2013

Characterizing the Implementation of CBT for Youth Anxiety

Meghan Smith
Virginia Commonwealth University

Follow this and additional works at: https://scholarscompass.vcu.edu/etd

Part of the Psychology Commons

© The Author

Downloaded from
https://scholarscompass.vcu.edu/etd/3239

This Thesis is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.
CHARACTERIZING THE IMPLEMENTATION OF CBT FOR YOUTH ANXIETY

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

By: MEGHAN MARIE SMITH

Bachelor of Arts, University of Virginia, 2010

Thesis and Academic Advisor: Bryce D. McLeod, Ph.D.
Associate Professor, Department of Psychology

Virginia Commonwealth University
Richmond, Virginia
November, 2013
# Table of Contents

List of Tables........................................................................................................ iv

List of Figures........................................................................................................ v

Acknowledgements............................................................................................ vi

Abstract.............................................................................................................. vii

Introduction....................................................................................................... 1

Literature Review............................................................................................... 7
  Youth Mental Health Problems...................................................................... 7
  Evidence-Based Treatments for Youth........................................................... 7
  Mixed Findings of EBT’s Performance in Practice Settings......................... 8
  Reasons for the Science-Practice Performance Gap.................................. 10
  Dissemination and Implementation Research.............................................. 13
  Using Implementation Research to Examine the Science-Practice Performance Gap.......................................................... 17
  Using a Treatment Integrity Framework to Measure Implementation Integrity.... 18
  Conceptual Model of Therapeutic Change in Therapy............................... 20
  Evaluating Implementation of Cognitive Behavioral Therapy for Youth Anxiety.... 22
  Benchmarking Implementation Integrity...................................................... 28
  The Therapy Processes Observational Coding System - Revised Strategy Scale... 30
  Problem Statement...................................................................................... 33
  Hypotheses..................................................................................................... 34

Method............................................................................................................... 35
  Participants................................................................................................. 35
  Treatment Methods................................................................................... 41
  Diagnostic and Symptom Measures............................................................ 42
  Assessment Procedures............................................................................... 45
  Treatment Adherence and Differentiation Measure................................... 46
  Coding and Session Sampling Procedures................................................. 49
  Comparing the Method and Results of the Two Parent Studies.................. 50

Results............................................................................................................. 52
  Preliminary Analyses.................................................................................. 52
  Primary Analyses....................................................................................... 60

Discussion....................................................................................................... 99
  Was There a Difference in CBT Delivery in Research and Practice Settings?... 101
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Demographic and Descriptive Information by Parent Study</td>
<td>36</td>
</tr>
<tr>
<td>2.</td>
<td>Results of Parent Studies</td>
<td>52</td>
</tr>
<tr>
<td>3.</td>
<td>Comparison of Continuous Demographic Variables Across Conditions</td>
<td>56</td>
</tr>
<tr>
<td>4.</td>
<td>Comparison of Categorical Demographic Variables Across Conditions</td>
<td>58</td>
</tr>
<tr>
<td>5.</td>
<td>Results of Hierarchical Model of Change for the TPOCS-RS Family Subscale</td>
<td>76</td>
</tr>
<tr>
<td>6.</td>
<td>Results of Condition Contrasts in the Means Model of the TPOCS-RS Family Subscale</td>
<td>77</td>
</tr>
<tr>
<td>7.</td>
<td>Results of Hierarchical Model of Change for the TPOCS-RS Psychodynamic Subscale</td>
<td>78</td>
</tr>
<tr>
<td>8.</td>
<td>Results of Condition Contrasts in the Linear Growth Curve Model of the TPOCS-RS Psychodynamic Subscale</td>
<td>79</td>
</tr>
<tr>
<td>9.</td>
<td>Results of Hierarchical Model of Change for the TPOCS-RS Coping Cat Subscale</td>
<td>84</td>
</tr>
<tr>
<td>10.</td>
<td>Results of Condition Contrasts in the Linear Growth Curve Model of the TPOCS-RS Coping Cat Subscale</td>
<td>85</td>
</tr>
<tr>
<td>11.</td>
<td>Results of Hierarchical Model of Change for the TPOCS-RS Cognitive Subscale</td>
<td>93</td>
</tr>
<tr>
<td>12.</td>
<td>Results of Condition Contrasts in the Means Model of the TPOCS-RS Cognitive Subscale</td>
<td>94</td>
</tr>
<tr>
<td>13.</td>
<td>Results of Hierarchical Model of Change for the TPOCS-RS Behavioral Subscale</td>
<td>95</td>
</tr>
<tr>
<td>14.</td>
<td>Results of Condition Contrasts in the Linear Growth Curve Model of the TPOCS-RS Behavioral Subscale</td>
<td>96</td>
</tr>
<tr>
<td>15.</td>
<td>Results of Hierarchical Model of Change for the TPOCS-RS Client-Centered Subscale</td>
<td>97</td>
</tr>
<tr>
<td>16.</td>
<td>Results of Condition Contrasts in the Means Model of the TPOCS-RS Client-centered Subscale</td>
<td>98</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Conceptual Model of Therapeutic Change in Therapy</td>
<td>21</td>
</tr>
<tr>
<td>2.</td>
<td>TPOCS-RS Family Extensiveness Scores by Condition</td>
<td>73</td>
</tr>
<tr>
<td>3.</td>
<td>TPOCS-RS Psychodynamic Extensiveness Scores by Condition</td>
<td>75</td>
</tr>
<tr>
<td>4.</td>
<td>TPOCS-RS Coping Cat Extensiveness Scores by Condition</td>
<td>83</td>
</tr>
<tr>
<td>5.</td>
<td>TPOCS-RS Cognitive Extensiveness Scores by Condition</td>
<td>88</td>
</tr>
<tr>
<td>6.</td>
<td>TPOCS-RS Behavioral Extensiveness Scores by Condition</td>
<td>89</td>
</tr>
<tr>
<td>7.</td>
<td>TPOCS-RS Client-centered Extensiveness Scores by Condition</td>
<td>90</td>
</tr>
</tbody>
</table>
Acknowledgements

I would like to thank Dr. McLeod for working with me to develop and polish this thesis. I would also like to thank my parents, lab members, and graduate student colleagues for their unwavering support.
Abstract
CHARACTERIZING THE IMPLEMENTATION OF CBT FOR YOUTH ANXIETY

By Meghan M. Smith, B.A.

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

Virginia Commonwealth University, 2013

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

Although evidence-based treatments (EBTs) often perform well in research settings, when EBTs are delivered in practice settings they sometimes fail to outperform usual care (UC). One reason for this could be that therapists in practice settings may not follow the EBT protocol as closely or may deliver more therapeutic interventions that align with other treatment domains. I tested this possibility in the context of cognitive-behavioral treatment (CBT) for youth anxiety by comparing how CBT was delivered in practice (i.e., community-based clinics) and research (i.e., lab) settings. A diverse sample of youth (aged 7-15) received one of two treatments to address primary presenting problems of anxiety: CBT in a lab (N = 51), CBT in clinics (N = 18), or UC in clinics (N = 22). The extensiveness of therapeutic interventions delivered was measured using the Therapy Process Observational Coding System for Child Psychotherapy – Strategies Scale (McLeod & Weisz, 2010). Two coders independently coded the same available sessions. Analyses investigated mean-level differences in interventions session-by-session, over time between the three groups. Results indicated that CBT delivery differed across groups, over time. Future researchers should use this characterization to further implementation and process-outcome research, and eventually either enhance quality control efforts or alter aspects of EBT protocols (Garland, Hulbert, & Hawley, 2006).
Introduction

Youth mental health problems are alarmingly common and costly, with prevalence estimates in youth as high as 20 percent, costing the United States almost 10 billion dollars in 2006 (Soni, 2006; U.S. Department of Health & Human Services, 2011). However, many youth with significant mental health problems do not receive psychosocial treatment or mental health treatment that is considered adequate (Fulda, Lykens, Bae, & Singh, 2009; Kataoka, Zhang, & Wells, 2002). More than 90 percent of youths who receive public psychosocial services are in programs that have no evidence supporting their impact (Bickman, 2008). Thus, there has been a push from mental health researchers and practitioners to resolve this glaring and critical need.

Over the past 25 years, mental health researchers have taken actions to provide youth the psychosocial treatment they require, by developing evidenced-based treatments (EBTs) and evaluating them in efficacy studies (i.e., controlled randomized clinical trials conducted in research settings; Weisz, Jensen-Doss, & Hawley, 2006). A wealth of efficacy studies indicate that there are a number of EBTs that significantly reduce youth symptomology from pre- to post-treatment (Chorpita & Dalieden, 2009; McLeod & Weisz, 2004; Weersing & Weisz, 2002). These studies suggest that EBTs can be beneficial to youth, and offer hope to the millions of youth suffering from mental health problems.

However, though EBTs have been found to perform well in efficacy trials, EBTs have not always performed as expected once they have been transported to practice settings (e.g., Southam-Gerow et al., 2010; Weisz et al., 2006). When delivered in effectiveness studies (i.e., naturalistic tests in practice settings), EBTs sometimes fail to outperform treatment as usual, or usual care (UC; Southam-Gerow et al., 2010). Thus, although the highly structured lab setting
where EBT development takes place may help researchers refine the EBT, it also may make transportation of the EBT more challenging. In particular, different therapist, client, and contextual characteristics in the two settings could hinder EBT transportation and adoption, but scant research has been conducted to begin to understand this issue and conclusions have been formed mainly on speculation (Bickman, 2000; Garland, Bickman, & Chorpita, 2010; Weisz et al., 2006).

Dissemination and implementation (D&I) research emerged in part to investigate how characteristics of practice settings may influence EBT performance. This field of research emphasizes conducting studies that identify the factors involved in moving an EBT from a research setting to a practice setting (Chambers, Ringeisen, & Hickman, 2005; Schoenwald & Hoagwood, 2001). D&I researchers ask: given the current problems with effectiveness studies, how can we facilitate future research in transporting and implementing EBTs in practice settings?

We cannot know “why” outcomes of EBTs sometimes differ in the two settings until we figure out “how,” or in what ways the therapy sessions differ (Doss, 2004; Kazdin, 2000). Before we can foster future research in transporting and implementing EBTs in practice settings, and potentially improve mental health care in practice settings, we need to understand what is actually occurring in the therapy sessions in these two contexts (Garland et al., 2010). In particular, researchers need to better understand the extent that elements of an EBT are delivered according to the treatment model in practice settings (i.e., implementation integrity; McLeod, Southam-Gerow, Bair, Rodriguez, & Smith, 2013). For instance, researchers have questioned whether one reason EBTs do not consistently outperform UC is because the EBTs are not delivered as designed in practice settings. Treatment integrity measures (i.e., measures that
assess the extent that the treatment was delivered as intended) give D&I researchers a framework to use to explore these types of questions; treatment integrity measures characterize implementation integrity in an organized, consistent way.

One particularly important aspect of implementation integrity (McLeod et al., 2013), and thus a central focus of D&I research, is the therapist’s delivery of therapeutic interventions. Given that pre-treatment dissimilarities in research and practice settings exist (e.g., therapist characteristics; Garland et al., 2010), variations in the way therapists deliver therapeutic interventions could exist as well (McLeod et al., 2013). For instance, therapists in practice settings may add or omit therapeutic interventions that are prescribed by (i.e., are included in) the treatment protocol, and these alterations could impact EBT implementation. However, I was not aware of any study that both (a) investigated if EBT delivery differences exist, and (b) delineated what the specific EBT delivery differences were, if they did exist.

The current study therefore aimed to help researchers begin to understand therapeutic intervention delivery differences across settings, by characterizing implementation integrity. Specifically, the current study examined the implementation of cognitive behavioral therapy (CBT) for youth anxiety, a well-regarded EBT (Silverman, Pina, & Viswesvaran, 2008). This mental health treatment was ideal to study because a number of studies demonstrate its efficacy but it sometimes performs no better than UC when implemented in practice settings (Barrington, Prior, Richardson, & Allen, 2005; Southam-Gerow et al., 2010). Some researchers assert that this science-practice performance discrepancy may occur due to differences in the delivery of therapeutic interventions (McLeod et al., 2013). For example, a past effectiveness study suggested that therapists who were trained to deliver CBT in a practice setting may have delivered low doses of CBT (Southam-Gerow et al., 2010). However, these researchers could not
draw conclusions from this finding since the study had no “gold standard” extensiveness level to compare it with. Since there was not an adequate comparator to help judge whether CBT implementation might have been high/low, the findings could not be used by future researchers to initiate quality control efforts to improve aspects of CBT implementation, if/where necessary.

The current study aimed to overcome this shortcoming through a process called benchmarking. In business, benchmarking uses a “gold standard” comparison point to assess an organization’s performance or progress in a particular domain (Lai, Huang, & Wang, 2011). The current study compared CBT implementation in a practice setting to a “gold standard” benchmark – that is, CBT delivery in a research setting. It is important to note that although therapists delivering CBT in research settings may not produce “gold standard” client outcomes, they deliver interventions in the “gold standard” way, or the way the interventions were originally designed. The current study therefore used benchmarking to compare the extent that interventions were delivered as designed in research vs. practice-settings, without suggesting that “gold-standard” delivery necessarily is positive or negative for client outcomes. The Therapy Process Observational Coding System for Child Psychotherapy – Revised Strategies Scale (TPOCS-RS; McLeod & Weisz, 2010) was used to characterize implementation integrity and delivery. The TPOCS-RS was a well-suited measure to use to compare treatment delivery in the two settings because it includes a wide range of therapeutic interventions that map onto the treatment domains commonly found in youth mental health treatment (McLeod & Weisz, 2010). Thus, it was ideal for measuring the delivery of interventions that fell either within or outside the CBT framework. Moreover, it is observational and includes a rating scale that assesses for the frequency and depth with which the interventions were delivered; therefore, the measure is relatively objective and thorough, respectively (DiMatteo, 2004; Hill, 1991; Mowbray, Holter,
Teague, & Bybee, 2003; Noell et al., 2005). This tool allowed this study to identify the type of therapeutic interventions that therapists delivered in each setting, the extent that they delivered these interventions, and the degree to which these interventions fit or did not fit the CBT framework.

