-R-SF r* = .68, *p* < .001). In addition, *General Client-Centered Focus* and *General Psychodynamic Focus* both had minimal relationships with the full session scores, falling in the "small" association range (*r* = .12, *p* = .02; *r* = .13, *p* = .009, respectively). Similar to the five-minute segments, the majority of items exhibited large associations to full set scores.

Summary-observation segment differences. Overall, the five- and 15-minute segments produced a consistent pattern of associations to full session, four set scores. Across segments, the same items had either large (five items) or no/small association (two items) with full session item scores. Both observations therefore were able to capture the majority of intended theoretical content between TIES-CBT-YA and full scale instruments.

Summary-item differences. When considering both brief segment correlation findings together, all items had strong correlations with the full session, four set scores except *General Client-Centered Focus* and *General Psychodynamic Focus*. These items had no and small associations with full session scores, respectively. Thus, the TIES-CBT-YA items generally had a strong association with the full session scales that were intended to be theoretically similar (i.e., ICBT interventions).

Research question five – discriminant validity. I explored my fifth research question: what is the shortest observation segment that shows sensitivity to differences in integrity between ICBT and UC in a way that approximates the full session, four set scores (i.e., discriminant validity)?

Full session, entire set scores. For *Skill-building-Adherence* analyses, findings with the full session, entire set data indicated that therapists in the standard manualized treatments (K-SMT and CS-SMT) delivered significantly higher doses of skill-building interventions than those in the modular manualized treatment (K-SMT vs. CS-MMT γ = 1.22, $p < .001$, $d = 0.67$, 95% CI [0.55, 0.79]; CS-SMT vs. CS-MMT γ = 1.13, $p < .001$, $d = 0.62$, 95% CI [0.49, 0.75]) and CS-UC (K-SMT vs. CS-UC γ = 2.57, $p < .001$, $d = 1.41$, 95% CI [1.29, 1.53]; CS-SMT vs CS-UC γ = 2.49, $p < .001$, $d = 1.37$, 95% CI [1.25, 1.49]), and therapists in CS-MMT delivered significantly higher doses of skill-building interventions than those in CS-UC (γ = 1.36, $p < .001$, $d = 0.75$, 95% CI [0.62, 0.88]); the dosages delivered in K-SMT and CS-SMT did not significantly differ ($p = .660$). In comparison, for *Exposure-Adherence* analyses, results indicated that scores in K-SMT were significantly higher for exposure than those in all other groups (K-SMT vs. CS-SMT γ = 1.73, $p < .001$, $d = 0.95$, 95% CI [0.85, 1.05]; K-SMT vs. CS-MMT γ = 1.42, $p < .001$, $d = 0.78$, 95% CI [0.67, 0.89]; K-SMT vs. CS-UC γ = 2.51, $p < .001$, $d = 1.38$, 95% CI [1.26, 1.49]). Therapists in CS-MMT and CS-SMT did not significantly differ in their exposure dosages ($p = .160$), and therapists in CS-UC delivered significantly less exposure interventions than all groups (CS-UC vs. CS-SMT γ = -0.78, $p < .001$, $d = 0.43$, 95% CI [0.31, 0.55]; CS-UC vs. CS-MMT γ = -1.08, *p* < .001, *d* = 0.59, 95% CI [0.46, 0.73]). For *Skill-Building-Competence*, therapists in K-SMT delivered more competent skill-building interventions than those in all other groups (K-SMT vs. CS-SMT γ = 1.09 *p* < .001, *d* = 0.92, 95% CI [0.77, 1.08]; K-SMT vs. CS-MMT γ = 1.45, *p* < .001, *d* = 1.23, 95% CI [1.06, 1.39]; K-SMT vs. CS-UC γ = 2.20, *p* < .001, *d* = 1.86, 95% CI [1.64, 2.09]); therapists in CS-MMT and CS-SMT did not significantly differ in their skill-building competence (*p* = 0.090), and therapists in CS-UC delivered significantly less competent skill-building interventions than all groups (CS-UC vs. CS-SMT γ = -1.10, *p* < .001, *d* = 0.93, 95% CI [0.70, 1.16]; CS-UC vs. CS-MMT γ = -0.75, *p* = .013, *d* = 0.64, 95% CI [0.39, 0.88]). For *General Psychodynamic Focus* analyses, results indicated that therapists in K-SMT, CS-SMT, and CS-MMT did not differ in the extent that they delivered psychodynamic interventions (K-SMT vs. CS-SMT *p* =.582; K-SMT vs. CS-MMT *p* = .340; CS-SMT vs. CS-MMT *p* = .670). All therapists in manualized groups delivered significantly lower doses of psychodynamic interventions than those in CS-UC (CS-UC vs. K-SMT γ = 0.19, *p* < .001, *d* = 0.65, 95% CI [0.45, 0.86]; CS-UC vs. CS-SMT γ = 0.17, *p* < .001, *d* = 0.58, 95% CI [0.45, 0.72]; CS-UC vs. CS-MMT γ = 0.15, *p* = .001, *d* = 0.52, 95% CI [0.38, 0.66]). For *General Family Focus*, analyses indicated that therapists in CS-MMT delivered significantly higher doses of family interventions than therapists in K-SMT (γ = 0.63, *p* = .009, *d* = 0.46, 95% CI [0.29, 0.63]) and CS-SMT (γ = 0.80, *p* = .001, *d* = 0.58, 95% CI [0.42, 0.75]). Therapists in CS-UC delivered significantly higher doses of family interventions than therapists in CS-SMT (γ = 0.56, *p* = .018, *d* = 0.41, 95% CI [0.25, 0.57]). Therapists did not differ in family intervention delivery in CS-UC and CS-MMT (*p* = .332), CS-UC and K-SMT (*p* = .093), or K-SMT and CS-SMT (*p* = .408). For *General Client-Centered Focus,* therapists in CS-UC delivered significantly more client-centered interventions than all other groups (CS-UC vs. K-SMT γ = 0.97, *p* < .001, *d* = 1.00, 95% CI [0.81, 1.19]; CS-UC vs. CS-SMT γ = 0.70, *p* < .001, *d* = 0.72, 95% CI [0.54, 0.91]; CS-UC vs. CS-MMT γ = 0.62, *p* = .004, *d* = 0.64, 95% CI [0.43, 0.85]); no other groups significantly differed in their dosage of client-centered interventions. Finally, for *VTAS-R-SF* analyses, therapists in K-SMT had a significantly stronger alliance with the child than therapists in CS-UC (γ = 0.69, *p* < .001, *d* = 0.92, 95% CI [0.68, 1.16]) and CS-MMT (γ = 0.42, *p* = .029, *d* = 0.56, 95% CI [0.32, 0.80]). Therapists in CS-SMT had a significantly stronger alliance than therapists in CS-UC (γ = 0.46, *p* = .019, *d* = 0.61, 95% CI [0.36, 0.87]); therapists in all other groups did not differ in alliance. As was expected, full session, entire set findings indicated that therapists in the ICBT groups adhered more to the ICBT protocol and delivered ICBT interventions more skillfully than therapists in CS-UC.

Full session, four set scores. The majority of full session, four set item scores paralleled the full session, entire set scores in group comparison significance patterns and effect sizes. In particular, *Adherence-Skill-building*, *Skill-building-Competence, Exposure-Competence, General Psychodynamic Focus,* and *General Client-Centered Focus* items all had group comparison findings that were similar to the full session, entire set scores. The group comparison findings were discrepant across the sets for *Adherence-Exposure*, *General Family Focus*, and *VTAS-R-SF* items. For *Adherence-Exposure*, unlike in the entire set scores, therapists in CS-MMT delivered significantly higher exposure dosages than those in CS-SMT (γ = 0.98, *p* < .001, *d* = 0.54, 95% CI [0.39, 0.70]), and therapists in CS-SMT did not deliver significantly different exposure dosages than those in CS-UC (*p* = .434). For *General Family Focus*, in contrast to the entire set scores, therapists in CS-UC delivered significantly higher doses of family interventions than therapists in both CS-SMT (γ = 0.94, $p < .001$, d = 0.90, 95% CI [0.66, 1.14]) and K-SMT (γ = 1.06, $p < .001$, d = 1.02, 95% CI [0.78, 1.26]). Finally, for *VTAS-R-SF* analyses, findings indicated that therapists in K-SMT had stronger alliance than therapists in CS-SMT (rather than therapists in CS-MMT; γ = 0.44, p = .047, d = 0.68, 95% CI [0.34, 1.02]). When these comparisons were considered for the purpose of this study, all adherence and competence items met group comparison expectations except *Adherence-Exposure*; results for this item indicated that therapists in the standard manualized treatment (CS-SMT) did not deliver significantly higher dosages of exposure interventions than therapists who received no training in ICBT (CS-UC). Thus, the four sessions selected for this study may have limited conclusions that can be drawn about the extent to which exposure interventions are delivered.

Five-minute segments. For the *Skill-building-Adherence* item in the five-minute segment analyses, all six group comparisons had the same significance patterns as the full session, four set scores (i.e., all groups were significantly different from one another except K-SMT and CS-SMT); even though the effect sizes were somewhat lower in the five-minute scores than the full session scores, all significant items fell in the “small” to “medium” association range (d = 0.21-.0.60). For *Exposure-Adherence*, two group comparisons had significance levels that differed from the full session scores; specifically, therapists in K-SMT did not deliver significantly more interventions than those in CS-SMT (p = .056) and CS-MMT (p = .383) and those findings had lower effect sizes than the full session scores (d = 0.15 and d = 0.07, respectively). *Skill-building-Competence* was discrepant from the full session, four set scores, with four group comparison findings that differed from the full session, four set scores. The discrepancies were that therapists in K-SMT did not deliver skill-building interventions more competently than therapists in CS-SMT (p = .138, d = 0.33 95% CI [0.12, 0.55]), therapists in

CS-UC did not deliver interventions significantly more competently than therapists in CS-MMT (p = .498, d = 0.26 95% CI [0.00, 0.67]) or CS-SMT (p = .141, d = 0.53 95% CI [0.18, 0.89]), and therapists in CS-SMT delivered skill-building interventions more competently than therapists in CS-MMT (γ = 0.97, p = .004, d = 0.81, 95% CI [0.54, 1.02]). It is important to note that *Skill-Building-Competence* for K-SMT vs. CS-SMT had a small effect size (and the confidence interval did not include zero) and CS-UC vs. CS-SMT had a medium effect size, indicating that these significance levels may have been influenced by sample size. For *General Psychodynamic Focus*, all group comparisons had the same significance levels as the full session, four set scores; though the effect sizes were smaller for the five-minute scores, all significant group comparisons had effect sizes that fell within the small-medium range. *General Family Focus* and *General Client-Centered Focus* both had somewhat discrepant findings when compared to the full session scores. Specifically*,* there were no significant differences between family intervention (*General Family Focus*) delivery dosage in K-SMT versus CS-MMT (p = .275, d = 0.11, 95% CI [0.00, 0.25]) or CS-SMT and CS-MMT (p = .231, d = 0.09 95% CI [0.00, 0.21]); the lower limit of the confidence intervals included zero, reflecting the lack of difference between these groups. There were also no significant differences in client-centered intervention (*General Client-Centered Focus*) dosage for therapists in K-SMT versus CS-UC (p = .362, d = 0.08 95% CI [0.00, 0.17]), CS-SMT and CS-UC (p = .883, d = 0.01, 95% CI [0.00, 0.09]), and CS-MMT and CS-UC (p = .202, d = 0.13 95% CI [0.03, 0.22]); again, the effect sizes were minimal. Lastly, *VTAS-R-SF* had one significance level difference when compared to the full session scores; therapists in K-SMT did not have significantly stronger alliance scores than therapists in CS-SMT (p = .137, d = 0.36 95% CI [0.13, 0.60]). However, this comparison had a “small” effect size, indicating that sample size may have influenced statistical significance for this item. Overall, two items had the same pattern of group comparison findings as the full session, four set scores; the rest had at least one difference. However, the effect sizes indicate that three comparisons may have had small differences despite non-significant *p* values. Consistent with the full session, four set scores and counter to expectations, therapists in CS-SMT did not deliver higher dosages of exposure interventions than therapists in CS-UC. In addition, therapists in CS-MMT or CS-SMT did not deliver significantly more competent skill-building interventions than therapists in CS-UC. Thus, two five-minute segment items were inconsistent with expectations, and one of these inconsistent items was also different from the full session, four set scores.