This study aimed to offer future researchers a comparative characterization (i.e., benchmarking) of implementation integrity to give D&I researchers insight into the areas of CBT implementation that met gold standard expectations and areas of CBT implementation that may have altered adoption. For example, if therapists in practice settings delivered more therapeutic interventions that were proscribed by the EBT program (i.e., interventions not included in the EBT protocols) than therapists in research settings and these interventions predicted poorer client outcomes, researchers could consider modifying therapist training sessions in practice settings (Frank, Kupfer, Wagner, & McEachran, 1991; Schoenwald et al., 2011). Conversely, if the delivery of more proscribed interventions in practice settings predicted better client outcomes, researchers could consider altering the EBT protocol, to incorporate the effective proscribed interventions (Garland et al., 2006). The current study aimed to describe how implementation integrity in practice settings compared to treatment delivery in research settings so that future researchers could either adjust how the EBT should be implemented in practice settings or adjust the EBT protocol itself to align with practice-based evidence.

In sum, the current study characterized the implementation of CBT. An observational measure designed to assess therapeutic interventions from a wide range of therapeutic domains, the TPOCS-RS, was used to benchmark the type and extensiveness of therapeutic interventions delivered. Then, implementation integrity over time was assessed, depicting the change in intervention delivery in all conditions as treatment progressed.
First, I will review the literature surrounding youth mental health problems, EBTs, D&I research, and treatment integrity research. I will present a model that overcomes limitations of implementation integrity studies in the past and optimizes the implementation integrity analysis proposed in the current study. Next, I will review the method, hypotheses, and analyses of the current study, which focused on the means and trajectories of the therapeutic intervention extensiveness levels delivered in regards to how they fell within and outside the CBT framework, across conditions. Finally, I will discuss the implications of the current study’s characterization of CBT implementation, including recommendations of how investigators could incorporate this information in future research and clinical endeavors.
Literature Review

Youth Mental Health Problems

Mental health problems in youth are prevalent, incapacitating, and oftentimes left untreated (Insel & Fenton, 2005; Kataoka et al., 2002; Mash & Barkley, 2003). Research indicates that around ten percent of youth aged 5-16 years old experience a mental health problem (e.g., generalized anxiety disorder, major depressive disorder, attention-deficit/hyperactivity disorder) at any one period (Mental Health Foundation, 2007). Half of all lifetime cases of mental health problems begin by age 14, but there are often long periods between the first onset of symptoms and when people seek psychosocial treatment (National Institute of Mental Health [NIMH], 2005). This suggests that youth suffering from mental health problems may not be getting the psychosocial treatment they require (NIMH, 2005). Moreover, the majority of youth that do seek help for mental health problems receive psychosocial treatments that are not empirically supported (Bickman, 2008). Taking steps to ensure that the youth who seek mental health services in practice settings receive effective psychosocial treatments is therefore a national public health priority (US Department of Human Health and Services, 2011).

Evidence-Based Treatments for Youth

Fortunately, many efficacious psychosocial treatments (i.e., treatments that have produced significant pre- to post-treatment symptom reduction in controlled research settings) exist for a wide range of youth mental health problems (Chorpita & Southam-Gerow, 2006; Jacobs et al., 2010; Nathan & Gorman, 2002; Piacentini et al., 2010; Webster-Stratton, Reid, & Beauchaine, 2011). At the beginning of the millennium, about 1,500 controlled research studies
had been conducted investigating the efficacy of psychosocial treatments for a variety of youth mental health problems (Kazdin, 2000), and many more have been conducted since then. In these controlled treatment settings, the magnitude of the treatment effect approaches a size that has been considered large by some researchers (d-type effect sizes around .70; Cohen, 1988); in percentile terms, this means that the average treated youth is doing better after treatment than approximately 76% of the control group youth, averaging across outcome measures (Weisz, Jensen, & McLeod, 2005). These efficacy trials therefore demonstrate that EBTs have the potential to put youth and their families on a more positive developmental course.

**Mixed Findings of EBT’s Performance in Practice Settings**

Although EBTs have generally produced beneficial effects in efficacy trials, it remains unclear how EBTs compare to UC, in both research and practice settings. Past research has suggested that UC has a marginal impact on youths’ symptom severity and functioning status (Bickman, Noser, & Summerfelt, 1999; Weiss, Catron, Harris, & Phung, 1999). Thus, in an effort to improve UC, researchers have attempted to transport efficacious EBTs to practice settings. However, in some studies, EBTs delivered in practice settings have produced worse or approximately the same youth symptom and functioning outcomes as UC; findings from these studies raise the question of whether EBTs can consistently produce better client outcomes than UC (e.g., Clarke et al., 2005; Southam-Gerow et al., 2010; Weisz et al., 2006).

A recent meta-analysis of randomized controlled studies comparing EBTs with UC in practice settings reflects the current status of effectiveness research in youth psychosocial treatment (Weisz et al., 2006). Weisz and colleagues (2006) attempted to answer the question of whether EBTs produce better outcomes than UC by synthesizing the findings of 32 studies that
directly compared EBTs delivered in practice settings with UC. Overall, the authors found that EBTs outperformed UC. However, there was marked study-to-study variability. Although most studies found that EBTs outperformed UC, in a subset of studies, EBTs did not outperform UC (Weisz et al., 2006).

A closer look at some of the studies in the meta-analysis helps illustrate that EBTs delivered in practice settings have performed both significantly better and worse than UC. In the studies conducted by Borduin, Henggeler, Blaske, and Stein (1990) and Kazdin, Esvedt-Dawson, French, and Unis (1987), the effect sizes were high (0.83 and 1.12, respectively), meaning that the average youth treated with an EBT was better off after treatment than around 80% and 85% of youths who received UC, respectively. However, Guarion (1992) and Hawkins, Jensen, Catalono, and Wells (1991) found negative effect sizes (-0.86 and -0.51, respectively), meaning that the average youths who received UC were actually better off than around 80% and 69% of youths who received an EBT, respectively. Clearly, EBTs delivered in practice settings do not always outperform UC.

Together, these findings demonstrate that although EBTs have produced significant positive outcomes in research settings, they have not consistently outperformed UC, as researchers expected. Before we take further implementation actions, we need to better understand why EBTs do not consistently outperform UC (Hoagwood & Olin, 2002). A few researchers have posited that EBTs sometimes do not outperform UC because implementation may differ between practice and research settings as a result of contextual discrepancies between the settings.
**Reasons for the Science-Practice Performance Gap**

Understanding the differences in research vs. practice settings is integral to EBT implementation research. In fact, the contextual differences between the two settings may, in part, explain the existence of the science-practice performance gap. Researchers have found that research and practice settings have different therapist, client, and contextual characteristics (Bickman, 2008; Southam-Gerow, Weisz, & Kendall, 2003); they speculate that these differences may influence EBT implementation (e.g., therapists in practice settings adhere less to the EBT protocol; Garland et al., 2010; McLeod et al., 2013) and/or outcomes (e.g., less symptom reduction in practice settings; Bickman, 2008; McLeod et al., 2013). Therefore, it is important to recognize these contextual differences and investigate if they influence EBT implementation and outcomes.

One major discrepancy between research and practice settings is therapist characteristics. Therapists in research settings have smaller, more specific caseloads, with extensive supervision and experience in certain treatment domains whereas therapists in practice settings have more cases, limited supervision, and more exposure to a diverse array of psychosocial treatment domains (Garland, Hurlburt, Brookman-Frazee, Taylor, & Accurso, 2010; Southam-Gerow et al., 2003). Therapists in practice settings also tend to endorse a broader range of theoretical orientations than therapists in research settings (Garland et al., 2010; Southam-Gerow et al., 2003). In general, therapists in practice settings tend to have more diverse backgrounds and expertise in more treatment domains than therapists in research settings (Southam-Gerow et al., 2003). These differences could lead to dissimilarities in how therapists in practice settings implement the EBT and subsequently, differences in client outcomes (McLeod et al., 2013). For example, therapists in practice settings could deliver more therapeutic interventions that adhere
to other treatment domains because they have had more training or experience delivering interventions from other treatment domains, and thus feel more comfortable using them (McLeod, 2009). Therapists in practice settings may also differ in regards to their attitudes toward the use of a protocol (Garland et al., 2010). For example, therapists in practice settings, who have had more diverse training and exposure to theoretical orientations, may be less likely than therapists in research settings to endorse the use of a protocol in therapy. Thus, discrepant therapist characteristics could bring about implementation and outcome differences.

Client characteristics in the two settings differ as well and could contribute to outcome discrepancies. Clients in practice settings have more severe and comorbid cases than clients in research settings, and have been found to significantly differ from clients in research settings in terms of ethnicity, parental psychopathology, and significant life events (Hammen, Rudolph, Weisz, Rao, & Burge, 1999). Clients in practice settings have been found to exhibit higher levels of externalizing problems and are disadvantaged in terms of important demographic characteristics; specifically, youth in practice settings are poorer and more likely to have only one caretaking parent than youth in research settings (Southam-Gerow et al., 2003). Youth clients in practice settings are also more likely to receive other psychosocial services (Garland et al., 2010). Moreover, efficacy trials oftentimes exclude youth participants based on low IQ or medication use, contributing to the differences between samples in the two settings. Thus, the more complex cases seen in practice settings could prompt practice setting therapists to diverge from the EBT protocols more frequently than research setting therapists (McLeod et al., 2013), leading to different client outcomes than those seen in research settings.

Lastly, there are important contextual differences between the two settings. Whereas efficacy trials are engendered to support the study’s research goals (e.g., employees are devoted
to accomplishing goals), practice settings are designed to provide services (e.g., employees are devoted to providing a certain number of services to cover expenses; Southam-Gerow et al., 2003). The disparate funding sources play a role in how resources are allocated or supplied; funding of research in practice settings is much less than funding of research in other areas of health care (Southam-Gerow et al., 2003; Westfall, Mold, & Fagnan, 2007). In addition, providers may not fully accept EBTs and argue that studies with statistically significant findings are irrelevant to practice settings or could contribute to misuse by managed care (Kettlewell, 2004). Discrepant provider attitudes, in turn, may influence other systems such as how therapists communicate with clinic staff members, and the subsequent extent that therapists feel comfortable delivering core EBT interventions. For example, if some staff members in practice settings do not endorse the use of the protocol, the practice setting therapists may feel less comfortable requesting certain materials or asking staff members to help with core therapeutic tasks. Thus, there is a mismatch in organizational and systemic priorities, decision-making, and norms that could influence EBT implementation and client outcomes (Southam-Gerow et al., 2003; Southam-Gerow, Marder, & Austin, 2008).

In sum, there are therapist, client, and contextual differences between research and practice settings that may influence implementation integrity and outcomes. In practice settings, therapists have more diverse training and expertise, families seeking help present with more complex cases, and organizations do not always have sufficient funding and employee support (Southam-Gerow et al., 2003). Thus, therapists in practice settings may feel the need to implement the EBT differently. For example, therapists in practice settings may stray more frequently from the EBT protocol due to their diverse training background or in order to appropriately accommodate practice-based challenges, in turn influencing the EBT’s
performance either negatively or positively (Garland et al., 2006; Weisz et al., 2006). However, little is known about exactly how the contextual differences influence aspects of EBT implementation, and how different aspects of EBT implementation, in turn, may influence client outcomes for better or worse.

**Dissemination and Implementation Research**

Over the last decade, dissemination & implementation (D&I) research emerged to investigate how contextual factors in the two different settings influence implementation and outcomes (Chorpita et al., 2002; Southam-Gerow et al., 2008). Previous efficacy/effectiveness research was not designed to answer “why” there are mixed EBT performance findings because this research only measured the “before” and “after” components of a psychosocial treatment study. As a result, how these contextual factors did, or did not, influence treatment outcomes, were speculative (Garland et al., 2010). D&I research attempts to overcome this shortcoming by investigating EBT *implementation* thereby making the *processes* at play “during” the study the outcomes of interest (McLeod et al., 2013; Southam-Gerow et al., 2011).

D&I researchers conduct research that falls into two broad categories. The first category is implementation research, research that investigates how factors in practice settings influence EBT effectiveness and adoption (Schoenwald & Hoagwood, 2001; Southam-Gerow et al., 2008). For example, some D&I researchers assess how organizations interact with one another and the extent that inter-organizational interactions influence EBT adoption; if one organization delivers an EBT and one organization does not, the exchange of information about EBT implementation or encouragement by the other organization’s providers to implement aspects of the EBT could influence both organizations in terms of their training, supervision, and EBT delivery (Aarons,
Hulbert, & Horwitz, 2011; Sobo, Bowman, Aarons, Asch, & Gifford, 2008). Other D&I researchers examine how organizational culture (i.e., norms) and climate (i.e., individual opinions of the workplace) influence EBT implementation. For example, organizational norms (e.g., whether everyone uses a protocol), mechanisms in place to spread knowledge throughout an organization (e.g., meetings, email, web-sites), and employee opinions of the decision-making structure or people in positions of power (e.g., attitudes toward protocol use, whether employees agree with the leader’s decision to use an EBT protocol) may influence EBT adoption (Aarons et al., 2011; Glisson & Schoenwald, 2005). D&I researchers even examine the extent that in-session therapy processes influence the implementation of an EBT (e.g., delivery of certain therapeutic interventions, alliance; McLeod et al., 2013). In general, implementation researchers have found that environments and therapists that do not have necessary pre-existing requirements/skills (e.g., therapists who lack active listening skills) and are not open to change may struggle to use evidence-based innovations (Aarons et al., 2011; Glisson & Schoenwald, 2005). Implementation studies thus investigate how a broad range of factors influence an EBT’s effectiveness and adoption.