Fifteen-minute segments. Fifteen-minute segments group comparison findings somewhat better resembled the full session scores. For both *Skill-building-Adherence* and *General Psychodynamic Focus*, the 15-minute scores had the same pattern of significance level findings as the full session scores and all significant findings had effect sizes in the “small” to “large” association ranges. In addition, for *Exposure-Adherence*, only one of the group comparison significance findings differed from full session scores; therapists in K-SMT did not deliver significantly more interventions than those in CS-MMT and the effect size was trivial (p = .820, d = 0.04 95% CI [0.00, 0.22]). *Skill-building-Competence* similarly only had one group comparison difference; unlike in the full session scores, therapists in CS-SMT delivered significantly more competent skill-building interventions than those in CS-MMT (γ = 0.65, p = .026, d = 0.50, 95% CI [0.28, 0.72]). *General Family Focus* and *General Client-Centered Focus* least resembled the full session scores; none of the group comparisons were significant across these items (with four group comparison differences in family interventions and three in client-centered interventions); however, it is important to note that, although these comparisons were not significant, a number of these effect sizes fell in the “small” to “medium” range ($d = 0.22$ - 0.47), indicating that sample size may have influenced statistical significance. Lastly, there were three group comparison findings that differed for the *VTAS-R-SF* item; therapists in CS-SMT and CS-MMT had stronger alliances with the child than those in CS-UC (CS-SMT vs. CS-UC $\gamma = 0.45$, $p = .044$, $d = 0.52$, 95% CI [0.26, 0.77]; CS-MMT vs CS-UC $\gamma = 0.49$, $p = .049$, $d = 0.56$, 95% CI [0.29, 0.84]). Only two items had the same pattern of group comparisons as full session scores and the rest had at least one difference. In accordance with the full session, four set scores and opposite to study expectations, therapists in CS-SMT did not deliver higher dosages of exposure interventions than therapists in CS-UC. However, unlike in the five-minute scores and consistent with the full session scores, therapists in CS-MMT delivered significantly more competent skill-building interventions than therapists in CS-UC.

Summary-observation segment differences. Overall, the five- and 15-minute scores both had two items that had the same pattern of group comparison scores of the full session, four set scores. All other items had at least one group comparison difference; however, a number of non-significant findings had “small” to “medium” effect sizes when the full session group comparisons were significant. For this study, the most important consideration in the discriminant analyses was whether the TIES-CBT-YA instrument could differentiate between treatments that were intended to be different (i.e., ICBT vs UC). For full session, four set scores (the gold standard comparator), this expectation was not met for all items; therapists in CS-SMT did not significantly differ from therapists in CS-UC in the dosage of exposure interventions delivered (i.e., *Exposure-Adherence*). Both five- and 15-minute scores exhibited the same significance pattern as the full session, four set score for *Exposure-Adherence*. Thus, it may be that the four sessions selected for the study did not optimally represent the range of exposure interventions across treatment groups. However, the five-minute scores had one additional finding that did not meet the group comparison significance and effect size expectations; therapists in CS-MMT did not significantly differ from therapists in CS-UC in the skill with which they delivered skill-building interventions (i.e., *Skill-building Competence*). Thus, the 15-minute segment may have *slightly* better met discriminant validity expectations and approximated the full session, four set scores.

Summary-item differences. Across brief segments, *Skill-building-Adherence* and *General Psychodynamic Focus* best approximated treatment group differences (i.e., they have all the same group comparison findings); all other items had at least one group comparison difference across observations. When assessing *Exposure-Competence*, there were not enough exposure sessions coded in CS-UC, so HLM analyses could not be conducted for full session or brief segment scores. Thus, alternative means were used to compare the groups. In full session, four set segments, one-way ANOVA analyses of treatment group comparisons identified that there were differences in *Exposure-Competence* across groups ($F = 42.12$, $p < .001$). Bonferonni post-hoc analyses indicated that therapists in K-SMT delivered exposure interventions more skillfully than those in CS-SMT ($t = 2.13$, $p < .001$) and CS-MMT ($t = 1.97$, $p < .001$); CS-UC was not included due to limited data. In contrast, there were no differences in *Exposure-Competence* across treatment groups in both five- ($F = 1.23$, $p = .301$) and 15- ($F = 1.68$, $p = .18$) minute segments. Overall, the group comparison findings indicated that *Skill-building-Adherence* and *General Psychodynamic Focus* had the highest discriminant validity and *Exposure-Adherence* had the lowest discriminant validity across brief segments.

Table 6

Multilevel Models Comparing Treatment Group Difference Scores across Observation Segments

Unconditional Means Models	Coefficient	S.E.	ES 95% [CI]
Skill-building - Adherence			
Full session, entire set			
K-SMT (1) vs. CS-UC (0)	2.57***	0.22	1.41 [1.29,1.53]
K-SMT (1) vs. CS-SMT (0)	0.90	0.19	0.49 [0.39,0.60]
K-SMT (1) vs. CS-MMT (0)	1.22***	0.22	0.67 [0.55,0.79]
CS-SMT (1) vs. CS-UC (0)	2.49***	0.22	1.37 [1.25,1.49]
CS-SMT (1) vs. CS-MMT (0)	1.13***	0.23	0.62 [0.49,0.75]
CS-MMT (1) vs. CS-UC (0)	1.36***	0.24	0.75 [0.62,0.88]
Full session, four set			
K-SMT (1) vs. CS-UC (0)	2.59***	0.32	1.42 [1.24,1.59]
K-SMT (1) vs. CS-SMT (0)	-0.39	0.29	0.21 [0.05,0.37]
K-SMT (1) vs. CS-MMT (0)	0.70*	0.32	0.38 [0.21,0.56]
CS-SMT (1) vs. CS-UC (0)	2.98***	0.29	1.63 [1.47,1.79]
CS-SMT (1) vs. CS-MMT (0)	1.09**	0.33	0.60 [0.42,0.78]
CS-MMT (1) vs. CS-UC (0)	1.89***	0.37	1.03 [0.83,1.23]
Five minute sessions			
K-SMT (1) vs. CS-UC (0)	1.19***	0.18	0.57 [0.48,0.65]
K-SMT (1) vs. CS-SMT (0)	0.06	0.18	0.03 [0.00,0.11]
K-SMT (1) vs. CS-MMT (0)	0.75***	0.19	0.36 [0.27,0.45]
CS-SMT (1) vs. CS-UC (0)	1.25***	0.18	0.60 [0.51,0.68]
CS-SMT (1) vs. CS-MMT (0)	0.81***	0.18	0.39 [0.30,0.47]
CS-MMT (1) vs. CS-UC (0)	0.44*	0.19	0.21 [0.12,0.30]
Fifteen minute sessions			
K-SMT (1) vs. CS-UC (0)	3.71***	0.31	1.79 [1.64,1.94]
K-SMT (1) vs. CS-SMT (0)	0.43	0.30	0.21 [0.06,0.35]
K-SMT (1) vs. CS-MMT (0)	1.91***	0.32	0.92 [0.77,1.08]
CS-SMT (1) vs. CS-UC (0)	3.27***	0.28	1.58 [1.44,1.71]
CS-SMT (1) vs. CS-MMT (0)	1.47***	0.39	0.71 [0.52,0.90]
CS-MMT (1) vs. CS-UC (0)	1.80***	0.29	0.87 [0.73,1.01]
Exposure - Adherence			
Full session, entire set			
K-SMT (1) vs. CS-UC (0)	2.51***	0.21	1.38 [1.26,1.49]
K-SMT (1) vs. CS-SMT (0)	1.73***	0.18	0.95 [0.85,1.05]
K-SMT (1) vs. CS-MMT (0)	1.42***	0.20	0.78 [0.67,0.89]
CS-SMT (1) vs. CS-UC (0)	0.78***	0.22	0.43 [0.31,0.55]
CS-SMT (1) vs. CS-MMT (0)	-0.30	0.21	0.16 [0.05,0.28]
CS-MMT (1) vs. CS-UC (0)	1.08***	0.24	0.59 [0.46,0.73]
Full session, four set			
K-SMT (1) vs. CS-UC (0)	2.04***	0.16	1.13 [1.04,1.22]
K-SMT (1) vs. CS-SMT (0)	1.98***	0.17	1.09 [1.00,1.19]

K-SMT (1) vs. CS-MMT (0)	1.00**	0.33	0.55 [0.37,0.73]
CS-SMT (1) vs. CS-UC (0)	0.06	0.07	0.04 [0.00,0.07]
CS-SMT (1) vs. CS-MMT (0)	-0.98***	0.28	0.54 [0.39,0.70]
CS-MMT (1) vs. CS-UC (0)	1.04***	0.30	0.57 [0.41,0.74]
Five minute sessions			
K-SMT (1) vs. CS-UC (0)	0.26***	0.22	0.19 [0.03,0.35]
K-SMT (1) vs. CS-SMT (0)	0.21	0.11	0.15 [0.07,0.23]
K-SMT (1) vs. CS-MMT (0)	0.10	0.12	0.07 [0.00,0.16]
CS-SMT (1) vs. CS-UC (0)	0.05	0.11	0.04 [0.00,0.12]
CS-SMT (1) vs. CS-MMT (0)	-0.31***	0.11	0.22 [0.14,0.30]
CS-MMT (1) vs. CS-UC (0)	0.36**	0.12	0.26 [0.17,0.35]
Fifteen minute sessions			
K-SMT (1) vs. CS-UC (0)	0.77**	0.24	0.54 [0.37,0.71]
K-SMT (1) vs. CS-SMT (0)	0.70**	0.23	0.49 [0.33,0.65]
K-SMT (1) vs. CS-MMT (0)	0.06	0.25	0.04 [0.13,0.22]
CS-SMT (1) vs. CS-UC (0)	0.07	0.23	0.05 [0.00,0.21]
CS-SMT (1) vs. CS-MMT (0)	-0.76**	0.24	0.54 [0.37,0.70]
CS-MMT (1) vs. CS-UC (0)	0.83**	0.25	0.58 [0.41,0.76]
Skill-building - Competence			
Full session, entire set			
K-SMT (1) vs. CS-UC (0)	2.20***	0.27	1.86 [1.64,2.09]
K-SMT (1) vs. CS-SMT (0)	1.09***	0.18	0.92 [0.77,1.07]
K-SMT (1) vs. CS-MMT (0)	1.45***	0.20	1.23 [1.06,1.40]
CS-SMT (1) vs. CS-UC (0)	1.10***	0.27	0.93 [0.70,1.16]
CS-SMT (1) vs. CS-MMT (0)	0.35	0.21	0.29 [0.12,0.47]
CS-MMT (1) vs. CS-UC (0)	0.75*	0.29	0.63 [0.39,0.88]
Full session, four set			
K-SMT (1) vs. CS-UC (0)	2.64***	0.22	2.37 [2.16,2.56]
K-SMT (1) vs. CS-SMT (0)	1.29***	0.24	1.15 [0.94,1.37]
K-SMT (1) vs. CS-MMT (0)	1.44***	0.20	1.29 [1.11,1.46]
CS-SMT (1) vs. CS-UC (0)	1.36***	0.32	1.21 [0.93,1.50]
CS-SMT (1) vs. CS-MMT (0)	0.15	0.22	0.13 [0.00,0.33]
CS-MMT (1) vs. CS-UC (0)	1.21***	0.19	1.08 [0.91,1.25]
Five minute sessions			
K-SMT (1) vs. CS-UC (0)	1.04*	0.44	0.86 [0.50,1.23]
K-SMT (1) vs. CS-SMT (0)	0.40	0.26	0.33 [0.12,0.55]
K-SMT (1) vs. CS-MMT (0)	1.36***	0.33	1.13 [0.86,1.41]
CS-SMT (1) vs. CS-UC (0)	0.64	0.43	0.53 [0.18,0.89]
CS-SMT (1) vs. CS-MMT (0)	0.97**	0.32	0.81 [0.54,1.08]
CS-MMT (1) vs. CS-UC (0)	0.32	0.48	0.27 [0.00,0.67]
Fifteen minute sessions			
K-SMT (1) vs. CS-UC (0)	2.13***	0.39	1.64 [1.34,1.94]
K-SMT (1) vs. CS-SMT (0)	0.64*	0.25	0.49 [0.30,0.68]
K-SMT (1) vs. CS-MMT (0)	1.29***	0.28	0.99 [0.78,1.21]
CS-SMT (1) vs. CS-UC (0)	1.49***	0.38	1.15 [0.85,1.44]
CS-SMT (1) vs. CS-MMT (0)	0.65*	0.28	0.50 [0.28,0.72]