Importantly, findings from implementation studies can lead to modifications in EBT protocol, implementation procedures, or both. For example, if data suggested that organizational culture negatively influenced the use of evidence-based innovations (a process-level outcome), researchers could either adjust implementation procedures to accommodate the organizational climate, or attempt to adjust the organizational climate to accommodate the evidenced-based procedures (Aaron et al., 2011). For example, some organizations may strictly prohibit off-site sessions. In one EBT (i.e., CBT for youth anxiety), the protocol instructs therapists to expose clients to their unique fears, oftentimes implying that the therapist hold a therapy session outside
of the clinic. In these types of situations, implementation researchers could consider including examples of exposure substitutions in the protocol (e.g., use of virtual exposure techniques; i.e., an implementation procedure) or negotiating with clinic staff to adjust off-site session policies (i.e., an organizational factor). The ultimate goal of D&I research is to modify the EBT and/or develop implementation procedures (i.e., instructions that outline the appropriate training and supervision practices of the EBT) so that optimal performance is achieved (evinced by multiple levels of outcomes) in diverse practice settings and the EBT can consequently be distributed on a large-scale level, a process referred to as dissemination (Southam-Gerow et al., 2008). However, D&I research emphasizes extensive testing of the EBT in the implementation study phase, since this phase could alter aspects of the EBT or implementation procedures in crucial ways and has been largely overlooked in psychosocial treatment research (Schoenwald et al., 2011).

The second category of D&I research, dissemination research, investigates how an EBT should be marketed so it can be distributed on a large-scale level. This research takes place only after the implementation procedures have been successfully defined and validated in the implementation stage (Southam-Gerow et al., 2008). The dissemination stage consists of a test of a dissemination intervention: a set of procedures that outline how both the psychosocial treatment and implementation procedures should be presented to organizations, providers, and stakeholders to best foster the EBT’s adoption (Southam-Gerow et al., 2008). For example, some researchers test how using different mediums of communication to present the EBT and implementation procedures influences EBT adoption (Rogers, 2003); mass media channels may be best suited to increase information and awareness within a system, whereas interpersonal channels may be best suited to persuade individuals to accept the new EBT and implementation procedures, because of its bidirectional nature (Southam-Gerow et al., 2008). Dissemination
research also tests how elements of the social system (e.g., norms, decision-making structure) influence EBT adoption. Since an opinion leader (i.e., an individual who has opinions that are respected in an organization) often dictates the social norms (i.e., typical attitudes about how others should behave) in a system, the way that an EBT is presented to the opinion leader (e.g., if the idea is consistent with the system’s values; the EBT representative seems similar to the opinion leader) could be critical to the subsequent adoption or rejection of the EBT (Southam-Gerow et al., 2008). Optimal dissemination may entail presenting the psychosocial treatment through multiple mediums (e.g., mass media, interpersonal), in a way that appears consistent with the system’s norms and the community’s needs, and emphasizes improvement upon the interventions currently in place in the system (Southam-Gerow et al., 2008). Adequate marketing of an EBT and its implementation procedures is essential for spreading the EBT throughout the community; this dissemination phase is the last step in transporting an efficacious EBT from research to practice settings.

Overall, D&I research overcomes past efficacy/effectiveness studies because it assesses the processes at work in EBT implementation, rather than only pre and post-treatment factors. Both categories of D&I research contribute uniquely to the knowledge base. First, implementation studies investigate how factors influence EBT implementation and outcomes, and those findings are used to inform and modify EBT implementation procedures. Then, dissemination studies investigate how best to market those validated EBT implementation procedures, so that the EBT can be widely distributed to diverse practice settings. However, given that the field has not sufficiently identified how certain factors influence EBT implementation and outcomes, implementation research remains the current focus of many D&I investigators.
Using Implementation Research to Examine the Science-Practice Performance Gap

Implementation research holds promise for helping investigators identify reasons for the science-practice performance gap. Unlike previous psychosocial treatment research (i.e., effectiveness studies) that evaluated an EBT’s success by comparing client-level outcomes of EBTs to those of UC in practice settings, implementation research evaluates an EBT’s success by comparing multiple levels of outcomes (e.g., client and therapist levels; system and organizational levels; therapy process levels) in both settings (Southam-Gerow et al., 2008). Implementation research attempts to identify how the contextual differences in the two settings influence the processes involved in adopting and adequately executing EBTs, something previous psychosocial treatment studies rarely assessed, or evaluated only for manipulation check purposes (Southam-Gerow et al., 2008). Identifying implementation (process-level) differences in the two settings is the first step, and therefore the key to answering “why” EBT performance in research and practice settings sometimes differs (Doss, 2004; Kazdin, 2000).

One critical focus of implementation research is the extent to which the elements of an EBT are delivered according to the treatment model in practice settings, also referred to as implementation integrity (McLeod et al., 2013). D&I researchers have begun to focus upon implementation integrity because contextual differences in research and practice settings may lead to differences in how the therapeutic interventions contained in EBTs are delivered in those settings. Indeed, type of treatment, rendition of protocol, and setting characteristics have all been found to influence the degree to which therapists deliver certain therapeutic interventions (Barber, Foltz, Crits-Christoph, & Chittams, 2004; Garland et al., 2010; Hill, O’Grady, & Elkin, 1992; Malik, Beutler, Alimohamed, Gallagher-Thompson, & Thompson, 2003). For example, Malik et al. (2003) found that therapists using three different renditions of a cognitive therapy
(CT) protocol (those formatted for individuals, couples, and groups) significantly differed in their delivery of core CT interventions; interventions were more heavily focused on behavior or insight, depending on the condition. However, researchers still know little about the nature of implementation integrity because it is dynamic and requires extensive resources for measurement (Garland et al., 2006).

It could be that EBTs are not consistently outperforming UC in practice settings because EBTs are not implemented as designed in practice settings (i.e., delivery of the EBT does not match the treatment protocol). For example, therapists in practice settings may diverge from the protocol more than therapists in research settings because they have been trained in other treatment domains and feel more comfortable delivering interventions from those domains (Garland et al., 2010), they deliver interventions that appropriately fulfill their clients’ unique needs (Garland et al., 2006), or they are not provided sufficient training in the EBT (Weisz et al., 2009). It is therefore important for researchers to examine implementation integrity so they can conclude whether it is the EBT (i.e., elements or structure of the EBT protocol) or implementation integrity (i.e., the delivery of the EBT elements) that may be contributing to the science-practice performance discrepancy (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005; McLeod & Southam-Gerow, 2013).

**Using a Treatment Integrity Framework to Measure Implementation Integrity**

Implementation integrity is important to assess, but researchers need a way to operationalize it (Schoenwald et al., 2011). With some adjustments, treatment integrity (i.e., an assessment of the extent to which a treatment was delivered as intended) provides a framework that D&I researchers can use to operationalize implementation integrity (McLeod et al., 2013). Specifically, treatment integrity breaks the delivering of EBTs into four components that
together, and in isolation, explain how EBTs produce therapeutic change: the relational component, competence, treatment adherence, and treatment differentiation (McLeod, Southam-Gerow, & Weisz, 2009; Perepletchikova & Kazdin, 2005). Each of these treatment integrity components has a unique, pivotal role in successful psychosocial treatment (McLeod et al., 2013).

The “relational” component of treatment integrity consists of the youth-therapist alliance and client involvement (McLeod et al., 2009). The youth-therapist alliance refers to affective aspects of the youth-therapist relationship such as their affective bond, participation in tasks, and agreement on therapy goals (McLeod & Weisz, 2005; Shirk & Saiz, 1992). Client involvement, on the other hand, refers to the extent that the youths participate in therapeutic activities, self-disclose, initiate the discussion of challenging topics, and psychologically engage in the therapeutic tasks (Braswell, Kendall, Braith, Carey, & Vye, 1985; Chu & Kendall, 2004; Eugster & Wampold, 1996). These factors capture aspects of treatment delivery that help the youth engage in mental health treatment, so he or she can be more receptive to the therapeutic interventions.

The three “technical” components of treatment integrity focus on either the therapist’s skill (competence) or the therapeutic interventions that therapists deliver (treatment adherence and treatment differentiation; McLeod et al., 2009). Competence refers to the therapist’s degree of skill and responsiveness when delivering therapeutic interventions (Perepletchikova, Treat, & Kazdin, 2007). Treatment adherence refers to the extent to which the therapist’s delivery of therapeutic interventions corresponds to the treatment’s specified protocol (Perepletchikova & Kazdin, 2005). Lastly, treatment differentiation refers to the extent to which the treatments differ from one another in important categories (Perepletchikova et al., 2007). Altogether, these factors
of treatment integrity constitute the implementation integrity information outlined in the
treatment protocols.

Treatment integrity enables D&I researchers to operationalize implementation by
separating it into four distinct components. The “relational component” of treatment integrity
makes up the aspects of implementation that in part allows the technical elements to work;
therapists can follow the EBT protocol, but they will only produce positive outcomes if the youth
is engaged in treatment. The three “technical components” constitute the aspects of
implementation that are explicitly delineated in the EBT protocols; these components are
necessary to give the sessions structure and depth. Combined, these components of treatment
integrity have a substantial position in the overall process of therapeutic change (McLeod, Islam,
& Wheat, 2013).

**Conceptual Model of Therapeutic Change in Therapy**

Figure 1 demonstrates where treatment integrity components reside in a model depicting
how elements of psychosocial treatment are thought to produce change (McLeod, Islam, &
Wheat, 2013). Pre-treatment factors such as therapist, client, parent, and setting characteristics
(labeled “therapy inputs” in the model) impact *relationship factors, therapeutic interventions,*
*and therapist competence* (all labeled “treatment delivery” in the model). These treatment
delivery components, conceptualized as in-session implementation integrity components in the
current study, in turn influence client cognitions, behavior skills, or parenting (labeled “change
mechanisms” in the model; Doss, 2004; Kazdin, 2000). Finally, these change mechanisms lead
to changes in post-treatment factors such as symptoms, functioning, client perspectives,
environments, and systems (labeled “outcomes” in the model).
Overall, Figure 1 illustrates how implementation integrity could play an important role in the association between different therapy inputs and outcomes in implementation studies. Different therapy inputs in research and practice settings (e.g., therapists with diverse training backgrounds in practice settings) could lead to differences in implementation integrity (e.g., less treatment adherence in practice settings). These implementation integrity differences could then lead to differences in change mechanisms (e.g., less/more change in client cognitions in practice settings). Given that change mechanisms are the medium through which psychosocial treatment produces change (Doss, 2004; Kazdin, 2000), altered change mechanisms could ultimately influence the outcomes of psychosocial treatment (e.g., less/more symptom reduction from EBTs implemented in practice settings). Treatment integrity, as seen in Figure 1, is the framework D&I researchers can use to characterize implementation integrity. That is, treatment integrity shows D&I researchers specific areas where implementation integrity is carried out as expected and specific areas where implementation integrity is not executed as expected (Schoenwald et al., 2011). Figure 1 demonstrates how treatment integrity measures provide the process-level
outcomes that D&I researchers require to better understand implementation integrity and identify areas to investigate further (Schoenwald & Hoagwood, 2001; Southam-Gerow et al., 2008).

The present study used a treatment integrity measure to characterize implementation integrity and generate knowledge to further future implementation research. Specifically, the current study focused on examining the *therapeutic intervention* (i.e., treatment adherence and differentiation) component of implementation integrity. Although the therapeutic relationship and competence also constitute implementation integrity and are critical to successful treatment (Barber, Sharpless, Klostermann, & McCarthy, 2007; McLeod, 2011), interest in therapeutic intervention information has recently increased; stakeholders seek this information to ascertain that the program they have funded is actually being implemented as designed in the practice setting (McLeod et al., 2013; Schoenwald et al., 2011). Moreover, competence, the last technical factor of treatment integrity, assumes adherence, but adherence does not assume competence (McGlinchey & Dobson, 2003; Perepletchikova et al., 2007); unfortunately, an analysis of all implementation integrity components was outside the scope of the current study. Therefore, gathering treatment adherence and differentiation information in the two settings was a good first step and foundation for characterizing implementation integrity.

**Evaluating Implementation of Cognitive Behavioral Therapy for Youth Anxiety**

To conduct this implementation integrity research, the treatment was carefully selected. A treatment must fit the study’s goals, for example by allowing for meaningful cross-setting comparisons. For these reasons, cognitive behavioral therapy (CBT) for youth anxiety best suited the goals of the current implementation integrity research.

For one, CBT for youth anxiety, in comparison to a number of EBTs, has received substantial empirical support in efficacy trials, and is considered “well established” (Silverman et
al., 2008). Indeed, CBT has been found to be superior to a control group at improving clinical outcomes for numerous anxiety disorders in dozens of efficacy studies (Chorpita & Southam-Gerow, 2006). CBT for youth anxiety has performed at a consistent high standard in research settings, and it was therefore worthwhile to try to test this EBT in a multitude of diverse settings.

Moreover, despite its marked performance in research settings, CBT was ideal because it has not always performed as expected when delivered in practice settings. Barrington and colleagues (2005) were some of the first researchers to compare CBT for youth anxiety to UC in an effectiveness study in Australia. They randomly assigned 54 youth with anxiety disorders to UC or CBT in a practice setting, but they did not randomly assign therapists to UC or CBT; therapists who had previous experience in CBT were assigned to the CBT condition. Therapists who had no experience in CBT training and preferred delivering non-CBT interventions were assigned to the UC condition. They found that although a significant amount of youth participants in both conditions no longer met diagnostic criteria for an anxiety disorder (78.00% in CBT, 69.00% in UC), CBT did not outperform UC. However, these findings were difficult to interpret because the lack of therapist randomization and extensive CBT training may have contributed to these results (Southam-Gerow et al., 2010). For instance, without therapist randomization, there was no way to know whether therapist effects (e.g., competence, motivation) accounted for the findings. Confounding therapist and training factors lowered the study’s internal validity.

Southam-Gerow and colleagues (2010) attempted to improve upon the Barrington et al. (2005) study by investigating if CBT would outperform UC in a fully randomized effectiveness trial (i.e., both therapists and clients were randomized to conditions). Unlike Barrington and colleagues (2005), the authors took steps to ascertain that training and supervision were
standardized for therapists in the CBT condition, enhancing the internal validity of their findings. However, despite their efforts, their findings mirrored those of Barrington and colleagues (2005). Although almost three quarters of youth participants in both conditions no longer met diagnostic criteria for their primary anxiety disorder post-treatment (66.70% in CBT, 73.70% in UC), CBT performed statistically equal to UC. Thus, CBT for youth anxiety has not performed as expected in all effectiveness studies.

Southam-Gerow et al. (2010) sought to understand why CBT did not outperform UC. They therefore used treatment integrity measures to assess implementation integrity, examining both treatment adherence and differentiation. They used an 11-item treatment integrity checklist (Kendall, 1994) to measure the CBT adherence of 11 cases from the CBT condition sample (45.80% of CBT condition sample). For these 11 cases, they listened to all of the available sessions from each case; they required the selected cases to have at least six sessions to provide sufficient sampling. One of two doctoral-level clinical psychologists coded tapes from all 11 cases, and the second doctoral-level clinical psychologist independently coded 9.00% of those tapes. Coders achieved adequate reliability (κ = 0.82). They found that a high percentage of the CBT interventions prescribed by the treatment protocol (i.e., included in the protocol) were delivered by therapists in the CBT condition (98.90% contained expected procedures).

Importantly, they also used the Therapy Process Observational Coding System for Child Psychotherapy - Strategies Scale (TPOCS-S; McLeod & Weisz, 2010), an observational measure of therapeutic interventions, to measure treatment adherence and differentiation between the CBT and UC conditions. This 31-item scale contains a broad range of therapeutic interventions that fall within classifications of five treatment domains subscales (Behavioral, Cognitive, Psychodynamic, Client-centered, and Family). Sessions were randomly sampled from each case.
that had pre- and post-treatment outcome data (UC = 19, CBT = 14), with one session sampled from the first and second halves of therapy. There were 70 total sessions coded, each was independently double-coded by graduate students, and scores showed acceptable interrater reliability (ICCs > 0.80).

The TPOCS-S enabled Southam-Gerow and colleagues to overcome a number of limitations faced by Barrington and colleagues (2005). For one, the authors used the TPOCS-S to assess treatment adherence. Two subscales were created from the TPOCS-S to assess CBT interventions. One subscale was entitled the CBT-Coping Cat subscale and it included six CBT therapeutic interventions found in the Coping Cat protocol (Cognitive Education, Cognitive Distortion, Coping Skill, Relaxation, Respondent, and Operant; described in more detail in the Methods). Thus, the TPOCS-S allowed the authors to compare the extensiveness of interventions specifically prescribed to the Coping Cat program (i.e., excluding general therapeutic interventions found in CBT programs) delivered across conditions. The authors also created a CBT subscale that included therapeutic interventions typically found in CBT programs (Cognitive Education, Cognitive Distortion, Coping Skills, Functional Analysis, Relaxation, Respondent, Operant, Modeling, Homework, Role Playing). The TPOCS-S therefore provided the authors with important information about how therapists incorporated a wide variety of CBT interventions.

Furthermore, unlike the treatment integrity checks used in the Barrington et al. study, the TPOCS-S also allowed the authors to assess treatment differentiation because it includes a wide range of therapeutic interventions from five different treatment domains (McLeod et al., 2013). Two features of the study necessitated that a differentiation check incorporate an especially extensive range of therapeutic interventions. First, therapists in practice settings have a more
diverse training history and therefore have a higher likelihood of using therapeutic interventions from other treatment domains when implementing an EBT (Garland et al., 2010; McLeod & Weisz, 2010). Second, families seen in practice settings present with more complex cases (e.g., youths have higher rates of comorbidity; Southam-Gerow et al., 2003); consequently, a therapist may believe it is more appropriate to incorporate therapeutic interventions from other treatment domains when implementing an EBT (Southam-Gerow et al., 2003). Thus, the TPOCS-S gave the authors a way to more exhaustively describe the extensiveness of both prescribed and proscribed therapeutic interventions delivered across conditions.

In addition, the TPOCS-S employs an observational method rather than a self-report method (i.e., participants report data on checklists, dimensional ratings, or diaries). Direct observation, the process in which data are gathered through in-vivo observation, videotape, or audiotape, advantageously fits a diverse array of contexts, and can include measures of multiple domains in one observation (Lane, Bocian, MacMillan, & Gresham, 2004). Although self-reports allow for efficient delivery by personnel with less training than studies using direct observation, this method may be more biased than direct observation; researchers have found that therapists tend to overestimate treatment integrity on self-report measures (DiMatteo, 2004; Noell et al., 2005). The TPOCS-S was therefore a more objective measure than the one used in the Barrington study and a number of past studies.

Finally, the TPOCS-S assesses for variability in implementation integrity. Unlike Barrington and colleagues, who used a measure to assess treatment adherence that tallied frequencies of therapeutic interventions delivered, the TPOCS-S allowed Southam-Gerow et al. to assess the extent that therapists delivered each therapeutic intervention on a scale of 1-7 (1 = not at all, 7 = extensively). This feature was especially valuable given that therapists in both
conditions may have been delivering the same interventions, but with different extensiveness (i.e., frequency and thoroughness). Extensiveness ratings allow researchers to make more descriptive, meaningful comparisons across groups than presence/absence or frequency ratings.

Southam-Gerow and colleagues (2010) presented a number of important findings using the TPOCS-S. They found that UC therapists scored higher on the Client-Centered subscale \((M = 3.66, SD = 0.84, 7\) point scale) compared to the CBT-generic (i.e., interventions typically found in CBT programs; \(M = 2.02, SD = 0.35\)), Psychodynamic (\(M = 2.13, SD = 0.85\)), and Family (\(M = 2.37, SD = 1.10\)) subscales. They also found that therapists in the UC scored significantly higher than therapists in the Coping Cat (CC) condition on the Psychodynamic (UC \(M = 2.13, SD = 0.85\); CC \(M = 1.27, SD = 1.27\)), Family (UC \(M = 2.37, SD = 1.10\); CC \(M = 1.29, SD = 0.39\)), and Client-Centered (UC \(M = 3.66, SD = 0.84\); CC \(M = 3.03, SD = 0.45\)) subscales. Additionally, they found that although CC therapists scored significantly higher than UC therapists on the CBT-Coping Cat Scale, (UC \(M = 1.62, SD = 0.35\); CC \(M = 2.24, SD = 0.52\)), the CBT-generic subscale scores did not significantly differ between the UC and CC groups (UC \(M = 1.90, SD = 0.42\); CBT \(M = 2.02, SD = 0.35\)). These findings indicate that UC therapists delivered higher levels of therapeutic interventions from other treatment domains (i.e., more proscribed interventions) than CC therapists but also that CC therapists did not deliver a significantly higher dosage of generic CBT interventions than UC therapists. Overall, therapists in the CC condition of the YAS study delivered a remarkably low dose of CBT and this could have contributed to the non-significant client-level outcome differences.

However, this interpretation was tentative for two reasons. First, only 70 sessions were coded so the findings could not be interpreted conclusively. More importantly, there was no comparison point; there was no way to determine if the extensiveness of interventions delivered
by the therapists in the CBT condition was lower than the level it was designed to be (McLeod et al., 2013). Thus, before we can answer this question, and further D&I research, we need to identify an implementation integrity comparison point. We need to investigate how CBT for youth anxiety implementation compares to the expected CBT for youth anxiety treatment delivery.

**Benchmarking Implementation Integrity**

The process of comparing the implementation of CBT for youth anxiety to CBT for youth anxiety delivered in research settings can be considered “benchmarking.” The term benchmarking refers to when an optimal comparison point is used to determine an organization’s performance or progress in a particular area (Lai et al., 2011). Benchmarking has been used by a number of fields (e.g., education, healthcare, business) and is a very helpful tool for understanding performance (McLeod et al., 2013). Many organizations compare markers of quality in a particular domain (e.g., customer satisfaction with services) to a benchmark (e.g., customer satisfaction with services of a competing business). This process enables organizations to identify areas of performance that are strengths and weaknesses so that implementation efforts can be increased and concentrated in the areas that require the most development (McLeod et al., 2013).

There are several steps involved in the benchmarking process. One of the first steps in creating benchmarks is to select a suitable comparator (Kennedy, Allen, & Allen, 2002). For example, in a business, a comparator may be a similar high-performing department in another organization. The next step is to operationalize the benchmark. In the instance of a business interested in a “customer satisfaction with services” benchmark, this might consist of measuring customers’ self-reported satisfaction ratings from web surveys administered by the two
organizations (McLeod et al., 2013). Finally, the benchmark can be viewed as constructive feedback; the business can use the quantitative information they acquire regarding their standing relative to the “gold standard” level (attained in the similar high-performing department) to determine whether actions should be taken toward improvement (McLeod et al., 2013).

Unfortunately, few psychosocial treatment studies have conducted benchmarking studies, and those that have focused mostly on client-level outcomes (Franklin, Abramowitz, Kozak, Levitt, & Foa, 2000; Persons, Bostrom, & Bertagnolli, 1999; Wade, Treat, & Stuart, 1998; Warren & Thomas, 2001; Weersing & Weisz, 2002), not therapist behavior (i.e., implementation integrity) or process-level outcomes. D&I researchers have, however, brainstormed ways in which the mental health field could use benchmarking to assess implementation integrity (McLeod et al., 2013). Specifically, in the mental health field, researchers could capitalize on the important information afforded by benchmarking implementation integrity by choosing one of the following benchmark comparators: (a) research standard (i.e., efficacy trials), (b) high-performing external unit (i.e., similar unit in another organization), and (c) high-performing internal unit (i.e., unit within the same organization; McLeod et al., 2013). These comparators could be selected according to the study’s goals, and used to better interpret the outcomes of EBT implementation (McLeod et al., 2013).

In particular, choosing an efficacy trial as a benchmark comparator may have value for D&I researchers. Researchers could get a first glimpse at whether EBT implementation in practice settings roughly meets the “gold standard” delivery observed in research settings (Schoenwald et al., 2011). Stakeholders in particular would appreciate this knowledge given their financial investment in an EBT’s effectiveness. Moreover, this information could guide future D&I researchers in initiating quality control changes (e.g., therapist training and
supervision) or altering the EBT program itself (McLeod et al., 2013). For instance, it could be that aspects of “gold standard” EBT delivery do not predict “gold standard” client outcomes, suggesting that we need to change the way EBTs are designed. Thus, although an efficacy benchmark may not necessarily allow researchers to compare EBT implementation to optimal EBT delivery (i.e., delivery that produces the best client outcomes), it allows researchers to compare EBT implementation to EBT delivery as it was designed.

The current study therefore aimed to add to D&I literature by being the first to compare CBT implementation for youth anxiety to an efficacy study benchmark. An efficacy study was the comparator that best suited the present study’s goals because it afforded an examination of whether EBTs are implemented as designed in practice settings (Southam-Gerow et al., 2010). Benchmarking implementation integrity using efficacy study data showed researchers where delivery in practice settings aligned or diverged from “gold standard” delivery in research settings, to identify areas of implementation that need to be investigated in future research and clinical endeavors.

After selecting the appropriate comparator, the next step in a benchmarking study is to operationalize the benchmark. The instrument selected for benchmarking should be determined by what the study aims to measure. The current study aimed to benchmark implementation integrity, so an instrument that assesses the extensiveness of both treatment adherence and treatment differentiation was used. In fact, the current study used a measure that meets a number of specific requirements, and has provided critical information in past studies (i.e., Southam-Gerow et al., 2010).

**The Therapy Process Observational Coding System – Revised Strategies Scale**

To meet the needs of the current study, a measure needed to (a) assess CBT interventions
found in the treatment protocol (i.e., adherence), (b) assess a wide range of interventions (i.e., differentiation), (c) be an observational measure, and (d) assess variability in implementation. As described in the Southam-Gerow et al. (2010) study, the TPOCS-S (McLeod & Weisz, 2010) met these requirements. For these reasons, a slightly revised version of the TPOCS-S, the *Therapy Process Observational Coding System – Revised Strategies Scale* (TPOCS-RS), was a natural fit for the current study.

Importantly, the TPOCS-RS could be used to assess treatment adherence. It contains a list of six core CBT interventions that can be placed on a CBT-Coping Cat subscale. This component is critical for characterizing the extent to which the prescribed therapeutic interventions are delivered in research and practice settings. As aforementioned, knowledge of adherence data has been increasingly sought by stakeholders (McLeod et al., 2013; Schoenwald et al., 2011). The TPOCS-RS allowed the current study to depict this important information and bring researchers a step toward furthering D&I research.

TPOCS-RS also best suited the purposes of the present study because it was an ideal measure of treatment differentiation (McLeod et al., 2013). The TPOCS-RS includes a wide range of therapeutic interventions, requisite to studying therapists in practice settings who have a diverse training history (Garland et al., 2010). Since the TPOCS-RS includes a number of therapeutic interventions that align with treatment domains other than CBT, I obtained a clearer, more exhaustive evaluation of how delivery of interventions prescribed by the CBT program in practice settings compares to the “gold standard” level of prescribed interventions delivered in research settings (Waltz, Addis, Koerner, & Jacobson, 1993).

In particular, although past studies have examined the extent to which therapists adhere to or depart from the treatment protocol, few researchers have additionally assessed the extent that
therapists deliver *proscribed* therapeutic interventions, as defined in the current study as therapeutic interventions that were purposely not included in the protocol due to being implicit violations of the treatment domain (McLeod et al., 2013). Past studies have explored whether therapists diverge from the protocol in patterned ways (e.g., Barber et al., 2004), but few have indicated whether these different interventions are proscribed (e.g., homework in Psychodynamic therapy, or interpretation in CBT). In a review of studies assessing treatment adherence, differentiation, and competence, Bhar and Beck (2009) found that no study explicitly mentioned proscribed (i.e., prohibited) therapeutic interventions in their treatment protocols, and only four of the nine studies assessed treatment differentiation. Treatment differentiation was poorly defined in the majority of the studies they reviewed (e.g., Cooper, Murray, Wilson, & Romaniuk, 2003;Crits-Christoph et al., 1999; Svartberg, Stiles, & Seltzer, 2004).