CS-MMT (1) vs. CS-UC (0)	0.84*	0.40	0.65 [0.34,0.95]
General Psychodynamic Focus			
Full session, entire set			
K-SMT (1) vs. CS-UC (0)	-0.19***	0.06	0.65 [0.45, 0.86]
K-SMT (1) vs. CS-SMT (0)	-0.02	0.04	0.07 [0.00,0.21]
K-SMT (1) vs. CS-MMT (0)	-0.04	0.04	0.14 [0.00,0.28]
CS-SMT (1) vs. CS-UC (0)	-0.17***	0.04	0.59 [0.49,0.72]
CS-SMT (1) vs. CS-MMT (0)	0.02	0.04	0.07 [0.00,0.21]
CS-MMT (1) vs. CS-UC (0)	0.15**	0.04	0.52 [0.38,0.66]
Full session, four set			
K-SMT (1) vs. CS-UC (0)	-0.21***	0.06	0.84 [0.60,1.08]
K-SMT (1) vs. CS-SMT (0)	-0.01	0.02	0.04 [0.00, 0.12]
K-SMT (1) vs. CS-MMT (0)	0.01	0.06	0.04 [0.00, 0.28]
CS-SMT (1) vs. CS-UC (0)	-0.20***	0.09	0.80 [0.44,1.16]
CS-SMT (1) vs. CS-MMT (0)	0.01	0.02	0.04 [0.00, 0.12]
CS-MMT (1) vs. CS-UC (0)	-0.21***	0.06	0.84 [0.60, 1.08]
Five minute sessions			
K-SMT (1) vs. CS-UC (0)	-0.08**	0.03	0.42 [0.26, 0.58]
K-SMT (1) vs. CS-SMT (0)	0.00	0.03	0.00 [0.00, 0.16]
K-SMT (1) vs. CS-MMT (0)	0.00	0.03	0.00 [0.00, 0.16]
CS-SMT (1) vs. CS-UC (0)	-0.08**	0.03	0.43 [0.29,0.55]
CS-SMT (1) vs. CS-MMT (0)	0.00	0.03	0.00 [0.00, 0.16]
CS-MMT (1) vs. CS-UC (0)	-0.08**	0.03	0.42 [0.26,0.58]
Fifteen minute sessions			
K-SMT (1) vs. CS-UC (0)	-0.09*	0.04	0.41 [0.23, 0.59]
K-SMT (1) vs. CS-SMT (0)	0.02	0.03	0.09 [0.00,0.23]
K-SMT (1) vs. CS-MMT (0)	0.02	0.04	0.09 [0.00,0.23]
CS-SMT (1) vs. CS-UC (0)	-0.11**	0.04	0.50 [0.32,0.68]
CS-SMT (1) vs. CS-MMT (0)	0.00	0.04	0.00 [0.00,0.18]
CS-MMT (1) vs. CS-UC (0)	-0.11**	0.04	0.50 [0.32,0.68]
General Family Focus			
Full session, entire set			
K-SMT (1) vs. CS-UC (0)	-0.38	0.22	0.28 [0.12,0.44]
K-SMT (1) vs. CS-SMT (0)	0.17	0.21	0.12 [0.00,0.28]
K-SMT (1) vs. CS-MMT (0)	-0.63**	0.23	0.46 [0.29,0.63]
CS-SMT (1) vs. CS-UC (0)	0.56	0.22	0.41 [0.25,0.57]
CS-SMT (1) vs. CS-MMT (0)	-0.80***	0.23	0.58 [0.42,0.75]
CS-MMT (1) vs. CS-UC (0)	0.24	0.25	0.18 [0.00,0.36]
Full session, four set			
K-SMT (1) vs. CS-UC (0)	-1.06***	0.25	1.02 [0.78,1.26]
K-SMT (1) vs. CS-SMT (0)	-0.11	0.24	0.11 [0.00, 0.34]
K-SMT (1) vs. CS-MMT (0)	-0.85**	0.27	0.82 [0.55,1.08]
CS-SMT (1) vs. CS-UC (0)	0.94***	0.25	0.90 [0.66,1.14]
CS-SMT (1) vs. CS-MMT (0)	-0.74**	0.27	0.71 [0.45,0.97]
CS-MMT (1) vs. CS-UC (0)	0.20	0.29	0.19 [0.00,0.47]
Five minute sessions			

K-SMT (1) vs. CS-UC (0)	-0.24*	0.11	0.27 [0.15,0.40]
K-SMT (1) vs. CS-SMT (0)	-0.03	0.12	0.03 [0.00,0.18]
K-SMT (1) vs. CS-MMT (0)	-0.10	0.12	0.11 [0.02,0.25]
CS-SMT (1) vs. CS-UC (0)	-0.22*	0.10	0.25 [0.14,0.36]
CS-SMT (1) vs. CS-MMT (0)	0.08	0.11	0.09 [0.00,0.22]
CS-MMT (1) vs. CS-UC (0)	-0.14	0.11	0.16 [0.03,0.28]
Fifteen minute sessions			
K-SMT (1) vs. CS-UC (0)	-0.41	0.24	0.39 [0.16,0.63]
K-SMT (1) vs. CS-SMT (0)	-0.11	0.23	0.11 (0.00,0.33]
K-SMT (1) vs. CS-MMT (0)	-0.49	0.26	0.47 [0.22,0.72]
CS-SMT (1) vs. CS-UC (0)	-0.30	0.23	0.29 [0.06,0.51]
CS-SMT (1) vs. CS-MMT (0)	0.38	0.04	0.00 [0.00,0.18]
CS-MMT (1) vs. CS-UC (0)	0.08	0.25	0.08 [0.00,0.32]
Client-centered			
Full session, entire set			
K-SMT (1) vs. CS-UC (0)	-0.97***	0.18	1.00 [0.81,1.19]
K-SMT (1) vs. CS-SMT (0)	-0.28	0.17	0.29 [0.11,0.46]
K-SMT (1) vs. CS-MMT (0)	-0.36	0.19	0.37 [0.18,0.57]
CS-SMT (1) vs. CS-UC (0)	-0.70**	0.18	0.72 [0.54,0.91]
CS-SMT (1) vs. CS-MMT (0)	-0.07	0.19	0.07 [0.00,0.27]
CS-MMT (1) vs. CS-UC (0)	-0.62**	0.20	0.64 [0.43,0.85]
Full session, four set			
K-SMT (1) vs. CS-UC (0)	-0.84***	0.21	0.88 [0.55,1.12]
K-SMT (1) vs. CS-SMT (0)	-0.13	0.20	0.14 [0.00,0.35]
K-SMT (1) vs. CS-MMT (0)	-0.16	0.23	0.17 [0.00,0.41]
CS-SMT (1) vs. CS-UC (0)	-0.71**	0.23	0.75 [0.51,0.99]
CS-SMT (1) vs. CS-MMT (0)	-0.03	0.23	0.03 [0.00,0.27]
CS-MMT (1) vs. CS-UC (0)	-0.67**	0.24	0.71 [0.45,0.96]
Five minute sessions			
K-SMT (1) vs. CS-UC (0)	-0.15	0.16	0.08 [0.01,0.17]
K-SMT (1) vs. CS-SMT (0)	0.13	0.16	0.07 [0.00,0.16]
K-SMT (1) vs. CS-MMT (0)	0.08	0.17	0.04 [0.00,0.14]
CS-SMT (1) vs. CS-UC (0)	0.02	0.15	0.01 [0.00,0.09]
CS-SMT (1) vs. CS-MMT (0)	-0.20	0.16	0.11 [0.02,0.20]
CS-MMT (1) vs. CS-UC (0)	-0.23	0.17	0.13 [0.03,0.22]
Fifteen minute sessions			
K-SMT (1) vs. CS-UC (0)	-0.31	0.32	0.18 [0.00,0.38]
K-SMT (1) vs. CS-SMT (0)	-0.51	0.31	0.31 [0.12,0.49]
K-SMT (1) vs. CS-MMT (0)	0.35	0.23	0.22 [0.01,0.42]
CS-SMT (1) vs. CS-UC (0)	-0.21	0.30	0.13 [0.00,0.31]
CS-SMT (1) vs. CS-MMT (0)	0.16	0.37	0.10 [0.00,0.28]
CS-MMT (1) vs. CS-UC (0)	0.05	0.32	0.03 [0.00,0.22]
VTAS-R-SF			
Full session, entire set			
K-SMT (1) vs. CS-UC (0)	0.69***	0.18	0.92 [0.68, 1.16]
K-SMT (1) vs. CS-SMT (0)	0.23	0.17	0.31 [0.08,0.53]

K-SMT (1) vs. CS-MMT (0)	0.42*	0.18	0.56 [0.32,0.80]
CS-SMT (1) vs. CS-UC (0)	0.46*	0.19	0.61 [0.36, 0.87]
CS-SMT (1) vs. CS-MMT (0)	0.19	0.19	0.25 [0.00,0.51]
CS-MMT (1) vs. CS-UC (0)	0.27	0.20	0.36 [0.09,0.63]
Full session, four set			
K-SMT (1) vs. CS-UC (0)	0.79***	0.18	1.22 [0.93,1.49]
K-SMT (1) vs. CS-SMT (0)	0.44*	0.22	0.68 [0.34,1.02]
K-SMT (1) vs. CS-MMT (0)	0.42	0.24	0.65 [0.28,1.02]
CS-SMT (1) vs. CS-UC (0)	0.35	0.19	0.54 [0.25,0.83]
CS-SMT (1) vs. CS-MMT (0)	-0.01	0.21	0.02 [0.00,0.34]
CS-MMT (1) vs. CS-UC (0)	0.37	0.21	0.57 [0.25,0.89]
Five minute sessions			
K-SMT (1) vs. CS-UC (0)	0.61***	0.20	0.76 [0.52,1.01]
K-SMT (1) vs. CS-SMT (0)	0.29	0.19	0.36 [0.13,0.60]
K-SMT (1) vs. CS-MMT (0)	0.27	0.21	0.34 [0.08,0.60]
CS-SMT (1) vs. CS-UC (0)	0.34	0.22	0.43 [0.15,0.70]
CS-SMT (1) vs. CS-MMT (0)	0.02	0.21	0.03 [0.00,0.29]
CS-MMT (1) vs. CS-UC (0)	0.31	0.19	0.39 [0.15,0.63]
Fifteen minute sessions			
K-SMT (1) vs. CS-UC (0)	0.68**	0.22	0.78 [0.53, 1.03]
K-SMT (1) vs. CS-SMT (0)	0.23	0.21	0.26 [0.02,0.51]
K-SMT (1) vs. CS-MMT (0)	0.19	0.22	0.22 [0.00,0.48]
CS-SMT (1) vs. CS-UC (0)	0.45*	0.22	0.52 [0.26,0.77]
CS-SMT (1) vs. CS-MMT (0)	-0.04	0.23	0.05 [0.00,0.31]
CS-MMT (1) vs. CS-UC (0)	0.49*	0.24	0.56 [0.29,0.84]

Note. S.E. = Standard Error; ES = effect size (Cohen's *d*); K-SMT = individual cognitive-behavioral treatment (standard manualized) delivered in Kendall et al. study; CS-SMT = individual cognitive-behavioral treatment (standard manualized) delivered in Child Steps; CS-MMT = individual cognitive-behavioral treatment (modular manualized) delivered in Child Steps. CS-UC = usual care delivered in Child Steps; * *p* < .05; ** *p* < .01; *** *p* < .001

Research question six – predictive validity. Finally, I assessed whether the efficient observational treatment integrity instrument can predict client outcomes as well as the full set instruments. I investigated: what is the shortest observation segment that predicts improvements in symptom rates across the efficient treatment integrity scores (i.e., predictive validity)?

Client-level item distributions. Before running HLM analyses, I assessed if data met normal assumptions at the client level. For the full session, entire data set, six items had skewness and kurtosis that fell within the -1.50/1.50 range (Blanca et al. 2013): *Skill-building-Adherence* (range = 1.00-5.58, skewness statistic = -0.69, kurtosis statistic = -0.29), *Skill-*

building-Competence (range = 2.25-6.08, skewness statistic = -0.44 , kurtosis statistic = -0.82), *Exposure-Adherence* (range = 1.00-4.78, skewness statistic = 0.04, kurtosis statistic = -1.36), *Exposure-Competence* (range = 2.30-6.00, skewness statistic = -0.71, kurtosis statistic = -0.78), *General Client-Centered Focus* (range = 1.83-4.52, skewness statistic = 0.64, kurtosis statistic = 0.20) and *VTAS-R-SF* (range = 0.77-3.80, skewness statistic = -0.98, kurtosis statistic = 0.99). In contrast, *General Psychodynamic Focus* and *General Family Focus* fell outside the -1.5/1.5 skewness and kurtosis range (skewness statistic = 2.22, kurtosis statistic = 5.12; skewness statistic = 1.42, kurtosis statistic = 2.30, respectively). Although *Exposure-Adherence* and *Exposure-Competence* met assumptions of normality*,* exposure was scored for only 78 of 106 clients and of those, seven were missing post-treatment symptom (CBCL) data; *Exposure-Adherence* had low average scores (*M* = 2.55) and *Exposure-Competence* had significant missing data (26.42%). Similarly, *Skill-building-Competence* had missing data (15.0%), as this item was scored for 99 out of 106 clients and an additional nine were missing post-treatment symptom data. The same pattern of client level item distributions and missing data held for full session, four set and brief segment scores. Given the significant amount of missing post-treatment symptom and treatment integrity data, predictive validity analyses were not run.

Discussion

Review of Study Aims

The goal of the current study was to take a first step in developing an observational treatment integrity instrument that balances efficiency (i.e., brief, easy to administer and use) and effectiveness (i.e., validity evidence; Glasgow & Riley, 2013; Schoenwald et al., 2011). The study assessed whether it is possible to reduce the time required to code treatment sessions for treatment integrity. If coders could produce treatment integrity ratings with strong validity evidence in shorter time frames, then results would suggest that it is worthwhile for future researchers to conduct similar, field tests of this type of instrument. The methodology for developing and systematically validating this instrument could serve as a template for future research.

Specifically, to develop the brief treatment integrity instrument, the TIES-CBT-YA, I selected items/subscales from adherence, competence, differentiation, and alliance instruments that have evidence of score validity and conceptual importance (Aim 1). Next, the TIES-CBT-YA was used to code two samples of recorded sessions: four sessions per client from an efficacy trial (Kendall et al., 2008) and four sessions per client from an effectiveness trial (Weisz et al., 2012). The coded data were then used to establish score reliability and validity (Aim 2) and draw conclusions about the optimal observational coding length (Aim 3) of the TIES-CBT-YA.

In this section, I summarize study findings, interpret relevant results, and ultimately offer implications and directions for future research. I first review my hypotheses for findings and describe whether or not the data support those hypotheses. Then, I describe whether five- or 15-minute segments better approximate the full session scores for each reliability or validity analysis. Next, I characterize items within TIES-CBT-YA, reviewing unique patterns of findings across items. Finally, I discuss implications of the findings for implementation researchers, supervisors, and clinicians and provide recommendations for future directions.

Were Hypotheses Regarding TIES-CBT-YA Development Supported?

First, I hypothesized that TIES-CBT-YA items identified from each treatment integrity domain (adherence, competence, treatment differentiation, alliance) would capture full variability in treatment integrity scores. When the range of TIES-CBT-YA items were compared to the corresponding full session, entire set scores and full session, four set scores, the ranges were comparable; all three items that had restricted range on the TIES-CBT-YA instrument also had restricted range in the respective full set instruments used in TIMS. In addition, only one item that met assumptions of normality in the full session scores did not meet assumptions of normality for the brief observation segments. Thus, TIES-CBT-YA items captured *most* of the expected variability in the full session subscales and my hypothesis was largely supported. One item, *General Psychodynamic Focus*, had particularly elevated skewness/kurtosis across full and brief data sets. This item should be interpreted with caution as it may not capture what is intended, the extent to which therapists deliver psychodynamic interventions. As aforementioned, this item was maintained in this pilot instrument for descriptive purposes, to guide future researchers interested in assessing treatment differentiation. Overall, these findings indicate that the TIES-CBT-YA instrument has the capacity to produce adherence, competence, and alliance scores that have normal distributions despite brief coding periods.

My second hypothesis was that TIES-CBT-YA items would produce scores (across observation segments) that were strongly associated with full session, four set scores. This hypothesis was only partially supported, as all TIES-CBT-YA items except *General Client-Centered Focus* and *General Psychodynamic Focus* were strongly correlated with the full session, four set scores. Thus, the TIES-CBT-YA items intended to assess the delivery of ICBT for youth anxiety (i.e., adherence, competence, alliance) appeared to have the largest association with the full session scores. Past studies that have compared adherence and competence to ICBT across brief and full session segments have similarly found that the scores fall within the “large” association range (Weck et al., 2014). However, no known previous studies have compared treatment differentiation scores across brief and full session segments. These items did not meet assumptions of normality or establish sufficient reliability in this study so it is difficult to interpret this finding. It could be that these items have poor psychometric strength for five- or 15-minute segments of an ICBT session. On the other hand, the items may not have been observed enough times to perform a fair test. Future researchers are encouraged to test similar differentiation items, with different samples of clients (with anxiety), therapists, or coders to ascertain whether these items could be retained in future pragmatic observational treatment integrity instruments.

My final hypothesis was that TIES-CBT-YA items would be able to discriminate treatment integrity across treatment settings and treatment types (ICBT, UC). I therefore investigated the extent to which group comparison findings for five- and 15-minute segment scores approximated those for full session scores in significance levels and effect sizes. Overall, my hypothesis was partially supported. Two out of eight items had the same pattern of significance and magnitude for the group comparison findings as full session scores. Though effect sizes were generally lower in the brief segment scores than the full session scores across items, a few brief segment, non-significant group comparisons had “small” to “medium” effect sizes when compared to the full session group comparisons that were significant; thus, the small sample size may have influenced a few of the statistical significance differences between data sets. Nevertheless, at least half of the items did not predict the same pattern of significance or magnitude for treatment differences across brief and full session segments. It could be that watching only a specific portion of the session limits access to subtle, systematic differences in ICBT intervention delivery. For example, perhaps therapists are generally more similar across ICBT programs when they are engaged in the “work” phase (i.e., middle portion) of treatment. Conversely, it could be that seven out of the 12 items selected for TIES-CBT-YA (i.e., two subscales of the adherence instrument [CBAY-A; Southam-Gerow et al., 2016]; two subscales of the competence instrument [CBAY-C; McLeod et al., 2016], three items of the treatment differentiation instrument [TPOCS-RS; McLeod et al., 2015]) were coded differently for brief segments than the full scale subscales/items and thus have low discriminant validity. Future researchers interested in using TIES-CBT-YA may consider selecting different portions of the session for coding or retesting the discriminant validity with different samples to see if these findings are limited to the current sample.

I also expected that therapists in the ICBT conditions would have significantly higher adherence and competence scores than therapists in UC across brief observation segments, given therapists' training differences. These expectations were not met for both brief and full session, four set scores on *Exposure-Adherence;* there was no difference between the dosage of therapists' exposure delivery in CS-SMT vs. CS-UC for all three of these data sets. In contrast, the expected difference between CS-SMT vs. CS-UC exposure intervention was present for the full session, entire set scores. Past research has found that five to ten sessions must be included in analyses to produce stable estimates of treatment integrity in cocaine dependency psychosocial treatment (Dennhag, Gibbons, Barber, Gallop, & Crits-Christoph, 2012). If this finding generalizes across treatments, the four sessions selected for this study may have limited the discriminant validity of TIES-CBT-YA. Alternatively, it could be that the scoring strategy for *Exposure-Adherence,* taking the highest score in the subscale, influenced group comparison findings. For example, on average therapists in CS-UC may have delivered minimal doses of exposure throughout the brief segment (e.g., rating of 1.5), but when they did deliver some exposure (e.g., rating of a 3), it may have been at a high enough dose to approximate the average dose delivered by therapists in CS-SMT. Finally, current study results indicated that exposure interventions were delivered, on average, at low doses across all treatment groups. Past researchers have theorized that exposure interventions may be more challenging to deliver (Levita, Duhne, Girling, & Waller, 2016), so it could be that a number of therapists, regardless of treatment group, had difficulty delivering these interventions or felt uncomfortable doing so (Garland et al., 2010). Low variability in exposure dose may have contributed to the similarities across theoretically distinct treatment groups. Therefore, exposure dose may be an area primed for more intensive therapist training or supervisory feedback, and thus an important item to assess in a pragmatic observational treatment integrity instrument. Overall, my expectations for treatment differences were not supported and warrant investigation in future studies.

Which Segment Best Approximated Full Session Scores for Each Validity Analysis?

Item distribution. Both five- and 15-minute observation segments had relatively congruent patterns of item distributions; most notably, the same three items did not meet assumptions of normality across observations. However, the five-minute segments had less client level outliers for the *General Client-Centered Focus* item. Past studies have found that raters make accurate ratings of teacher competence in less than 30 seconds (Ambady & Rosenthal, 1993). It could be that the coders of the five-minute segments made more accurate ratings of an intervention less likely to change over the course of a session (i.e., therapists' warmth). In addition, coders may have made ratings in a more systematic way (i.e., a more normal distribution pattern) for this item when they had less information about the therapist (e.g., not having to consider how the therapist delivers other interventions; Weck et. al., 2014). Overall, these findings suggest that the length of the brief segment coded (i.e., five- or 15-minutes) had a trivial influence on item distributions. Future research could investigate whether the type of interventions selected for pragmatic observational treatment integrity instruments influences the minimal observation time required to make accurate treatment integrity ratings.

Reliability evidence. There were some differences in reliability scores across observation segments. For example, five-minute segments had one more item (than 15-minute segments) that fell in the "poor" agreement range. In addition, six items in the 15-minute segments fell within one agreement category of the full session reliability scores; in contrast, only two items in the five-minute segments fell within one agreement category of the full session reliability scores. Thus, the 15-minute segment reliability scores somewhat better approximated the full session reliability scores than the five-minute scores. It is possible that coders were more likely to agree on ratings when they had a longer time frame and more opportunities to detect a given intervention or skill (e.g., if they did not detect that the therapist was engaging in skill-building at one time-point, they may detect it during the second time-point [10 minutes later], and score it similarly to the other coder who saw both interactions). Of note, both observation segments produced adherence and competence inter-rater agreement scores that fell within the "fair" agreement range or above. This finding parallels a past study that found some items had less than "good" agreement when comparing the adherence and competence scores in the 18-20-minute middle segment of ICBT treatment to the full session adherence and competence scores (Weck et al., 2014). Similar to that study, I used an ICC Model 1 (due to only a portion of sessions being double-coded), which provides a lower reliability estimate than ICC Model 2 (used when all sessions are double-coded; Shrout & Fleiss, 1979). Indeed, by using the ICC Model 1, at least three items within each brief observation segment fell one agreement category below what the agreement category would have been using the ICC Model 2. Future investigations could explore how brief treatment integrity inter-rater reliability scores differ using less limited reliability analysis methods.

Sensitivity to change. For both observation segments, most items resembled the full session, four set scores in the magnitude and direction of intervention dose change. Nonetheless, 15-minute segment scores had one more item than the five-minute segment scores that better resembled the full session, four set scores in growth model fit and slope direction. Thus, the 15-minute segment appeared to *slightly* better approximate the full session scores on sensitivity to change than the five-minute segment. However, both segments had only one item with the same quadratic intervention trajectory as the full session, entire set scores, *Exposure-Adherence*; therapists delivered exposure dosages increasingly as the course of treatment progressed and their exposure intervention delivery stayed the same or dipped slightly at the end of treatment. It makes sense that therapists would deliver more exposure interventions as the course of treatment progressed, given that exposures are intended to be delivered in the second half of ICBT for youth anxiety (Kendall & Hedtke, 2006). Thus, since the ICBT protocol encourages therapists to deliver skill-building interventions in the first half of treatment, it is somewhat surprising that neither the five- nor the 15-minute segments exhibited significant changes in *Skill-building-Adherence* over time (as they did for both full session set scores). Watching brief segments of sessions may limit coders' ability to accurately assess extensiveness of skill-building interventions at different stages of treatment. For example, it would be expected that coders who watch an entire session would see both skill-building and exposure interventions delivered late in treatment, but it is likely that they would see a higher proportion of exposure interventions than they would see skill-building interventions. This expectation was met for the full session, four set scores, as therapists' dose of skill-building interventions decreased as treatment progressed. Perhaps a coder who only watches the middle portion of ICBT sessions that are late in treatment (e.g., sessions 10-16) would mostly see therapists preparing clients for exposure tasks by engaging in skill-building (e.g., reminding the client to take deep breathes or use coping thoughts during the exposure). Though the 15-minute segments appeared marginally more sensitive to change than the five-minute segments, neither had as much change as would be expected given the entire set scores.