Moreover, a few past studies comparing two treatments considered therapeutic interventions to be proscribed if they adhered to the competing protocol under analysis (e.g., Carroll et al., 1998; Hogue et al., 1998; Hogue, Dauber, Samuolis, & Liddle, 2006). However, these studies failed to take into account therapeutic interventions that were not prescribed to any of the treatment domains under analysis. The TPOCS-RS overcomes these limitations by taking into account non-prescribed therapeutic interventions that fall under three distinct treatment domains (Client-centered, Psychodynamic, Family), rather than just one competing treatment domain. Benchmarking the therapeutic interventions delivered using the TPOCS-RS enabled me to characterize when therapists in practice settings “add” therapeutic interventions that were not included in the treatment protocol and not expected in the delivery of CBT. Overall, the TPOCS-RS allowed the current study to measure treatment differentiation in a comprehensive manner with the intention of furthering D&I research.
Finally, the TPOCS-RS met the current study’s requirements of being observational and assessing for variability. As previously described, these capacities were essential for ensuring that the study was objective and thorough, respectively (Hill, 1991; Mowbray et al., 2003). By objectively describing the degree to which interventions were delivered in both settings, I aimed to clarify important questions regarding the science-practice performance gap, as well as help to inform future quality control efforts.

The TPOCS-RS fulfilled all of the requirements asked of a measure in the current study. Now that I have reviewed the relevant youth mental health, EBT, D&I, and treatment integrity literature, presented a model of implementation integrity, and described the measure that allowed for the current study’s analysis, I will remind the reader of the central problem the present study aimed to address.

**Problem Statement**

Efficacious EBTs may not produce the same beneficial effects for youth seen in practice settings as they do in research settings. When certain EBTs are transported and delivered in practice settings, they sometimes perform equal to or worse than treatment typically delivered in practice settings, sometimes called UC (Guarton, 1992; Southam-Gerow et al., 2010). Thus, the many efficacious treatments that researchers spend years in the lab perfecting are not always performing as expected in practice settings so youth in practice settings may not be receiving the quality of care they deserve.

EBTs may sometimes fail to perform as expected in practice settings because practice settings differ from research settings in a variety of ways. In particular, we know that that there are therapist, client, and contextual differences between the two settings that could influence
EBT implementation and outcomes, but we do not know how exactly these factors make a
difference (Garland et al., 2010; Southam-Gerow et al., 2003). Therefore, researchers do not
have the requisite information to identify ways to improve EBT implementation or the EBT
protocol itself, and subsequently begin to close the science-practice performance gap.

The current study aimed to help researchers begin to understand the science-practice
performance gap through an exploration of implementation integrity. In essence, I aimed to
characterize what actually takes place within the therapy session. I used CBT for youth anxiety
to test the hypothesis that EBT implementation differs from EBT delivery in research settings.
Specifically, I compared treatment adherence and differentiation levels in practice settings to
treatment adherence and differentiation levels attained in research settings. This preliminary
benchmarking study aimed to show researchers and stakeholders areas where CBT
implementation approximately met or fell short of “gold standard” CBT delivery so that,
eventually, future researchers could use this knowledge to inform quality control efforts (e.g.,
improve CBT training in practice settings) or alter the treatment protocol (to best suit the
practice-setting context).

Hypotheses

It was hypothesized that therapists who delivered CBT for youth anxiety in practice
settings would more often add therapeutic interventions or fail to adhere to the CBT protocol
than therapists who delivered CBT in research settings (Bickman, 2000; Garland et al., 2010;
Weisz et al., 2006). More specifically, it was hypothesized that (a) therapists in the CBT practice
setting condition, herein referred to as YAS-CC, would deliver significantly more therapeutic
interventions proscribed by the Coping Cat program (i.e., significantly higher scores on the
TPOCS-RS Family and Psychodynamic subscales) than therapists in the CBT research setting condition (Southam-Gerow et al., 2010), herein referred to as ICBT, (b) therapists in the UC practice setting condition, herein referred to as YAS-UC, would deliver significantly more therapeutic interventions proscribed by the Coping Cat program than therapists in both the YAS-CC and ICBT conditions (i.e., significantly lower scores on the TPOCS-RS Family and Psychodynamic subscales; Garland et al., 2010; Weisz et al., 2006), (c) therapists in YAS-CC would deliver significantly less therapeutic interventions prescribed by the Coping Cat program than therapists in ICBT (i.e., significantly lower scores on the TPOCS-RS CBT- Coping Cat subscale; Southam-Gerow et al., 2010), and (d) therapists in YAS-UC would deliver significantly less therapeutic interventions prescribed by the Coping Cat program than therapists in both YAS-CC and ICBT (i.e., significantly lower scores on the TPOCS-RS CBT-Coping Cat Scale; Garland et al., 2010).

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 Therapy (ICBT), Family Cognitive Behavioral Therapy (FCBT), and Family-based education/support/attention (FESA), at Temple University (Kendall, Hudson, Gosch, Flannery-Schroeder, & Suveg, 2008). The second study, the Youth Anxiety Study (YAS), compared the effectiveness of CBT to UC in six outpatient community mental health clinics in Los Angeles County (Southam-Gerow et al., 2010). In the current study, participants were drawn from the individual CBT condition of the Kendall et al. study (i.e., ICBT; \(N = 55\))
and the entire sample of the YAS study (YAS-UC \(N = 24\) and YAS-CC \(N = 24\)). See Table 1 for descriptive information for each parent study.

Table 1

*Demographic and Descriptive Information by Parent Study*

<table>
<thead>
<tr>
<th>Variable</th>
<th>ICBT</th>
<th>YAS</th>
<th>Significant Difference (Y/N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Youth age in years ((SD))</td>
<td>10.37 ((1.88))</td>
<td>10.90 ((2.10))</td>
<td>N</td>
</tr>
<tr>
<td>Percent of female youths</td>
<td>41.80</td>
<td>56.20</td>
<td>Y</td>
</tr>
<tr>
<td>Percent of Caucasian youths</td>
<td>83.64</td>
<td>38.50</td>
<td>Y</td>
</tr>
<tr>
<td>Percent of youths with principle diagnoses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAD</td>
<td>47.00</td>
<td>12.50</td>
<td>-</td>
</tr>
<tr>
<td>SAD</td>
<td>38.00</td>
<td>37.50</td>
<td></td>
</tr>
<tr>
<td>SP</td>
<td>-</td>
<td>22.90</td>
<td></td>
</tr>
<tr>
<td>SOP</td>
<td>45.00</td>
<td>27.01</td>
<td></td>
</tr>
<tr>
<td>Total diagnoses at initial assessment ((SD))</td>
<td>3.56 ((1.74))</td>
<td>3.20 ((1.80))</td>
<td>N</td>
</tr>
<tr>
<td>Therapist age in years ((SD))</td>
<td>missing</td>
<td>33.67 ((9.59))</td>
<td>-</td>
</tr>
<tr>
<td>Therapist years of training ((SD))</td>
<td>missing</td>
<td>4.40 ((2.20))</td>
<td>-</td>
</tr>
</tbody>
</table>
Percent of therapist degrees

<table>
<thead>
<tr>
<th>Degree Level</th>
<th>Missing</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Worker</td>
<td>27.30</td>
<td></td>
</tr>
<tr>
<td>Masters-level</td>
<td></td>
<td>51.50</td>
</tr>
<tr>
<td>Doctoral-level</td>
<td></td>
<td>9.10</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>12.10</td>
</tr>
</tbody>
</table>

Note. ICBT = Individual Cognitive Behavioral Therapy condition of Kendall et al. study, YAS = Youth Anxiety Study, Y = yes, N = no, GAD = generalized anxiety disorder, SAD = separation anxiety disorder, SOP = social phobia, SP = specific phobia. In the ICBT study, a number of youths had more than one primary diagnosis.

**Kendall et al. study.** Participants in the ICBT condition of the Kendall et al. study consisted of youth that were recruited through multiple sources (e.g., clinics and practitioners, public/nonpublic schools, flyers, media). Their parents contacted the Child and Adolescent Anxiety Disorders Clinic (CAADC) at Temple University (where the study was conducted) and were given a brief phone screen. If the youth was between ages 7 and 14 and the parent was interested in participating in the study, an intake interview was scheduled. 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. A composite approach (using the “or” rule) recommended by the ADIS-C/P (Silverman & Albano, 1996), was used to determine the youth’s diagnoses; the diagnosis was assigned if the youth and/or parent reported that the youth was experiencing symptoms associated with a particular diagnosis. If the youth met diagnostic criteria for an anxiety disorder, the clinician assigned a clinician severity rating (CSR) that ranged from 1 to 8 (4 or greater indicated the presence of an anxiety disorder). The principal disorder was determined by the severity of the CSR on the integrated youth and parent report. Exclusion criteria included: a disabling medical condition, psychotic symptoms, intellectual disability, the youth’s taking medication for anxiety or depression, or the youth’s
participation in concurrent psychosocial treatment. In addition, one parent was required to be fluent in the English language. Parents provided informed consent and the youth provided assent. 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 in the ICBT condition consisted of 161 youths and their parents who were referred by community sources between 2000 and 2006 and were diagnosed with a primary anxiety disorder (generalized anxiety disorder [GAD], separation anxiety disorder [SAD], social phobia [SOP]). From Kendall et al.’s sample, the current study drew from the 55 participants assigned to receive ICBT (attrition of five youths). Youth participants were excluded from the current sample \( (n = 4) \) if they received treatment from multiple therapists and/or if they had TPOCS-RS data from fewer than two sessions (since it is required to have at least two time points to assess for treatment integrity over time). Therefore 51 youth remained in the final sample.

Of the 51 youth participants, ages 7-14 \( (M = 10.36, \ SD = 1.90) \), 39.20% were females and 86.30% were Caucasian. Based on information collected with the ADIS, 37.30% of the youth were diagnosed with a principal diagnosis of GAD, 29.40% with SAD, and 33.30% with SOP (approximately 15 youth had more than one principal diagnosis; diagnoses had equal CSR ratings and were therefore deemed “co-principal”). 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. Participants reported family income as below $20,000 (2.13%), between $20,000 and $40,000 (17.01%), between $40,000 and $60,000 (19.15%), between $60,000 and $80,000 (19.15%), and above $80,000 (42.56%).
**Therapist participants.** In the Kendall et al. parent study, ICBT was delivered by 18 masters- and doctoral- level therapists who were recruited through the CAADC. Two therapists were removed from the parent sample’s ICBT condition due to exclusion criteria (e.g., cases involving multiple therapists). The data regarding therapist’s age in the ICBT group were not reported but the therapists 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.

**Youth anxiety study.** Participants in the YAS parent study consisted of clinically referred families that called one of the six community mental health clinics in the Los Angeles area. The families were asked to describe the youths’ problems that had led them to seek services. If the problems included symptoms of anxiety disorders, the youth was between 8 and 15 years of age, and parental consent was obtained, the families were invited to participate in a standardized diagnostic interview. If the youth (a) met DSM-IV criteria for a primary anxiety disorder (SAD, social phobia [SOP], SP, GAD) then they were invited to participate in the study. Unlike the Kendall et al. study, in which diagnoses were determined by the ADIS-C/P, diagnoses for the YAS study were determined based on the combined parent and youth report of the Diagnostic Interview Schedule for Children Version 4.0 (DISC 4.0). The project team, clinic staff, and family met to discuss the youth’s diagnosis, referral problem, symptoms, and functioning, and to determine whether anxiety was a priority in treatment. Youths with an intellectual disability or psychotic symptoms were excluded from the study, but youths receiving medication were included to appropriately represent the population and setting of youths receiving services in practice settings. Medication was tracked during treatment; decisions about
medication were made according to standard clinic procedures in both UC and CBT conditions. In addition, families were allowed to seek additional mental health services during the study as they saw fit. The institutional review board approved the study and compensation was provided to study participants.

Youth participants. In YAS, the parent sample consisted of 48 youth participants ages 8-15 ($M = 10.90$, $SD = 2.10$). These youths enrolled in the study and met diagnostic criteria for an anxiety disorder as determined by standardized diagnostic interviews and parents who judged that an anxiety disorder should be the primary treatment focus. As in the Coping Cat study, participants were excluded from the current sample if they received treatment from multiple therapists ($n = 2$) and/or if they had TPOCS-RS data from fewer than two sessions ($n = 8$). Therefore, 40 youths were retained in the sample, 19 youths in YAS-CC ($M$ age = 11.22, $SD$ age = 2.29; 36.80% Caucasian; 68.40% female) and 21 in YAS-UC ($M$ age = 10.44, $SD$ age = 1.91; 33.30% Caucasian; 47.60% female). Primary diagnoses determined by combined parent and youth structured interviews were 35.00% SAD, 25.00% SOP, 27.50% SP, 10.00% GAD, and 2.5% Agoraphobia. Approximately 35.00% of participants met criteria for the diagnosis of a comorbid anxiety disorder only, 60.00% met criteria for a comorbid non-anxiety disorder, and 5.00% were missing comorbidity information. Of the 40 youths, 57.50% were females, 35.00% were Caucasian, 5.00% were African American, 32.50% were Latino, and 7.50% self-identified as mixed race; eight participants elected not to report their ethnic identity. In regards to family income, 32.50% of families reported an annual family income under $15,000, 32.50% reported between $15,000 and $30,000, 5.00% reported between $30,000 and $45,000, 5.00% reported between $45,000 and $60,000, 2.50% reported between $60,000 and $75,000, 12.50% reported over $90,000, and 10.00% of family income information was not reported.
Therapist participants. The parent sample of the YAS study included 39 therapists who were recruited at one of six community mental health clinics; these therapists were clinic employees who volunteered to participate in the YAS study. They were randomly assigned to deliver either UC \((n = 21)\) or CBT \((n = 18)\). Nine therapists were excluded from the current sample due to exclusion criteria, thus 16 remained in UC and 14 remained in CBT. These therapists were mostly female \((86.70\%)\), averaged 33.97 \((SD = 10.13)\) years of age, 4.40 \((SD = 1.99)\) years of training, and 5.30 \((SD = 8.50)\) years of additional professional experience. Approximately half were Caucasian \((46.70\%)\), 30.00\% Latino, 6.70\% Asian/Pacific, and 13.30\% mixed ethnicity; one therapist did report his/her ethnicity. The therapists consisted of 36.70\% masters-level psychologists, 26.70\% social workers, 10.00\% doctoral-level psychologists, and 10.00\% other (e.g., marriage and family therapists); five therapists did not report their degrees.