Analyzing the full and brief sessions over the course of four sessions (rather than the full course of treatment; i.e., 16-20 sessions) may have limited my ability to detect intervention delivery change. Only three of eight TIES-CBT-YA items exhibited change over time when using four sessions, whereas seven of eight items changed over time when using the entire set of sessions. This finding is discrepant from a past study that detected change in intervention delivery when 96.0% of the clients had four sessions or less available for coding (Boswell et al., 2013) and a past study that detected change in intervention delivery with four assessment points (Henggeler et al., 2008). Researchers have recommended using at least three observations for growth curve modeling, so it is possible that the three clients who only had two sessions in the current study contributed to this assessment limitation (Curran, Obeidat, & Losardo, 2010). No previous study has compared specific intervention trajectories from entire data sets to a sample of sessions, so it is difficult to know whether my sampling method (i.e., choosing two sessions from the skill-building phase of treatment and two from the exposure phase of treatment) was flawed or if any sample of sessions less than the full set would similarly limit trajectory data. Again, it may be that more sessions were needed to establish accurate treatment integrity ratings across the course of treatment (Dennhag et al., 2012). However, it is important to note that the TIES-CBT-YA items were able to capture complex intervention delivery trajectories when the full session, entire set of sessions were analyzed. The only intervention that did not change over time was *General Client-Centered Focus*, which has been found to have no variability over time with a similar client sample and session set (Smith et al., in press). This trajectory is consistent with the ICBT protocol, which encourages therapists to provide consistent warmth and validation throughout the course of treatment (Kendall & Hedtke, 2006). However, it could also mean that this item has little clinical utility for those researchers interested in developing a supervisory feedback tool (i.e., who need to assess therapist's incremental improvement over the course of training; McLeod et al., 2013). Findings across items may indicate that researchers or supervisors who would like to use the TIES-CBT-YA instrument for regular feedback purposes may need to watch the full session, weekly.

Convergent validity. Overall, all TIES-CBT-YA items (across both observation segments) except *General Client-centered Focus* and *General Psychodynamic Focus* had "large" associations with the full session, four set scores. Most TIES-CBT-YA items significantly corresponded with the items that they were intended to be theoretically similar. Thus, five- and 15-minute observation segments had comparable convergent validity findings. This finding is consistent with a past study, that found that all adherence and competence items had "large" associations with the full session scores (Weck et al., 2014). Future supervisors, clinicians, or researchers thus have some evidence that including these four adherence/competence items in instruments that assess treatment integrity one time, in the middle of treatment sessions, may be worthwhile.

Discriminant validity. Overall, the 15-minute scores approximated the full session scores in significance and magnitude for a few more group comparison findings than the five-minute sessions. The five-minute scores also did not meet an important group comparison expectation that the 15-minute segments met. For five-minute segments (unlike all other session segments), therapists in CS-MMT did not significantly differ from therapists in CS-UC in the skill with which they delivered skill-building interventions. As mentioned, it may be that coders are able to assess subtle, but important differences in treatment delivery when they spend more time watching the session (Weck et al., 2014). Even 15-minute segment scores did not have the same pattern of group findings as full session scores for most items. This finding appears most important for researchers who need to use such an instrument for treatment group comparisons in randomized controlled studies; it could be that researchers who require the highest level of validity in their findings need to watch the entire session.

Predictive validity. Due to missing data, the predictive validity (i.e., the extent to which TIES-CBT-YA items predicted criterion-related scores at a future time point) of TIES-CBT-YA as a whole and the differences between five- and 15-minute observation segments could not be concluded. There have been mixed findings about the link between treatment integrity and client outcomes (Webb et al., 2010; Weck et al., 2014), so conclusions about how brief and full session treatment integrity scores predict outcomes would have been an important contribution to the literature. Moreover, some researchers would argue that predictive validity is one of the most important validity dimensions for treatment integrity assessment (e.g., Bond, Evans, Salyers, Williams, Kim, 2000). Thus, these findings may limit interpretations about the accuracy of TIES-CBT-YA in general. However, the pattern of missing data also raises questions for researchers interested in coding exposure sessions; for example, it could be more likely that coders rate exposures as “not present” or “somewhat” present when they do not know an exposure will be occurring in a segment (i.e., the design in TIMs and the current study) versus when they know an exposure will be occurring in a segment and are instructed to rate the adherence and competence levels of that exposure. Using this rationale, exposures may be most unlikely to be scored at high extensiveness or skill levels in random, five-minute segments, when coders have the least context clues about what may be occurring in a session (e.g., they need to decipher whether an interaction is a normal conversation versus a social exposure). Future researchers should consider manipulating this aspect of the study design (i.e., knowledge of which intervention will occur in the session) to better understand whether exposures can be accurately coded in a brief session segment. This would also minimize the likelihood that an exposure occurs outside of a session room, in turn, is not recorded, and thus is not coded altogether. Overall, future researchers who use this pilot study as a guide should include this validity dimension in their study to enhance study conclusions, drive treatment integrity assessment research, and potentially foster future implementation efforts.

How Did TIES-CBT-YA Items Perform Across Observation Segments?

Across TIES-CBT-YA observation segments, the *Skill-building-Adherence* item was the most similar to full session, four set scores. This item had normal distributions, little heterogeneity of the data, ICCs that fell within the "excellent" agreement range, a "large" association with full session scores, the same pattern of treatment group differences as full session scores, and the same pattern of predictions as full session scores. Researchers have wondered whether competence is more difficult to code than adherence (given that competence has a more complex definition and often requires adherence but not vice-versa; Barber, Sharpless, Klostermann, & McCarthy, 2007); it could be that adherence was easiest to code and that skill-building occurred more frequently than exposure interventions within the brief segments, allowing for more variability and consistency in *Skill-building-Adherence* scores. Alternatively, this item may have best represented the individual items within the full scale instrument (e.g., *Relaxation*, *Psychoeducation*; Southam-Gerow et al., 2016) that contributed to the full session subscale. Regardless, this item appears to be well-suited for inclusion in observational treatment integrity instruments that assess middle portions of sessions.

In contrast, all three TIES-CBT-YA treatment differentiation items did not meet assumptions of normality across five- and 15-minute segments and had low reliability and validity scores compared to the rest of the items. These items were retained in the sample due to the nature of this pilot study, to guide future researchers who are considering developing a pragmatic observational treatment integrity instrument. Importantly, *General Family Focus* and *General Psychodynamic Focus* did not meet assumptions of normality when the entire set of full sessions were scored (i.e., the TPOCS-RS general focus items validated in the TIMS; McLeod et al., 2016). These findings may suggest that family and psychodynamic interventions were delivered in small doses, to a small sample of clients, throughout the course of treatment. Given that therapists in three out of four treatment groups were trained in the ICBT protocol, interventions within psychodynamic and family domains may have been more skewed because therapists in ICBT groups did not diverge from the protocol (as intended); family and psychodynamic interventions may have thus been more infrequent in the shorter segments. Indeed, a past study that used the TPOCS-RS with TIMS data found that therapists delivering ICBT for youth anxiety delivered the strongest dose of cognitive-behavioral interventions, followed by client-centered, family, and psychodynamic interventions (Smith et al., in press). An additional TIMS study using the TPOCS-RS found that therapists in ICBT groups delivered family and psychodynamic interventions at low doses throughout the course of treatment; these authors encouraged future researchers to test this instrument with specific populations assigned to receive family and psychodynamic interventions (McLeod et al., 2016). Researchers interested in studying ICBT may therefore consider whether the sample they select warrants assessment of these types of differentiation items; it may be that therapists in ICBT rarely deliver these types of interventions.

General Client-Centered Focus performed most poorly in brief segments compared to the full session scores. Whereas *General Client-Centered Focus* in the full session, four set scores had normal distributions, “fair” inter-rater agreement, and validity scores that resembled entire set scores, *General Client-Centered Focus* in five- and 15-minute segments had outliers, “poor” agreement, no association with full session scores, and a slightly different pattern of group comparison findings when compared to entire set scores. Unlike the other differentiation items, this item was frequently coded, likely because therapist warmth, positive regard, and validation are encouraged in the ICBT protocol. There are a number of reasons why scores on this item may not have been as reliable and valid for the brief observation segments. For example, it may be difficult to assess some client-centered interventions (e.g., validation) in the middle segment of manualized ICBT sessions (e.g., due to therapists being focused on delivering the core ICBT components during the “work” phase of treatment) whereas other client-centered interventions (e.g., warmth) are easier to assess regardless of session portion (McLeod et al., 2013; McLeod & Weisz, 2010) . In addition, although coders achieved an ICC of .59 or above during certification, they were certified based on their agreement with expert opinions on full session codes; coders therefore may not have reached the level of agreement needed for certification if they had rated client-centered interventions in brief segments. Future testing of this item could ascertain if this finding regarding client-centered interventions generalizes across samples and certification procedures.

Strengths and Limitations

The current study had a number of strengths. First, I compared the same sample of clients used to code TIES-CBT-YA brief segments to the previously coded TIMS full sessions (both four session and 16-20 session sets); this enhanced my ability to interpret whether differences between brief and full scores could be related to session length or session number. In addition, the current study used the full VTAS-R-SF instrument to constitute the overall alliance score of the TIES-CBT-YA instrument; using all of these items likely improved the reliability and validity scores for alliance and enhanced implications for future research. Indeed, the VTAS-R-SF may be well-suited for integration into future pragmatic treatment integrity instruments. Finally, HLM analyses were conducted which allowed the current study to (a) account for missing data, (b) account for nesting, and (c) assess sensitivity to change; these factors are integral to determining whether the instrument can be used to detect differences or changes across clients, therapists, and time (Singer & Willet, 2003). On the whole, study findings underscore the importance for future researchers to replicate this instrument development process using the aforementioned methods that fortified the study's design.

Despite the advantages of the current study's design, there were some limitations that should be considered. First, coding resources were limited. Only 10% of sessions were double-coded, four sessions per client were coded, and additional time segments (i.e., 10, 20 minutes) could not be coded due to coder time and budget constraints. This may have limited the generalizability of findings. For example, some double-coded items had no variance; it is difficult to interpret whether the observation segment length or the small sample of double-coded sessions influenced these scores. Second, only four sessions from each client were coded; thus, the current study did not capture the full course of treatment. Given that full session, four set item scores exhibited significantly less change over time than full session, entire set scores, selecting four sessions likely limited the sensitivity to change analyses. Lastly, this instrument was developed in the context of assessing ICBT for youth anxiety; thus, treatment integrity results may not generalize to other treatment domains. Given that the stability of treatment integrity scores can vary based on type of therapeutic interventions delivered (Dennhag et al., 2012), differences might arise for developing pragmatic observational treatment integrity instruments across treatment domains. In general, though the current pilot study had some limitations, there were sufficient resources to ask initial questions and raise issues to be investigated in future studies.

The current study also had methodological limitations that are important to note. The TIES-CBT-YA item selection and scoring approach for each full scale item were based on expert opinion and conceptual rationale, and differed depending on the type of subscale (e.g., adherence and competence were scored by taking the highest item within the subscale; alliance was scored by averaging items). This is one way to approach item selection and subscale scoring (out of numerous potential approaches), and thus should be taken into account when interpreting findings. For example, given that there is no established way to score subscales on these treatment integrity variables (McLeod et al., 2013; Smith et al., in press), future researchers may consider testing other scoring strategies. Further, due to limited resources, I engaged in the coding even though I am the author; although I was blind to treatment condition while coding, I understood the purpose of the study and therefore may have been biased when making ratings (e.g., unintentionally rating five- or 15-minute ratings in a way that would correspond to the full session ratings). Finally, due to the significant amount of missing data in post-treatment CBCL scores, a number of items could not be assessed for predictive validity. This would have provided information about how well the instrument can generalize outside of treatment – to predict client outcomes, an important domain in theoretical constructs of treatment process (Fjermestad et al., 2015; Weck et al., 2014). Overall, current study findings should be interpreted cautiously by readers and researchers are encouraged to avoid these methodological issues in future studies.