Treatment Methods

CBT. In both the Kendall et al. and YAS studies, therapists were trained to use Coping Cat, a CBT program aimed at treating youth diagnosed with anxiety disorders (Kendall & Hedtke, 2006; Kendall, Kane, Howard, & Siqueland, 1990). Coping Cat includes 16-20 sessions conducted individually with the youth participant. This program teaches youths skills to manage anxiety using a FEAR acronym: (a) Feeling frightened? (youths identify anxious symptoms), (b) Expecting bad things to happen? (youths recognize anxious thoughts), (c) Actions and attitudes that can help (youths and therapists collaborate on a way to cope by changing negative self-talk and promoting coping behavior), and (d) Results and Rewards (therapists teach youths to self-reinforce). Therapeutic interventions such as graduated in-vitro and in-vivo exposure tasks, role playing, and contingent reinforcement are delivered throughout treatment, and homework is regularly assigned to the youth. Parents are included in treatment during sessions 4 and 9 of the
program to inform them about youth anxiety and their child’s treatment progress as well as to collect relevant information about the youth’s functioning outside of the sessions.

 Therapists in the Kendall et al. parent study were trained in all treatments (ICBT, FCBT, FESA). Therapists in the ICBT condition (the condition drawn for the current study) studied the Coping Cat protocol and participated in training (two 3-hr 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 treatment. Therapists in the CBT condition of the YAS parent study also studied the Coping Cat protocol and participated in a training workshop (one-day, 6-hours) before beginning supervised trial experience. They were supervised weekly, throughout treatment by one of two doctoral-level psychologists with expertise in CBT.

 Usual care. YAS 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.

 Diagnostic and Symptom Measures

 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 with established reliability (Silverman, Saavedra, & Pina, 2001) and convergent validity with other measures (Wood, Piacentini, Bergman, McCracken, & Barrios, 2002). This interview was used in the Kendall et al. study to assess the youth’s symptom correspondence to DSM-IV disorders. Independent evaluators assessed youth DSM-IV disorders and symptom counts based on separate interviews with the caregivers and the youth, and made ratings on the
ADIS-C/P Clinician’s Severity Rating scale (CSR; 0 = not at all, 4 = some, 8 = very, very much) for each assigned anxiety diagnosis. Clinical levels are at ratings of 4 and above.

Doctoral clinical psychology students were trained to conduct the semi-structured interviews using procedures recommended by the ADIS-C/P authors. They were required to achieve and maintain an inter-rater diagnostic reliability of 0.85 (Cohen’s K; Silverman & Albano, 1996). These trained independent 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. All treatment sessions were videotaped and monitored to ensure adherence to the ADIS-IV protocol. Specifics of the ADIS-C/P interviewing procedures and psychometric properties are provided elsewhere (Silverman & Albano, 1996).

**YAS. Diagnostic Interview Schedule for Children Version 4.0** (DISC 4.0). The DISC is a highly structured interview with rich psychometric data (Schwab-Stone, Shaffer, Dulcan, & Jensen, 1996; Shaffer, Fisher, Dulcan, & Davies, 1996). This interview was used in the YAS study to assess the youth participant’s symptom correspondence to DSM-IV disorders (Rubio-Stipec et al., 1996). The DISC is computer-administered and respondent-based, meaning the interviewer asks standard questions requiring simple answers from the interviewees (e.g., yes or no responses). The youth responses were combined with separate parent responses to determine diagnoses.

Pairs of evaluators conducted the structured assessment interviews (a research assistant and clinical psychology graduate student). These evaluators independently conducted separate assessment interviews with the parents and youth to assess DSM-IV diagnoses. They were trained to adhere to exact DISC-IV administration procedures, and were blind to condition (UC
vs. CBT). Interviews were recorded and randomly sampled by study supervisors throughout the study to assess administrative standardization. Specifics of the DISC 4.0 interviewing procedures can be found in Shaffer et al. (1996).

*Child Behavior Checklist for Ages 6-18* (CBCL/6-18; Achenbach, 1991). The CBCL/6-18 is a 118-item checklist designed to assess a wide range of emotional and behavioral problems and was collected in both Kendall et al. and YAS studies. The CBCL is a component of the Achenbach System of Empirically Based Assessment, a set of multi-informant measures designed to assess the adaptive and maladaptive functioning of people of all ages (Achenbach, 2012). In the current study, the school age assessment forms were used, the CBCL/6-18. For this measure, 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 measure generates *T* scores that reflect a youth’s status relative to others of the same gender and age (e.g., internalized distress).

This measure 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). Respondents with a *T* score of 65 or above on the broadband scales may be in need of treatment. Overall, the CBCL/6-18 has evinced high 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 used the broadband Total, Internalizing, and Externalizing subscales to describe the level of symptomatology across the samples.
Assessment Procedures

**Kendall et al. study.** If a youth met inclusion criteria and was invited 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 that there was a balance of participants across conditions. Cases were randomly assigned to therapists to control for potential therapist variables (i.e., double randomization: both therapists and clients were randomly assigned to condition). Therapists only delivered one of the three treatments in the study. Multi-rater, multi-domain assessments were administered pre-treatment (T1), post-treatment (T2) and at a 1-year follow up (e.g., ADIS, CBCL). Throughout treatment, techniques were used to ensure the post-treatment diagnosticians were blind to condition.

**YAS.** If youth met inclusion criteria and were invited to participate in the YAS parent study, youth participants were assigned to either UC or CBT using block randomization to ensure balance of clinic, gender, and bilingual characteristics. Therapists were also assigned to UC or CBT using a randomized block procedure to ensure balance of therapists and discipline representation. Therapists only delivered one of the two conditions in the study. The Diagnostic Interview Schedule for Children Version IV (DISC-IV; Shaffer et al., 1996) and the youth and parent responses to the Child Behavior Checklist (CBCL; Achenbach, 1991) were used to assess diagnoses and symptoms. These multi-rater, multi-domain assessments were conducted before the start of therapy (T1), and immediately following treatment (T2) for both conditions. The post-treatment assessments were conducted according to the individual case since treatment duration was allowed to vary in both conditions.
Treatment Adherence and Differentiation Measure

*Therapy Process Observational Coding System for Child Psychotherapy – Revised Strategies Scale* (TPOCS-RS; McLeod & Weisz, 2010). The TPOCS-S is a 31-item scale used to measure the extent that therapists deliver therapeutic interventions within five core treatment domains (Cognitive, Behavioral, Psychodynamic, Family, Client-Centered). It was originally designed to characterize UC, and was developed based off of the Therapy Process Checklist (Weersing, Weisz, & Donenburg, 2002) and the Therapist Behavior Rating Scale (Hogue, Liddle, & Rowe, 1996). Coders watch or listen to therapy tapes and rate the extensiveness of each therapeutic intervention item that considers two qualities: frequency and thoroughness (McLeod & Weisz, 2010). The 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. Specifically, coders rated each item on a seven point Likert-type scale with the anchors: 0 = *not at all*, 4 = *considerably*, 7 = *extensively*. For example, if an intervention was delivered at a low frequency, but at great depth, the coder might code that intervention item as a 4 or 5 in extensiveness.

Overall, the TPOCS-S has demonstrated sufficient internal consistency, inter-rater reliability, and construct validity (McLeod & Weisz, 2010; Southam-Gerow et al., 2010; Weisz et al., 2009; Wood, Piacentini, Southam-Gerow, Chu, & Sigman, 2006).

The current study used the original 31 TPOCS-S items (McLeod & Weisz), with the addition of 11 items defined in more detail below (also known as the TPOCS-RS). Three items were added to the Behavioral subscale (Behavioral Activation, Monitoring, Skill Building), one
was added to the Psychodynamic subscale (Resistance), two were added to the Family subscale (Family Member’s Roles, Parenting Skill), and five were added as general items (Advice, Assessment, Coaching, Questioning, Self-disclosure). At the subscale level, the present study analyzed all five original TPOCS-RS subscales (Cognitive, Behavioral, Psychodynamic, Family, Client-Centered). The current study also analyzed an additional subscale that was developed in the Southam-Gerow et al. study (2010): the CBT-Coping Cat subscale (Cognitive Education, Cognitive Distortion, Coping Skills, Relaxation, Respondent, and Operant) to assess treatment adherence. The current study did not analyze the general items on the TPOCS-RS (e.g., Assessment) because the primary goal of the study was to characterize prescribed and proscribed interventions; an analysis of general items fell outside the scope of the current study.

In the current study, item scores on the CBT-Coping Cat subscale were considered an adequate representation of “prescribed” therapeutic interventions delivered, since these therapeutic interventions constitute the core interventions used in Coping Cat. The following six items fall within the CBT-Coping Cat subscale and were thus deemed prescribed: Cognitive Education (teaches client cognitive model), Cognitive Distortion (teaches and/or encourages the client to identify and/or restructure cognitive distortions) Coping Skills (teaches client to use coping skills), Relaxation (teaches client to use relaxation), Respondent (develops a hierarchy, employs mastery ratings, or performs exposure), and Operant (teaches principles of operant interventions).

Items on two TPOCS-RS subscales, Psychodynamic and Family, were independently considered an appropriate representation of proscribed therapeutic interventions delivered, since these therapeutic interventions are not included in Coping Cat and on account of their omission could be considered “prohibited” by protocol developers (Schoenwald et al., 2011; Waltz et al.,
1993). The following 4 items fall within the Psychodynamic subscale and were thus considered proscribed: Transference *(discusses or interprets client’s interaction with the therapist)*, Explores Past *(discusses client’s past experiences)*, Resistance *(identifies and discusses client’s resistance to therapy and/or resistance to change, including motivation to be in therapy)*, and Interpretation *(comments on client behavior and/or relates that behavior to an aspect of the client’s characteristics, general functioning, and/or past experiences)*. The following six items fall within the Family subscale and were similarly considered proscribed: Targets Other Participants *(participants other than the target youth are asked to modify their affect, behavior, cognitions)*, Recruits Others *(attempts to recruit/retain parents and other supporters for future sessions)*, Parenting Style *(encourages parents to modify their parenting practices)*, Parenting Skill *(parent-focused interventions designed to improve ability to set limits, ensure adequate supervision, and monitor youth behavior, activities, and involvements)*, Multiparticipant *(establishes, teaches, and discusses in-session multiparticipant interactions)*, and Family Roles *(teaches or emphasizes how emotional/behavioral problems may be caused or maintained by the environmental context/family dynamics, etc.)*.

Items that fall on the Client-centered subscale, cognitive and behavioral items not on the CBT Coping Cat subscale, and the “general” items were not categorized as prescribed or proscribed, since many of these items are universally used and expected in treatment, or are allowed but not unique and fundamental to CBT for youth anxiety (Waltz et al., 1993). These items included interventions such as Validate Client *(validates clients feelings or treatment goals)*, Positive Regard *(responds to client in warm and compassionate manner)*, and Rehearsal *(encourages client to participate in hypothetical enactments)*. The differences between the client-centered therapeutic interventions delivered in all three settings were analyzed in the exploratory
component of the analyses since few past researchers have investigated contextual differences in this area of implementation integrity (Garland et al., 2010).

**Coding and Session Sampling Procedures**

**Coders.** Two female doctoral students in clinical psychology (one Asian-American and one Caucasian) comprised the coding team for the current study and averaged 25.18 years of age ($SD = 1.77$) at the start of the study. These coders were responsible for coding sessions for the extent that therapeutic interventions were delivered during therapy using the TPOCS-RS (McLeod & Weisz, 2010). The coders were blind to condition and coded sessions in a randomly assigned order.

**Coder training.** Coders trained for six months on the TPOCS-RS. Training entailed a thorough reading of the scoring protocol, coding practice sessions, and attending regular training meetings. To code tapes for the study, coders had to reach a reliability level that was considered adequate for each item (ICC (2,2) of at least 0.60; Cicchetti, 1994) on a set of 32-criterion recordings consensus coded by the study PIs. After coders were certified for coding, recordings were randomly assigned. During coding, regular meetings were held and inter-rater reliability was regularly assessed to prevent coder drift.

**Sampling of therapy sessions.** All available recordings, except the first and last session for each client, from both studies were selected from each case for coding and randomly assigned to coders. The final sample of TPOCS-RS scores consisted of 977 recordings that were coded by both coders (532 sessions in ICBT, 235 sessions in YAS-CC, 210 sessions in YAS-UC).

**Scoring of therapy sessions.** Each session (excluding the intake and termination sessions) was double coded. Since the average score across coders provided a more reliable
estimate of the actual therapeutic interventions delivered (due to its interval rating scale format),
the scores used in the analysis represented the average score of both coders for each item. Taking
the average of items in each measure, rather than using a total score, ensured that the score
applied in the analyses fell on the same metric as the original subscale (McLeod, Islam, &
Wheat, 2013).