Implications for Researchers and Clinicians

Some researchers have proposed that one way to resolve a number of treatment integrity assessment limitations is to carefully consider the purpose of the instrument and use of the findings (Schoenwald et al., 2011). For example, if researchers decide to use one treatment integrity instrument for evaluating therapist performance for payment/hiring, they may consider using or developing a treatment integrity instrument that focuses more on effectiveness (i.e., stronger validity evidence, accuracy) than efficiency (i.e., brief, easy to use; Schoenwald et al., 2011). Conversely, if researchers decide to use treatment integrity instruments for brief, informal training, as implementation researchers often elect to do in practice settings (e.g., providing therapists feedback during initial EBT training), they may consider using or developing an instrument that is more efficient than effective. The current study developed an instrument that, when considered as one tool, leaned more to the efficient side of the spectrum than the effective side. TIES-CBT-YA is brief and takes less than 20 minutes to code; however, three TIES-CBT-YA item scores did not meet established criteria for acceptable reliability and validity when the session was coded in 15-minute segments or less. In addition, most items were not sensitive to change over time or did not have the same pattern of treatment group differences when compared to full session scores. Thus, researchers or supervisors interested in assessing therapist's improvement in skills after each session or comparing treatment groups may need to code more than four sessions to test if TIES-CBT-YA, or a similar instrument, can even establish these validity dimensions in shorter time segments. Researchers and supervisors are encouraged to weigh the strengths and limitations of TIES-CBT-YA when considering the purpose of their study or instrument goal, whether that be investigating aspects of this pilot instrument with other samples, developing alternative pragmatic observational treatment integrity instruments, or using this study as a guide to informally assess therapist behavior in training.

Research implications. Given the pilot nature of this study, there are a number of implications for future research. Some TIES-CBT-YA items, such as adherence to skill-building interventions, competence in delivering skill-building interventions, and the child-therapist alliance, met assumptions of normality, established at least "fair" reliability (ICCs = .49 - .79), and had "large" associations with full set scores (e.g., *r*s = .50 - .72) when coded for 15-minutes. Researchers have suggested that not enough studies have measured the interactions of adherence, competence, and alliance (Barber et al., 2007). This is likely because researchers use three separate, lengthy observational instruments to assess these domains (e.g., TIMS; NIMH RO1 MH086529), and managing three different coding teams may be too difficult or expensive for most researchers. Pragmatic observational treatment integrity instruments with evidence of score construct validity have the potential to make this investigation feasible, by combining these domains into a brief tool. Given that TIES-CBT-YA did not detect significant change over time, future researchers who do not have resources to code more than four sessions could use this study design as a guide to develop an instrument that streamlines the one-time observational assessment of internal validity for their study (i.e., was the independent variable [adherence, competence, alliance] manipulated as intended?; Kazdin, 2003). They could also use this study as a guide to develop an instrument for the purpose of increasing stakeholder investments in EBT implementation (McLeod et al., 2013; Schoenwald et al., 2011); researchers could show stakeholders observational treatment integrity data after study completion to persuade them to increase their investments in implementation efforts. Lastly, future researchers may investigate (a) *if* TIES-CBT-YA psychometric findings replicate across samples, (b) *why* certain TIES-CBT-YA items exhibit higher reliability and validity scores than others, and (c) *what that means* for pragmatic treatment integrity instrument development and application. For example, it could be that only the items that were coded frequently, using the full range of the scale in full session instruments, are worthwhile for inclusion in pragmatic observational treatment integrity instruments. However, this speculation will not be supported until this study can be replicated with other samples.

EBT supervision and training implications. Supervisors and clinicians involved in EBT training can also use this study as a roadmap for treatment integrity instrument use and development. For example, current study findings indicated that trained graduate student coders watched five-minute segments of ICBT sessions and still gave therapists' adherence, competence, and alliance ratings that were strongly associated with the ratings of the full session. They watched 15-minute segments and still rated therapists' adherence, competence, and alliance in a way that other trained coders generally agreed upon. Implementation studies have found that feedback, modeling, and role-playing enhance EBT performance (Beidas & Kendall, 2010; Miller et al., 2004). Thus, supervisors who aim to use evidence-based methods of training, and give therapists informal feedback during their initial ICBT for youth anxiety training (i.e., they do not need to compare ratings across treatment groups or assess how well performance maps onto future outcomes), may consider using certain TIES-CBT-YA items to provide feedback. Using TIES-CBT-YA items would allow supervisors to optimize their time in role-plays and feedback, rather than methods that have less influence on EBT performance (e.g., didactic training; Miller et al., 2004). However, supervisors and trainers will need to wait until TIES-CBT-YA exhibits stronger validity evidence before using it for purposes that require assurance of accuracy, such as therapist certification or overall evaluation of the ICBT performance in community contexts (Carroll et al., 2002; Godley et al., 2011). Nevertheless, study findings provide initial evidence that supervisors and clinicians may be able to use pragmatic observational treatment integrity instruments, like TIES-CBT-YA, to improve EBT implementation assessment while still conserving time and resources.

Future Directions

Future implementation researchers can use findings from this pilot study to answer the next set of questions. For one, only four brief sessions were coded in the current study and only one brief session was coded in a similar, previous study (Weck et al., 2014); future research should assess how manipulating the number of sessions included in reliability and validity analyses influence a pragmatic treatment integrity instrument's psychometric strength, as this could have significant implications for research and practice. Second, two brief time-segments were coded in the current study and only one was coded in past studies (Weck et al., 2011; Weck et al., 2014); future researchers could further manipulate segment minutes to ascertain the optimal number needed to approximate full session reliability and validity scores. Third, I elected to use expert opinion and conceptual rationale to select items for TIES-CBT-YA, resulting in a few items that had skewed distributions and biased findings; future researchers may consider using both empirical and conceptual rationale to select items for pragmatic instruments, to avoid this issue. Finally, I employed four graduate student coders and double-coded 10% of sessions; future researchers are encouraged to select reliability analyses based on the purpose of their study.

In regards to manipulating the number of sessions coded, if future researchers could code all sessions using a pragmatic observational treatment integrity instrument, then include decreasing numbers of sessions in subsequent analyses, they could assess how many sessions produce optimal psychometric properties; this could define how many sessions are necessary to establish stable estimates of pragmatic treatment integrity ratings (Dennhag et al., 2012). Sensitivity to change and discriminate validity analyses appeared to vary across full session, entire set (i.e., approximately 16 sessions) scores and full session, four set scores in the current study; however, it is unknown how the brief segment scores of four sessions would compare to the brief segment scores within the entire set of sessions. Moreover, this investigation may have important clinical applications. For example, this could dictate the number of sessions that supervisors need to watch to use the instrument to (a) assess therapists' session-by-session behavior change and (b) provide constructive feedback. This type of research could improve the field's understanding of how many sessions are needed to establish stable estimates of both full treatment integrity ratings and pragmatic treatment integrity ratings.

For concurrent reasons, future research could manipulate coding minutes and coding segments (i.e., early, middle, end, random segments of the session) to ascertain the minimum or appropriate segment needed for the pragmatic treatment integrity instrument to achieve strong psychometric properties. The current study found that 15-minute segments slightly better resembled full session scores in reliability, sensitivity to change, and discriminant validity scores; however, the 15-minute segments had a number of differences from full session scores in those domains. Thus, it could be, for example, that coding 20 or 25 minutes is optimal, or coding random 15-minute segments produces better reliability or validity scores than other time points. If researchers and supervisors knew which portions of sessions to code for validity or feedback purposes, this may improve the generalizability of findings and ultimately enhance implementation efforts.

In addition, researchers may consider using both conceptual and empirical rationale to select items for future pragmatic observational treatment integrity instruments. A number of researchers in previous studies have used factor analysis or item response theory to systematically select items for inclusion in treatment integrity instruments, and have recommended this approach to item selection (Moyers, Manuel, Henrickson, & Miller, 2005; Shelef & Diamond, 2008). Using empirical rationale could allow researchers to identify items that are poor indicators of the underlying treatment integrity construct (Devellis, 2012). It would be interesting if this method would consistently eliminate treatment differentiation items across treatment types, given that these are not central to the treatment protocol.

Finally, future researchers may consider tailoring their reliability analyses based on their need to prioritize the reliability of their instrument versus their need to reduce resources and be feasible within clinical contexts. Researchers who have purposes that lean toward the effectiveness side of the spectrum may elect to double-code all sessions and employ expert coders blind to condition *and* study purpose in order to reduce coder bias and optimize reliability estimates. Conversely, implementation researchers who would like to use the pragmatic observational treatment integrity instrument in busy, community contexts may ask local supervisors to be the coders for the study (e.g., Ball et al., 2009; Martino et al., 2011), only requiring them to double code 10% of the sessions and using a one-way ICC model. Employing supervisor coders may enhance external validity of the findings and, perhaps, the sustainability of the instrument for use after the contracted EBT experts withdraw resources or consultation infrastructure from the clinic (Dorsey et al., 2013; Hogue et al., 2013).

Conclusion

Implementation researchers are attempting to change the fact that the current measurement of key implementation constructs has been convoluted by inefficient instruments, suited for use in research settings, but not always appropriate for use in community contexts. By developing more efficient (i.e., cost and time-effective) instruments, such as therapist self-report treatment integrity instruments, researchers aimed to resolve this issue; however, the validity evidence of self-report instruments has been questioned due to their subjective nature and the less than adequate correspondence with observational ratings (e.g., Chapman et al., 2013; Hogue et al., 2012; Martino, et al., 2009). The current study aimed to pilot the development of an efficient observational treatment integrity instrument for ICBT for youth anxiety. I developed a brief, 12-item observational treatment integrity tool from four observational treatment integrity instruments with validity evidence. I also characterized how the instrument performed in line with my expectations, and how the tool performed across two different brief coding segments, from the middle of treatment sessions. Finally, I presented future researchers, supervisors, and clinicians with a picture of how this study could be used as a roadmap for future pragmatic observational treatment integrity instrument development, with the hope that future researchers replicate aspects of the study with other samples and proposed methods. By doing so, the current study brings the field a step closer to enhancing EBT implementation assessment. We need to continue to develop treatment integrity instruments that are both efficient and effective in practice settings so that we can improve implementation efforts and ultimately enhance the quality of care provided to millions of individuals who seek care in community settings.

List of References

Aarons, G. A., Hurlburt, M., & Horwitz, S. M. (2011). Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*(1), 4–23. doi:10.1007/s10488-010-0327-7

Achenbach, T. M. (1991). *Manual for the child behavior checklist 4–18 and 1991 profile.* Burlington, VT: Department of Psychiatry, University of Vermont.

Achenbach, T. M., & Rescorla, L. (2001). *Manual for the ASEBA school-age forms and profiles.* Burlington: University of Vermont, Center for Children, Youth and Families.

Allen, J. A., 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, NY: Oxford University Press.

Ambady, N. & Rosenthal, R. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. Journal of Personality and Social Psychology, 64, 431-441. doi:10.1037/0022-3514.64.3.431

American Psychological Association. (2001). Publication manual of the American Psychological Association (5th ed.). Washington, DC: Author.

Ball, S. A., Martino, S., Nich, C., Frankforter, T. L., Van Horn, D., Crits-Christoph, P., … Carroll, K. M. (2009). Site matters: Multisite randomized trial of motivational enhancement therapy in community drug abuse clinics. *Journal of Consulting and Clinical Psychology*, *75*(4), 556-567. doi:10.1037/a0015304

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

Barrington, J., Prior, M., Richardson, M., & Allen, K. (2005). Effectiveness of CBT versus standard treatment for childhood anxiety disorders in a community clinic setting. *Behavior Change, 22*, 29-43. doi:10.1375/bech.22.1.29.66786

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:10.1111/j.1468-2850.2009.01187.x

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

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health, 25,* 464-469. doi:10.1111/j.1467-842X.2001.tb00294.x

Berwick, D. M. (2005). Broadening the view of evidence-based medicine. *Quality & Safety in Health Care*, *14*, 315–316. doi:10.1136/qshc.2005.015669

Blanca, M. J., Arnaue, J., Lopez-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology, 9,* 78-84. doi:10.1027/1614-2241/a000057

Bond, G. R., Evans, L., Salyers, M. P., Williams, J., & Kim, H. (2000). Measurement of fidelity in psychiatric rehabilitation. *Mental Health Services Research, 2,* 75-87. doi:10.1023/A:1010153020697

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research and Practice*, *16*(3), 252–260. doi:10.1037/h0085885

Boswell, J. F., Gallagher, M. W., Sauer-Zavala, S. E., Bullis, J., Gorman, J. M., Shear, M. K., … Barlow, D. H. (2013). Patient characteristics and variability in adherence and competence in cognitive-behavioral therapy for panic disorder. *Journal of Consulting and Clinical Psychology*, *81*(3), 443–454. doi:10.1037/a0031437

Braswell, L., Kendall, P. C., Braith, J., Carey, M. P., & Vye, C. S. (1985). “Involvement” in cognitive-behavioral therapy with children: Process and its relationship to outcome. *Cognitive Therapy and Research*, *9*(6), 611–630. doi:10.1007/BF01173021

Carpenter, K. M., Cheng, W. Y., Smith, J. L., Brooks, A. C., Amrhein, P. C., Wain, R. M., & Nunes, E. V. (2012). “Old dogs” and new skills: How clinician characteristics relate to motivational interviewing skills before, during, and after training. *Journal of Consulting and Clinical Psychology*, *80*(4), 560–573. doi:10.1037/a0028362

Carroll, K. M., Farentinos, C., Ball, S. A., Crits-Christoph, P., Libby, B., Morgenstern, J., … Woody, G. E. (2002). MET meets the real world: Design issues and clinical strategies in the Clinical Trials Network. *Journal of Substance Abuse Treatment*, *23*, 73–80. doi:10.1016/S0740-5472(02)00255-6

Chamberlain, P., Roberts, R., Jones, H., Marsenich, L., Sosna, T., & Price, J. M. (2012). Three collaborative models for scaling up evidence-based practices. *Administration and Policy in Mental Health and Mental Health Services Research*, *39*, 278–290. doi:10.1007/s10488-011-0349-9

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7–18. doi:10.1037/0022-006X.66.1.7

Chapman, J. E., McCart, M. R., Letourneau, E. J., & Sheidow, A. J. (2013). Comparison of youth, caregiver, therapist, trained, and treatment expert raters of therapist adherence to a substance abuse treatment protocol. *Journal of Consulting and Clinical Psychology*, *81*(4), 674–80. doi:10.1037/a0033021

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., Riese, S., Weisz, J. R., Grubbs, K., Becker, K., Krull, J. L., The Research Network on Youth Mental Health. (2010). Evaluation of the Brief Problem Checklist: Child and caregiver interviews to measure clinical progress. *Journal of Consulting and Clinical Psychology, 78,* 526-536. doi:101.1037/a0019602

Chorpita, B. F., & Weisz, J. R. (2005). *Modular approach to therapy for children with anxiety, depression, or conduct problems.* Honolulu, Hi, and Boston, MA: University of Hawaii at Manoa and Judge Baker Children's Center, Harvard Medical School.

Chu, B. C., & Kendall, P. C. (2004). Positive association of child involvement and treatment outcome within a manual-based cognitive-behavioral treatment for children with anxiety. *Journal of Consulting and Clinical Psychology*, *72*(5), 821–829. doi:10.1037/0022-006X.72.5.821

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

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

Cumming, G., & Finch, S. (2005). Inference by eye. *American psychologist, 60*, 170-180. doi:10.1037/0003-066X.60.2.170

Curran, P. J., Obediat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognitive Development, 11,* 121-136. doi:10.1010/15248371003699969

Decker, S. E., Carroll, K. M., Nich, C., Canning-Ball, M., & Martino, S. (2013). Correspondence of motivational interviewing adherence and competence ratings in real and role-played client sessions. *Psychological Assessment*, *25*(1), 306–12. doi:10.1037/a0030815

Dennhag, I., Gibbons, M. B. C., Barber, J. P., Gallop, R., & Crits-Christoph, P. (2012). How many treatment sessions and patients are needed to create a stable score of adherence and competence in treatment of cocaine dependence? *Psychotherapy Research, 22,* 475–488. doi:10.1080/10503307.2012.674790

DeVellis, R. F. (2012). *Scale development: Theory and applications*. London, England: Sage Publications.

DiMatteo, M. R. (2004). Variations in patients' adherence to medical recommendations: A quantitative review of 50 years of research. *Medical Care*, *42*, 200–209. doi:10.1097/01.mlr.0000114908.90348.f9

Dorsey, S., Pullmann, M. D., Deblinger, E., Berliner, L., Kerns, S. E., Thompson, K., … Garland, A. F. (2013). Improving practice in community-based settings: a randomized trial of supervision - study protocol. *Implementation Science*, *8*, 89-100. doi:10.1186/1748-5908-8-89

Doss, B. D. (2004). Changing the way we study change in psychotherapy. *Clinical Psychology: Science and Practice, 11*, 368-386. doi:10.1093/clipsy/bph094

Eugster, S. L., & Wampold, B. E. (1996). Systematic effects of participant role on evaluation of the psychotherapy session. *Journal of Consulting and Clinical Psychology*, *64*(5), 1020–1028. doi:10.1037/0022-006X.64.5.1020

Feingold, A. (2009). Effect sizes for growth curve modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychological Methods, 14,* 43-53. doi:10.1037/a0014699

Finch, W. H. (2010). Imputation methods for missing categorical questionnaire data: A Comparison of Approaches. *Journal of Data Science, 8*, 361-378. doi:10.1.1.477.213

Fjermestad, K., McLeod, B. D., Tully, C. B., & Liber, J. M. (2015). Therapist characteristics and interventions: Enhancing alliance and involvement with youth. *The oxford handbook of treatment processes and outcomes in psychology.* New York, NY: Oxford Press.

Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement, 5,* 105-122. doi:10.1177/014662168100500115

Garland, A. F., Bickman, L., & Chorpita, B. F. (2010). Change what? Identifying quality improvement targets by investigating usual mental health care. *Administration and Policy in Mental Health and Mental Health Services Research*, *37*(1-2), 15–26. doi:10.1007/s10488-010-0279-y

Garland, A. F., Brookman-Frazee, L., Hurlburt, M. S., Accurso, E. C., Zoffness, R. J., Haine Schlagel, R., & Ganger, W. (2010). Mental health care for children with disruptive behavior problems: A view inside therapists' offices. *Psychiatric Services*, *61*, 788–795. doi:10.1176/appi.ps.61.8.788

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

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

Glasgow, R. E. (2013). What does it mean to be pragmatic? Pragmatic methods, measures, and models to facilitate research translation. *Health Education & Behavior*, *40*, 257–265. doi:10.1177/1090198113486805

Glasgow, R. E., Fernald, D. H., & Green, L. A. (2005). Practical and relevant self-report for primary care research. *Annals of Family Medicine*, 73–82. doi:10.1370/afm.261

Glasgow, R. E., & Riley, W. T. (2013). Pragmatic measures: What they are and why we need them. *American Journal of Preventive Medicine*, *45*(2), 237–243. doi:10.1016/j.amepre.2013.03.010

Godley, S. H., Garner, B. R., Smith, J. E., Meyers, R. J., & Godley, M. D. (2011). A large scale dissemination and implementation model for evidence-based treatment and continuing care. *Clinical Pscyhology: Science and Practice*, 1–17. doi:10.1111/j.1468-2850.2011.01236.x

Hagermoser Sanetti, L. M., Chafouleas, S. M., Christ, T. J., & Gritter, K. L. (2009). Extending use of direct behavior rating beyond student assessment: Applications to treatment integrity assessment within a multi-tiered model of school-based intervention delivery. *Assessment for Effective Intervention*, *34*(4), 251–258. doi:10.1177/1534508409332788

Henggeler, S. W., Chapman, J. E., Rowland, M. D., Halliday-Boykins, C. A., Randall, J., Shackelford, J., & Schoenwald, S. K. (2008). Statewide adoption and initial implementation of contingency management for substance-abusing adolescents. *Journal of Consulting and Clinical Psychology*, *76*(4), 556–567. doi:10.1037/0022-006X.76.4.556

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. doi:10.1016/j.jsat.2007.09.002

Hogue, A., Dauber, S., & Henderson, C. E. (2012). Therapist self-report of evidence-based practices in usual care for adolescent behavior problems: Factor and construct validity. *Administration and Policy in Mental Health and Mental Health Services Research*, *41*, 126–139. doi:10.1007/s10488-012-0442-8

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

Hogue, A., Ozechowski, T. J., Robbins, M. S., & Waldron, H. B. (2013). Making fidelity an intramural game: Localizing quality assurance procedures to promote sustainability of evidence-based practices in usual care. *Clinical Psychology: Science and Practice*, *20*, 60–77. doi:10.1111/cpsp.12023

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2,* 88–110. doi:10.1080/19345740802539325

Kahler, C. W., Metrik, J., LaChance, H. R., Ramsey, S. E., Abrams, D. B., Monti, P. M., … Brown, R. A. (2008). Addressing heavy drinking in smoking cessation treatment: A randomized clinical trial. *Journal of Consulting and Clinical Psychology*, *76*(5), 852–862. doi:10.1037/a0012717

Karver, M. S., Handelsman, J. B., Fields, S., & Bickman, L. (2006). Meta-analysis of therapeutic relationship variables in youth and family therapy: The evidence for different relationship variables in the child and adolescent treatment outcome literature. *Clinical Psychology Review,* 26, 50−65. doi:10.1016/j.cpr.2005.09.001

Kazdin, A. E. (2000). Developing a research agenda for child and adolescent psychotherapy. *Archives of General Psychiatry, 57*, 829-835. doi:10.1001/archpsyc.57.9.829

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

Kendall, P. C., & Hedtke, K. (2006). *Cognitive-behavioral therapy for anxious children: Therapist manual* (3rd 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*(2), 282–297. doi:10.1037/0022-006X.76.2.282

Kendall, P. C., Kane, M., Howard, B., & Siqueland, L. (1990). *Cognitive–behavioral therapy for anxious children: Treatment manual.* Ardmore, PA: Workbook Publishing.

Kendall, P. C., & Ollendick, T. H. (2004). Setting the research and practice agenda for anxiety in children and adolescence: A topic comes of age. *Cognitive and Behavioral Practice*, *11*, 65−74. doi:10.1016/S1077-7229(04)80008-7

Lerner, M. D., McLeod, B. D., Mikami, A. Y. (2013). Preliminary evaluation of an observational measure of group cohesion for group psychotherapy. *Journal of Clinical Psychology, 69*, 191-208. doi: 10.1002/jclp.21933

Levita, L., Duhne, P. G. S., Girling, C., & Waller, G. (2016). Facets of clinician's anxiety and the delivery of cognitive-behavioral therapy. *Behaviour Research and Therapy, 77,* 157 –

161.

Lichstein, K. L., Riedel, B. W., & Grieve, R. (1994). Fair tests of clinical trials: A treatment implementation model. *Advances in Behaviour Research and Therapy, 16,* 1–29. doi:10.1016/0146-6402(94)90001-9

Litt, M. D., Kadden, R. M., & Kabela-Cormier, E. (2009). Individualized assessment and treatment program for alcohol dependence: Results of an initial study to train coping skills. *Addiction*, *104*(11), 1837–1848. doi:10.1111/j.1360-0443.2009.02693.x

Manuel, J. K., Hagedorn, H. J., & Finney, J. W. (2011). Implementing evidence-based psychosocial treatment in specialty substance use disorder care. *Psychology of Addictive Behaviors*, *25*(2), 225–237. doi:10.1037/a0022398

Marker, C. D., Comer, J. S., Abramova, V., & Kendall, P. C. (2013). The reciprocal relationship between alliance and symptom improvement across the treatment of childhood anxiety. *Journal of Clinical Child & Adolescent Psychology*, *42*(1), 22–33. doi:10.1080/15374416.2012.723261

Martino, S., Ball, S. A., Nich, C., Canning-Ball, M., Rounsaville, B. J., & Carroll, K. M. (2011). Teaching community program clinicians motivational interviewing using expert and train-the-trainer strategies. *Addiction*, *106*(2), 428–441. doi:10.1111/j.1360-0443.2010.03135.x

Martino, S., Ball, S., Nich, C., Frankforter, T. L., & Carroll, K. M. (2009). Correspondence of motivational enhancement treatment integrity ratings among therapists, supervisors, and observers. *Psychotherapy Research: Journal of the Society for Psychotherapy Research*, *19*, 181–193. doi:10.1080/10503300802688460

McLeod, B. D. (2011). Relation of the alliance with outcomes in youth psychotherapy; A meta-analysis. *Clinical Psychology Review, 31*, 603-616. doi:10.1016/j.cpr.2011.02.001

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.