Comparing the Method and Results of the Two Parent Studies

Kendall et al. study. Both therapists and participants were randomized to ICBT, FCBT,
or FESA in a university lab. Therapists in both the ICBT condition and FCBT condition used the
Coping Cat protocol for youth with anxiety (Kendall & Hedtke, 2006); however, the FCBT
condition included an additional component aimed to alter parental beliefs, expectations, distress,
and communication skills. Whereas parents attended two sessions in ICBT and were considered
“collaborators,” parents in FCBT attended all but two sessions and were considered “co-clients”
(Kendall et al., 2008). Therapists in the FESA condition used a protocol for family education,
support, and attention for youth with anxiety and also included a workbook to increase youth
engagement; this condition warranted attendance of both parents and youths in all sessions. All
conditions produced positive outcomes; 64.00%, 64.00%, and 42.00% of principal diagnoses
were no longer present for ICBT, FCBT, and FESA, respectively post-treatment. On the CBCL
Internalizing scale, mother-reports demonstrated no significant differences in the proportion of
youths in each condition exceeding the clinical cutoff versus returning within the normal limit at
post-treatment (ICBT $M = 62.17$, $SD = 8.30$; FCBT $M = 60.06$, $SD = 9.51$; FESA $M = 61.17$, $SD$
= 9.68). However, for father-reported CBCL data, youths in ICBT and FCBT conditions were
significantly more likely than youths in the FESA condition to fall within the normal range
(below the clinical cutoff of $T$ scores > 65) post-treatment (ICBT $M = 58.37$, $SD = 10.48$; FCBT
In general, ICBT and FCBT resulted in post-treatment outcomes that surpassed those of FESA in reducing the presence and principality of the anxiety disorder, and these outcomes were maintained at the one-year follow up. The ICBT and FCBT condition produced comparable outcomes, suggesting that extensive parental involvement in CBT may not be integral to positive outcomes. Overall, this study suggests that CBT outperforms FESA and similar programs.

**YAS.** In the YAS study, both therapists and participants were randomized to UC and CBT conditions and were treated in six community mental health clinics. Therapists in the CBT condition delivered the Coping Cat program (Kendall et al., 1990) whereas therapists in the UC condition used the treatment interventions they regularly used and thought to be effective. In both conditions, more than half of the youth no longer met criteria for an anxiety disorder by the end of treatment. However, UC and CBT groups did not significantly differ on diagnostic or symptom outcomes; 66.70% and 73.70% of youths who were diagnosed with a primary anxiety disorder at the beginning of the study no longer met diagnostic criteria for that disorder post-treatment in CBT and UC conditions, respectively. On the CBCL Internalizing scale (clinical cutoff of \( T > 65 \)), post-test \( T \) scores were not statistically different (CBT \( M = 58.87, SD = 9.03; \) UC \( M = 56.12, SD = 10.36 \)), a finding that did not support the authors’ hypothesis and indicates the need for future research. Nevertheless, results showed that youth in the CBT condition received fewer additional mental health services, suggesting that CBT may be more cost effective. See Table 2 for a summary and comparison of findings from both studies.
Table 2

*Results of Parent Studies*

<table>
<thead>
<tr>
<th></th>
<th>ICBT</th>
<th>YAS-CC</th>
<th>YAS-UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic Remission Rates of Primary Anxiety Disorder</td>
<td>64.00%</td>
<td>66.70%</td>
<td>73.70%</td>
</tr>
<tr>
<td>CBCL Internalizing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Treatment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>67.92</td>
<td>66.52</td>
<td>66.30</td>
</tr>
<tr>
<td>$SD$</td>
<td>8.25</td>
<td>9.23</td>
<td>8.23</td>
</tr>
<tr>
<td>Post-Treatment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>62.17</td>
<td>58.57</td>
<td>56.12</td>
</tr>
<tr>
<td>$SD$</td>
<td>8.30</td>
<td>9.03</td>
<td>10.36</td>
</tr>
</tbody>
</table>

*Note.* ICBT = CBT in research setting, YAS-CC = CBT in practice setting, YAS-UC = usual care. CBCL Internalizing data are reported in $T$ scores.

**Results**

Analyses progressed through two phases. First, preliminary analyses were conducted to ascertain that the data were reliable, valid, and representative of the data we intended to study. This included the calculation of ICCs, alphas, chi-square tests, linear regressions, ANOVAs, and paired sample $T$ tests. Next, primary analyses were conducted to finalize the sample, score treatment domain subscales, and test the study hypotheses. This phase mainly consisted of hierarchal linear modeling analyses.

**Preliminary Analyses**

Preliminary analyses were conducted to evaluate the psychometric properties of the TPOCS-RS, the distribution and patterns of missingness in the data, and the extent that the
sample represented the parent studies and differed across conditions.

**Evaluation of psychometric properties.** Interrater reliability was calculated for each TPOCS-RS item and subscale based on a two-way random effects model for the average of the two coders, ICC (2,2). 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. This model was chosen because the same two coders coded each session and the average of their ratings represent the unit of analysis (Shrout & Fleiss, 1979). For the full sample of recordings, interrater reliability ICCs ranged from 0 - .96 ($M = .77, SD = .18$) for all individual items and 0 - .96 ($M = .74, SD = .22$) for all individual items under analysis. The interrater reliability for 25 of the 42 items fell in the “excellent” range, 14 items fell in the “good” range, one fell in the “fair” range, and two items (Functional Analysis, Behavioral Activation) fell in the “poor” range. Interrater reliability varied across the three conditions (ICBT, YAS-CC, YAS-UC); all items fell within the “fair” to “excellent” range except for three items: Recruits Others, Functional Analysis, and Behavioral Activation. The interrater reliability for the Recruits Others item was acceptable except in the ICBT condition (ICC = .30). This likely occurred because Recruits Others had the smallest variance of all the conditions in the ICBT Condition (ICBT $\sigma^2 = 0.07$, YAS-CC $\sigma^2 = 0.19$, YAS-UC $\sigma^2 = 0.42$). Two items (Functional Analysis, Behavioral Activation) fell in the “poor” range, with ICCs below .40. An examination of the descriptive properties of the Functional Analysis ($M = 1.07, SD = 0.25$) and the Behavioral Activation ($M = 1.00, SD = 0.06$) items indicated that the low ICCs (ICCs = .30 and 0, respectively) were likely due to very small variation in the full sample of scores ($\sigma^2 = .06$ and $\sigma^2 = .004$, respectively). Given that these two items had small variation in the full sample (not just one condition like the Recruits Others item), they were
dropped from the TPOCS-RS and from all subsequent analyses. However removing these items did not influence the subscale ICCs (ICC = .94), likely due to low frequency.

Next, Cronbach’s alpha, or internal consistency, was calculated for each subscale (Cronbach, 1951). Following guidelines, alpha’s below .5 reflect “unacceptable” agreement, alphas between .5 and .6 reflect “poor” agreement, alphas from .6 to .7 reflect “questionable” agreement, alphas from .7 to .8 reflect “acceptable” agreement, alphas from .8 to .9 reflect “good” agreement, and alphas above .9 reflect “excellent agreement” (George & Mallory, 2003). For the full sample, the internal consistency of each subscale was above .59. All subscales fell in the “acceptable” agreement range except the Client-Centered subscale. On the Client-Centered subscale, the alpha fell in the upper limits of the “poor” agreement range (α = .59), suggesting that items on this subscale did not relate to each other as well as they did on other subscales. This low item agreement should be taken into consideration in the interpretation of the Client-centered subscale findings. Overall, although the Client-Centered subscale fell in the “poor” agreement range, the subscales reflected what has been deemed acceptable agreement ($M = .78$, $SD = .10$).

**Evaluation of sample representation.** Analyses were run to determine that the sample drawn for the current study did not significantly differ from either of the parent studies in regards to demographic, clinical, youth, and therapist characteristics. This test ensured that the current sample represented the parent sample, so we could generalize the findings to that sample. $T$-tests were conducted for continuous measurements and chi-square tests were conducted for categorical measurements. Overall, the final sample was representative of both of the parent samples. In particular, client’s age, gender, ethnicity, annual family income, primary diagnoses, comorbidity, and CBCL Total, Internalizing, and Externalizing $T$ scores in the final sample did
not significantly differ from the corresponding conditions in the parent samples. In the Kendall et al. study, therapist’s gender and ethnicity were not statistically different from the parent sample. In the YAS study, therapist’s age, gender, ethnicity, professional degree, years of training, and years of experience after training did not significantly differ from the parent sample. Also, number of sessions held or number of weeks in treatment did not significantly differ in the final YAS sample compared to the parent sample. Unfortunately, session number and weeks in treatment information was missing for three of the four excluded participants in the Kendall study; thus, the final and parent sample could not be compared on session number and weeks in treatment data. Overall, even with the removal of 14 participants from the parent studies (4 from the Kendall et al. study, 10 from the YAS study), the current sample still adequately represented the parent sample.

**Evaluation of distribution of data.** To examine if there were confounding variables, or variables that differed across conditions that could potentially account for change in the TPOCS-RS scores (other than time, therapist, and condition), one-way ANOVA and chi-square analyses were conducted. Analyses tested whether conditions differed in regards to demographic, clinical, youth, and therapist characteristics. See Table 3 for details of continuous variables and Table 4 for categorical variables. YAS-CC, YAS-UC, and ICBT conditions did not significantly differ in client’s age, gender, total number of diagnoses, total number of anxiety diagnoses, or CBCL Total and Internalizing T scores. However, all three conditions significantly differed in client ethnicity, family income, CBCL Externalizing T scores, and total number of externalizing disorders. In addition, in the Kendall et al. study, multiple anxiety disorders could be considered primary whereas in YAS, only one anxiety disorder was considered primary. The primary diagnoses across conditions therefore could not be meaningfully compared, but likely
significantly differ due the YAS study’s inclusion of participants with primary diagnoses of Specific Phobia. In regards to therapist characteristics, therapists did not significantly differ in gender or ethnicity across conditions. Lastly, number of sessions held did not significantly differ across all three conditions, but weeks in treatment significantly differed. Post hoc tests with a Bonferroni correction indicated that the clients in YAS-CC and YAS-UC conditions were in treatment for significantly more weeks (YAS-CC M = 26.21 weeks, YAS-UC M = 26.84 weeks) than clients in the ICBT condition (M = 19.52 weeks), p = .039 and p = .015, respectively. In all, conditions were comparable across a range of demographic and clinic factors, but they differed in client ethnicity, severity and number of externalizing disorders, family income, and weeks in treatment; these factors will have to be taken into account in the interpretation of the findings.

Table 3

*Comparison of Continuous Demographic Variables Across Conditions*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Condition</th>
<th>M</th>
<th>SD</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client’s age</td>
<td>ICBT</td>
<td>10.36</td>
<td>1.90</td>
<td>1.35</td>
<td>.264</td>
</tr>
<tr>
<td></td>
<td>YAS-CC</td>
<td>11.22</td>
<td>2.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>YAS-UC</td>
<td>10.44</td>
<td>1.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBCL Internalizing</td>
<td>ICBT</td>
<td>67.40</td>
<td>8.37</td>
<td>0.10</td>
<td>.904</td>
</tr>
<tr>
<td></td>
<td>YAS-CC</td>
<td>66.38</td>
<td>8.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>YAS-UC</td>
<td>66.82</td>
<td>8.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBCL Externalizing</td>
<td>ICBT</td>
<td>52.96</td>
<td>10.08</td>
<td>5.61</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td>YAS-CC</td>
<td>60.81</td>
<td>7.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>YAS-UC</td>
<td>59.41</td>
<td>9.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBCL Total</td>
<td>ICBT</td>
<td>63.18</td>
<td>8.44</td>
<td>0.39</td>
<td>.678</td>
</tr>
<tr>
<td></td>
<td>YAS-CC</td>
<td>YAS-UC</td>
<td>ICBT</td>
<td>YAS-CC</td>
<td>YAS-UC</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>--------</td>
<td>--------</td>
<td>-------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>YAS-CC</td>
<td>64.19</td>
<td>7.34</td>
<td>3.02</td>
<td>2.59</td>
<td>2.95</td>
</tr>
<tr>
<td>YAS-UC</td>
<td>65.06</td>
<td>6.46</td>
<td>1.45</td>
<td>1.18</td>
<td>1.16</td>
</tr>
<tr>
<td>Total Number of Diagnoses</td>
<td>ICBT</td>
<td>YAS-CC</td>
<td>YAS-UC</td>
<td>ICBT</td>
<td>YAS-CC</td>
</tr>
<tr>
<td></td>
<td>3.02</td>
<td>2.59</td>
<td>2.95</td>
<td>2.49</td>
<td>2.00</td>
</tr>
<tr>
<td>Total Number of Anxiety Diagnoses</td>
<td>ICBT</td>
<td>YAS-CC</td>
<td>YAS-UC</td>
<td>ICBT</td>
<td>YAS-CC</td>
</tr>
<tr>
<td></td>
<td>2.49</td>
<td>2.00</td>
<td>2.05</td>
<td>1.17</td>
<td>1.00</td>
</tr>
<tr>
<td>Total Number of Externalizing Diagnoses</td>
<td>ICBT</td>
<td>YAS-CC</td>
<td>YAS-UC</td>
<td>ICBT</td>
<td>YAS-CC</td>
</tr>
<tr>
<td></td>
<td>0.43</td>
<td>0.41</td>
<td>0.90</td>
<td>0.57</td>
<td>0.51</td>
</tr>
<tr>
<td>Number of Sessions</td>
<td>ICBT</td>
<td>YAS-CC</td>
<td>YAS-UC</td>
<td>ICBT</td>
<td>YAS-CC</td>
</tr>
<tr>
<td></td>
<td>15.92</td>
<td>16.74</td>
<td>15.71</td>
<td>1.43</td>
<td>6.20</td>
</tr>
<tr>
<td>Weeks in Treatment</td>
<td>ICBT</td>
<td>YAS-CC</td>
<td>YAS-UC</td>
<td>ICBT</td>
<td>YAS-CC</td>
</tr>
<tr>
<td></td>
<td>19.52</td>
<td>26.21</td>
<td>26.84</td>
<td>3.97</td>
<td>12.60</td>
</tr>
</tbody>
</table>

*Note.* ICBT = CBT in research setting, YAS-CC = CBT in practice setting, YAS-UC = usual care. CBCL Internalizing and CBCL Total data are reported in $T$ scores.
Table 4