McLeod, B. D., Smith, M. M., Southam-Gerow, M. A., Weisz, J. R., & Kendall, P. C. (2015). Measuring treatment differentiation for implementation research: The Therapy Process Observational Coding System for Child Psychotherapy Revised Strategies Scale. *Psychological Assessment*, *27*(1), 314–325. doi:10.1037/t30447-000

McLeod, B. D., Southam-Gerow, M. A., Rodriguez, A., Quinoy, A., Arnold, C., Kendall, P. C., & Weisz, J. R. (2016). Development and initial psychometrics for a therapist competence instrument for CBT for youth anxiety. *Journal of Clinical Child and 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 & Practice*, *20*, 14–32. doi:10.1111/cpsp.12020

McLeod, B. D., Southam-Gerow, M. A., & Weisz, J. R. (2009). Conceptual and methodological issues in treatment. *School Psychology Review*, *38*(4), 541–546.

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

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*(3), 436–443. doi:10.1080/15374411003691750

Miller, W. R., Sorensen, J. L., Selzer, J. A., & Brigham, G. S. (2006). Disseminating evidence-based practices in substance abuse treatment: A review with suggestions. *Journal of Substance Abuse Treatment*, *31*, 25–39. doi:10.1016/j.jsat.2006.03.005

Miller, W. R., Yahne, C. E., Moyers, T. B., Martinez, J., & Pirritano, M. (2004). A randomized trial of methods to help clinicians learn motivational interviewing. *Journal of Consulting and Clinical Psychology*, *72*(6), 1050–1062. doi:10.1037/0022-006X.72.6.1050

Moyers, T. B., Martin, T., Manuel, J. K., Hendrickson, S. M. L., & Miller, W. R. (2005). Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment*, *28*(1), 19–26. doi:10.1016/j.jsat.2004.11.001

Noell, G. H., Witt, J. C., Slider, N. J., Connell, J. E., Gatti, S. L., Williams, K. L., … Duhon, G. J. (2005). Implementation integrity following behavioral consultation in schools: A comparison of three follow-up strategies. *School Psychology Review*, *34*, 87–106.

Norcross, J. C., & Lambert, M. J. (2011). Psychotherapy relationships that work II. *Psychotherapy, 48*, 4–8. doi:10.1037/a0022180

Orimoto, T. E. (2012). Assessment of therapy practices in community treatment for children and adolescents. *Psychiatric Services*, *63*(4), 343. doi:10.1176/appi.ps.201100129

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

Pierson, H. M., Hayes, S. C., Gifford, E. V., Roget, N., Padilla, M., Bissett, R., … Fisher, G. (2007). An examination of the motivational interviewing treatment integrity code. *Journal of Substance Abuse Treatment*, *32*(1), 11–17. doi:10.1016/j.jsat.2006.07.001

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,* 65-78. doi:10.1007/s10488-010-0319-7

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), 119. doi:10.1186/1748-5908-7-119

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed.)*. Newbury Park, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling*. Chicago, IL: Scientific Software International.

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research: Methods and data analysis.* New York: McGraw-Hill.

Roth, A. D., Pilling, S., & Turner, J. (2010). Therapist training and supervision in clinical trials: Implications for clinical practice. *Behavioural and Cognitive Psychotherapy*, *38*(3), 291–302. doi:10.1017/S1352465810000068

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. doi:10.1007/s10488-010-0321-0

Sheidow, A. J., Donohue, B. C., Hill, H. H., Henggeler, S. W., & Ford, J. D. (2008). Development of an audiotape review system for supporting adherence to an evidence-based treatment. *Professional Psychology: Research and Practice*, *39*(5), 553–560. doi:10.1037/0735-7028.39.5.553

Shelef, K., & Diamond, G. M. (2008). Short form of the revised vanderbilt therapeutic alliance scale: Development, reliability, and validity. *Psychotherapy Research*, *18*, 433–443. doi:10.1080/1050330701810801

Shirk, S. R., & Saiz, C. C. (1992). Clinical, empirical, and developmental perspectives on the therapeutic relationship in child psychotherapy. *Development and Psychopathology*, *4*, 713–728. doi:10.1017/S0954579400004946

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428. doi: 0033-2909/79/8602-0420

Silverman, W. K., & Albano, A. M. (1996), *The anxiety disorders interview schedule for DSM-IV-child and parent versions*. San Antonio, TX: Graywind Publications.

Silverman, W. K., Pina, A. A., & Viswesvaran, C. (2008). Evidence-based psychosocial treatments for phobic and anxiety disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 37*(1), 105–130. doi:10.1080/15374410701817907

Silverman, W. K., Saavedra, L. M., & Pina, A. A. (2001). Test–retest reliability of anxiety symptoms and diagnoses with the Anxiety Disorders Interview Schedule for DSM-IV: Child and parent versions. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 937–944. doi:10.1097/00004583-200108000-00016

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Press.

Slepian, M. L., Bogart, K. R., & Ambady, N. (2014). Thin-slice judgments in the clinical context. *Annual Review of Clinical Psychology, 10*, 131-153. doi:10.1146/annurev-clinpsy-090413-123522

Smith, J. L., Carpenter, K. M., Amrhein, P. C., Brooks, A. C., Levin, D. L., Schreiber, E. A., … Nunes, E. V. (2012). Training substance abuse clinicians in motivational interviewing using live supervision via tele-conferencing. *Journal of Consulting and Clinical Psychology*, *80*(3), 450–464. doi:10.1037/a0028176.

Smith, J. D., Dishion, T. J., Shaw, D. S., & Wilson, M. N. (2013). Indirect effects of fidelity to the family check-up on changes in parenting and early childhood problem behaviors. *Journal of Consulting and Clinical Psychology, 81*, 962–974. doi:10.1037/a0033950

Smith, M. M., McLeod, B. D., Southam-Gerow, M. A., Jensen-Doss, A., Kendall, P. C., & Weisz, J. R. (in press). Does the delivery of CBT for youth anxiety differ across research and practice settings? *Behavior Therapy*.

Stallard, P., Myles, P., Branson, A. (2014). The Cognitive Behaviour Therapy Scale for Children and Young People (CBTS-CYP): Development and psychometric properties. *Behavioural and Cognitive Psychotherapy, 42,* 269-282. doi:10.1017/S135246581300115X

Southam-Gerow, M. A., & Mcleod, B. D. (2013). Advances in applying treatment integrity research for dissemination and implementation science: Introduction to special issue. *Clinical Psychology: Science and Practice*, *20*, 1–13. doi:10.1111/cpsp.12019

Southam-Gerow, M. A., McLeod, B. D., Arnold, C. A., Rodriguez, A., Cox, J. R., Reise, S. P., Bonifay, W. E., Weisz, J. R., & 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: 10.1034/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 and Adolescent Psychiatry*, *49*(10), 1043–1052. doi:10.1016/j.jaac.2010.06.009

Stein, M. D., Herman, D. S., & Anderson, B. J. (2009). A motivational intervention trial to reduce cocaine use. *Journal of Substance Abuse Treatment*, *36*(1), 118–125. doi:10.1016/j.jsat.2008.05.003

Van Selst, M., & Jolicoeur, P. (1994) A solution to the effect of sample size on outlier elimination. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 47*, 631-650. doi:10.1080/14640749408401131

Walkup, J. T., Albano, A. M., Piacentini, J., Birmaher, B., Compton, S. N., Sherrill, J. T., … Kendall, P. C. (2008). Cognitive behavioral therapy or a combination in childhood anxiety. *The New England Journal of Medicine*, *359*(26), 2753–2766. doi:http://dx.doi.org.proxy.library.vcu.edu/10.1056/NEJMoa0804633

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

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 78*(2), 200–211. doi: 10.1037/a0018912

Weck, F., Grikscheit, F., Höfling, V., & Stangier, U. (2014). Assessing treatment integrity in cognitive-behavioral therapy: Comparing session segments with entire sessions. *Behavior Therapy*, *45*(4), 541–552. doi:10.1016/j.beth.2014.03.003

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., … Mayberg, S. (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., 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

Weller, E. B., Weller, R. A., Rooney, M. T., & Fristad, M. A. (1999a). Children's Interview for Psychiatric Syndromes (ChIPS). Washington, DC: American Psychiatric Press.

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

Wood, J. J., Piacentini, J. C., Bergman, R. L., McCracken, J., & Barrios, V. (2002). Concurrent validity of the anxiety disorders section of the Anxiety Disorders Interview Schedule for DSM-IV: Child and parent Versions. *Journal of Clinical Child and Adolescent Psycholology, 31,* 335-342. doi:10.1207/153744202760082595

Wood, J. J., Piacentini, J. C., Southam-Gerow, M. A., Chu, B. C., & Sigman, M. (2006). Family cognitive behavioral therapy for child anxiety disorders. *American Academy of Child and Adolescent Psychiatry*, *45*, 314–321. doi:10.1097/01.chi.0000196425.88341.b0

Appendix

TREATMENT INTEGRITY – EFFICIENT SCALE FOR CBT FOR YOUTH ANXIETY (TIES-CBT-YA)

General information

1. Type of recording (circle one): Audio (0) Video (1) Delay (2)

2. Total time of recording assignment (5 or 15 minutes): _________ 3. Number of people in recording: _________

4. Who is on the recording:

__ [1] Therapist (0 = not present, 1 = present) __ [2] Client (0 = not present, 1 = present)

__ [3] Other (describe: _______________) __ [4] Other (describe: _______________)

__ [5] Other (describe: _______________) __ [6] Other (describe: _______________)

To Enter Other:

Female caregiver (1), Male caregiver (2), Sibling (3), Other adult fam member (4), Other child (5), Teacher (6), Confederate (7)

5. Quality of recording: Poor (0) Fair (1) Good (2) Excellent (3)

6. Time started coding: ____________ 7. Time finished coding: ____________

8. Other coders with you (circle one): 0 1 2 3 4 5

9. Were there any problems with this file (circle one)? No (0) Yes (1)

10. If there *was* a problem with the file, what type of problem was it?

__ Audio/visual issues (01) ___ File cut off before completion of session (02)

__ Major portion of tape was in Spanish (03) ___ Other issue (please describe) _______________ (04)

__ None (90)

11. Was the therapist out of session at all? No (0) Yes (1) 12. If so, how long was the therapist gone? ___min

Descriptors of people in recording. If audio, record the first 5 words spoken by the client. If video, provide visual descriptors:

ADHERENCE SCALE

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

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

Item							**Freq**	**Ext**	
1. Skill-building - Adherence									1
2. Exposure - Adherence									2

COMPETENCE SCALE

Using the grid provided below, please indicate COMPETENCE of any item observed for each **two and a half-minute** time segment. Use a "*" to indicate *an above average rating*, "x" to indicate *an average rating*, and "-" to indicate *a below average rating*. After watching the ENTIRE recording, use the 0-7 scale to assign a Competence rating (Comp) for all items that are present in at least ONE (1) time period.

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Not present		**Poor**		**Adequate**		**Very Good**	

Item							**Comp**	
3. Skill-building - Competence								3
4. Exposure - Competence								4

DIFFERENTIATION SCALE

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

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

Item							**Freq**	**Ext**	
5. General Psychodynamic Focus									5
6. General Family Focus									6
7. General Client-centered Focus									7

VANDERBILT THERAPEUTIC ALLIANCE SCALE—REVISED: SHORT FORM (VTAS-R-SF)

Instructions: Using the Likert scales provided below, please indicate **to what extent the client** engaged in the behavior described. Place the appropriate number from the Likert scale in the space provided next to each item.

0	1	2	3	4	5
No Evidence	Minimal Evidence	Between Minimal and Moderate Evidence	Moderate Evidence	Clear Evidence	Extensive Evidence

Was this a parent only session? (0) No 1 (Yes)

If not, how long was the child in the session: _______ min

________ 1. Indicate that she experiences the therapist as understanding and supporting her.

________ 2. Seem to identify with the therapist's method of working, so that he assumed part of the therapeutic task himself.

________ 3. Act in a mistrustful or defensive manner toward the therapist.

________ 4. Share a common viewpoint about the definition, possible causes, and potential alleviation of the adolescent's problems.

________ 5. Agree upon the goals and tasks for the session.

Vita

Meghan Smith was born on September 15, 1988 in Boston, Massachusetts. She graduated with high distinction from the University of Virginia in May 2010 with a Bachelor of Arts in Psychology. She then worked for a year as a research assistant at the Child and Family Development Lab at the University of Maryland. She earned her Master of Science degree in Clinical Psychology in 2013 from Virginia Commonwealth University in Richmond, Virginia. Meghan is currently completing a clinical internship at Virginia Treatment Center for Children and will graduate with her doctorate in August, 2017.