Comparison of Categorical Demographic Variables Across Conditions

<table>
<thead>
<tr>
<th>Variable</th>
<th>Condition</th>
<th>%</th>
<th>Chi-square</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client’s gender</td>
<td>ICBT</td>
<td>39.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(% Female)</td>
<td>YAS-CC</td>
<td>68.4</td>
<td>4.74</td>
<td>2</td>
<td>.094</td>
</tr>
<tr>
<td></td>
<td>YAS-UC</td>
<td>47.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client’s ethnicity</td>
<td>ICBT</td>
<td>86.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(% Caucasian)</td>
<td>YAS-CC</td>
<td>36.8</td>
<td>35.93</td>
<td>10</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>YAS-UC</td>
<td>33.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client’s family income</td>
<td>ICBT</td>
<td>56.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(% Above 60k per year)</td>
<td>YAS-CC</td>
<td>15.8</td>
<td>16.97</td>
<td>2</td>
<td>&lt; .001</td>
</tr>
<tr>
<td></td>
<td>YAS-UC</td>
<td>14.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Therapist’s gender</td>
<td>ICBT</td>
<td>75.0</td>
<td>.007</td>
<td>2</td>
<td>.996</td>
</tr>
<tr>
<td>(% Female)</td>
<td>YAS-CC</td>
<td>85.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>YAS-UC</td>
<td>81.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Therapist’s ethnicity</td>
<td>ICBT</td>
<td>75.0</td>
<td>9.65</td>
<td>6</td>
<td>.140</td>
</tr>
<tr>
<td>(% Caucasian)</td>
<td>YAS-CC</td>
<td>50.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>YAS-UC</td>
<td>43.8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. ICBT = CBT in research setting, YAS-CC = CBT in practice setting, YAS-UC= usual care.
Patterns of missingness. It was important to assess the pattern of missingness in the therapy sessions under analysis in order to rule out the possibility that missing session data rather than condition accounted for the findings (i.e., predicted TPOCS-RS scores). That is, of all of the sessions held throughout the course of the study, the percent of sessions missing from the analyses (e.g., never recorded, uncodable) was calculated. In the full sample, clients averaged 32.71% of sessions missing ($SD = 18.51\%$, $Range = 0.89$). The distribution of the percent of missing data fell below what some researchers consider moderate contamination (i.e., a high degree of heterogeneity in data, $< .92$; Blanca, Arnau, Lopez-Montiel, Bono, & Bendayan, 2013) and there were three outliers that were more than 2.5 standard deviations away from the mean (Van Selst & Jolicoer, 1999), with above 80% of sessions missing. Due to the descriptive nature of the study, all clients were included in analyses if they had two or more sessions; thus, the three clients who were missing above 80% of recorded sessions were retained in the sample. In general, conditions were missing a similar amount of session data. Clients in the ICBT condition averaged 34.49% of sessions missing ($SD = 17.44\%$, $Range = 0.75$), clients in YAS-CC averaged 26.91% of sessions missing ($SD = 15.88\%$, $Range = 0.75$), and clients in YAS-UC averaged 33.63% of sessions missing ($SD = 22.68\%$, $Range = .89$). Given that all conditions had similar amounts of missing session data, it is likely that missingness did not account for the effects of condition on TPOCS-RS scores.

It was also important to assess whether certain demographic characteristics predicted missingness, because if certain client or clinical characteristics accounted for missing session data, the data would not fully represent the sample it was designed to represent and could account for the findings (i.e., TPOCS-RS scores). Thus, linear regression and one-way ANOVAs were conducted to test if demographic characteristics predicted percent of sessions missing in the
full sample and each condition. No demographics predicted missingness in any analysis. In the full sample and in each condition, client’s age, gender, ethnicity, family income, number of anxiety disorders, number of externalizing disorders, and CBCL Total T scores did not significantly predict missingness. Thus, using Rubin’s (1976) taxonomy of missingness, we can at least conclude that the data are missing at random (MAR); the probability of missing data for a client is not a function of any of the aforementioned characteristics. Taken together, client and clinical factors were not associated with missing session data, so it is likely that missingness did not influence the pattern in TPOCS-RS findings.

The percent coded for each client in each phase of treatment (5 phases) was computed to ascertain that all the therapy components were sufficiently sampled (e.g., enough exposure sessions were coded). In the full sample, the percent coded for each client in each phase of treatment ranged from an average of 67.21 to 78.39 percent coded. Paired Samples T tests were conducted and indicated that there were no significant differences in the percent of sessions coded across all five phases. Therefore, we can assume that each stage of CBT treatment (i.e., engagement, skill building, exposures) is equally represented in the coded tapes, and the number of sessions coded does not account for the change in TPOCS-RS scores over time.

**Primary analyses**

The main analyses in the current study included decisions about how to score the treatment domain subscales, select the best time in treatment variable, identify and deal with outliers, assess for demographic, client, and therapist predictors, as well as run and interpret the hierarchical linear modeling. Therefore, arguments for each of these steps in the data analytic strategy are detailed below.
**Subscale scoring.** Subscale scores were calculated by taking the maximum extensiveness score of an intervention delivered within each subscale. In other words, each treatment domain subscale score equaled the highest extensiveness rating of one of the intervention items within that subscale. This scoring strategy was adopted because this study aimed to benchmark the delivery of CBT; since therapists delivering CBT are not expected to deliver multiple interventions in a session, this strategy more fairly represented the dosage delivered in each session. Other subscale scoring strategies would not have depicted the optimal intervention extensiveness as effectively. Specifically, subscale scoring strategies such as taking the average of items in each subscale would have penalized therapists who delivered high doses of few intervention items within a subscale while rewarding therapists who delivered low doses of numerous intervention items within a subscale. For example, if the current study had taken the average of each subscale: a session that received an extensiveness rating of 5 on *Cognitive Education*, but 1s on *Cognitive Distortion* and *Coping Skills* would have received the same subscale score (i.e., an extensiveness score of 2.33 on the Cognitive subscale) as a session that received extensiveness ratings of 3 on *Cognitive Education*, and 2s on *Cognitive Distortion* and *Coping Skills*. The current study took the maximum extensiveness score in each treatment domain subscale so that the sessions containing fewer, more in-depth interventions would be adequately represented; using the current study’s scoring strategy on the previous example, the first session would have received a Cognitive subscale score of 5 compared to 2.33. Taking the maximum extensiveness score within each subscale therefore allowed the current study to analyze the full depth of interventions delivered in each treatment domain subscale, suitable for the descriptive aims of the study. In addition, therapists are rarely expected to deliver more than one or two interventions in a single session; manual-guided treatments tend to focus on a limited
number of interventions in a session, so this approach is more consistent with how therapy is typically delivered. Since the current study aimed to measure treatment differentiation, this scoring approach depicted a more accurate picture of the extensiveness of proscribed interventions delivered.

It is important to note that extensiveness ratings of subscale focus items (e.g., *Behavioral Focus*) were excluded from subscale calculations. These items were excluded because they did not represent actual, specific interventions (e.g., interventions defined in treatment protocols). The current study aimed to describe the delivery extensiveness of well-known interventions and potentially create extensiveness benchmarks for future research. Therefore, excluding subscale focus items allowed the subscale scores to be more consistent with protocols and more applicable to future studies. In addition, subscale ICCs without the subscale focus items ($M = 0.89, SD = 0.08$) were marginally higher than subscale ICCs with the subscale focus items ($M = 0.85, SD = 0.12$). Overall, the subscale focus items provided little unique information that would add to current study conclusions or benefit future researchers.

**Modeling time.** Upon examination of the distribution of sessions over time, across individual clients and conditions, there was great variability. Specifically, sessions were not held in regular intervals, and sessions in the two YAS conditions were held less frequently, over a significantly greater amount of time. Therefore, although all three conditions had similar means of sessions held, the range in sessions held as well as the mean and range in duration of treatment in weeks differed considerably. In the full sample, clients averaged 16.04 sessions ($SD = 5.32$, $Range = 2 – 52$), with 22.61 weeks in treatment ($SD = 10.31$, $Range = 1.14 – 68.71$). Clients in the ICBT Condition averaged 15.92 sessions ($SD = 1.43$, $Range = 8 – 18$), with 19.52 weeks in treatment ($SD = 3.97$, $Range = 8.57 – 30$). Clients in the YAS-CC and YAS-UC Conditions, on
the other hand, averaged 16.74 (SD = 6.20, Range = 2 - 28) and 15.71 sessions (SD = 9.34, Range = 5 – 52), with a total of 26.21 (SD = 12.60, Range = 1.14 – 48.43) and 26.84 weeks (SD = 15.53, Range = 5 – 68.71) in treatment, respectively. In the full sample, the number of total sessions held demonstrated a level of skewness (statistic = 3.11) and kurtosis (statistic = 23.39) that some researchers consider to be extreme contamination (i.e., highly heterogeneous), even in a small sample (Blanca et al., 2013). Duration of treatment in weeks, on the other hand, was slightly less skewed (statistic = 1.55) and kurtotic (statistic = 4.58), still falling in the high and extreme range of contamination, respectively (Blanca et al., 2013). The number of sessions held demonstrated extreme contamination within each study condition (e.g., YAS-UC kurtosis statistic = 12.07) whereas the duration of treatment in weeks in each condition was slightly less contaminated (e.g., YAS-UC kurtosis statistic = 1.34).

I elected to examine the trajectory of intervention delivery by using the weeks in treatment variable. This variable provided the best representation of the course of treatment for a number of reasons. For one, in comparison to days in treatment, weeks in treatment were easier to interpret since sessions were expected to be held weekly. Days in treatment would have been more difficult for readers to conceptualize. Second, in comparison to session number, the weeks in treatment variable was more statistically appropriate. The “number of sessions held” variable was more skewed and kurtotic, and thus more contaminated than duration of weeks in treatment in each condition (Blanca et al., 2013). In addition, using weeks in treatment was a better conceptual fit for the three conditions under analysis. Although session number is potentially valuable for understanding the content of the sessions (i.e., interventions delivered as prescribed by the protocol) in the ICBT condition, I had concerns that session number did not map onto session content in the YAS-CC or YAS-UC conditions. Assessing weeks in treatment was
therefore equally applicable and meaningful in all three conditions. Finally, some clients in the YAS conditions were in treatment for over a year. The lapse in time could influence the types of interventions the therapists chose to deliver, and could therefore be important in the interpretation of extensiveness data. Thus, using weeks in treatment was the optimal way to analyze change in intervention delivery over time.

**Evaluation of outliers.** Prior to further analysis, the continuous independent and dependent variables were assessed for univariate outliers. First, the “weeks in treatment” variable was assessed. A conservative criteria was used for identifying outliers; sessions 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). At the individual client level, there were 2 clients above this cut off criteria that were in treatment for more than 51 weeks total (65 sessions altogether). At the session level, there were 36 sessions above the outlier cut off point; these clients were in treatment for more than 39 weeks. Of these 36 session outliers, 25 had been coded. Given the current study’s small sample size and the importance of analyzing as many sessions as possible, outliers were excluded at the session rather than individual client level to retain as many recordings in the sample as possible. Therefore, 25 rather than 65 sessions were excluded and 952 session recordings were retained in the sample. This adjusted the skewness and kurtosis so that the data were no longer extremely contaminated; after outlier removal, skewness (statistic = 0.85) fell within what has been considered a moderate positive range (i.e., 0.76 – 1.25) and kurtosis (statistic = 0.58) fell within what has been considered a slight positive range (i.e., 0.26 - 0.75; Blanca et al., 2013).
Once these recordings were removed, the TPOCS-RS subscale scores were assessed for univariate outliers at the session level using the same conservative criteria (i.e., > 2.5 standard deviations away from the mean). Only the Psychodynamic subscale indicated outliers; 37 sessions had scores on the Psychodynamic subscale that fell more than 2.5 standard deviations above the mean ($M = 1.45$, $SD = 0.77$). These sessions were checked to ensure that all fell within the acceptable extensiveness rating scale (i.e., there were no data entry errors). Since the mean of the Psychodynamic subscale was notably low, outliers consisted of sessions in which psychodynamic interventions were delivered on average to a “considerable” degree (TPOCS-RS extensiveness ranging from 3.5 – 6.0). These outliers were appropriately unique and therefore integral to the illustrative purposes of the current study. They were thus retained in the sample and will need to be explored in the interpretation of the findings.

**Assessment of client and therapist characteristic covariates.** In general, it is important to examine client, therapist, and context characteristics that may account for significant variation in criterion scores, and thus may have to be controlled or included as random effects in the hierarchical linear model building (Singer & Willet, 2003). ANOVA and chi square analyses were therefore conducted to guard against confounding variable effects by first examining whether demographic variables differed across conditions at the session level. The following client and clinical predictors were separately entered into the ANOVA: client age, CBCL Total $T$ scores, CBCL Internalizing $T$ scores, CBCL Externalizing $T$ scores, total number of diagnoses, total number of anxiety diagnoses, total number of externalizing diagnoses, sessions held, and weeks in treatment. The following client and clinical predictors were separately entered into the chi square analyses: gender, ethnicity, and family income. The following therapist predictors were entered into chi square analyses: therapist ethnicity and gender. At the session level, all
demographic variables except CBCL Internalizing $T$ scores and therapist gender significantly differed across conditions ($p < .001$). CBCL Internalizing $T$ scores and therapist gender were therefore the only predictors entered into a linear regression and ANOVA, respectively, to test if they significantly influenced TPOCS-RS subscale scores. CBCL Internalizing scores predicted TPOCS-RS scores only on the Psychodynamic subscale, such that the higher the CBCL score, the less psychodynamic interventions delivered ($t(866) = -4.24, p < .001$). Therapist gender significantly predicted scores on the Coping Cat, Behavioral, Psychodynamic, and Client-centered subscales. Specifically, males delivered psychodynamic interventions significantly more extensively ($M = 1.73, SD = .99$) than females ($M = 1.41, SD = .72; p < .001$) and females delivered Coping Cat ($M = 4.72, SD = 2.02$), behavioral ($M = 3.79, SD = 2.01$), and client-centered interventions ($M = 4.41, SD = 1.43$) significantly more extensively than males (Coping Cat $M = 4.26, SD = 2.15$; behavioral $M = 3.37, SD = 2.06$; client-centered $M = 3.68, SD = 1.17; p = .020, p = .032, p < .001$, respectively). However, given that there were many more female ($N = 798$) than male therapists ($N = 119$) represented in the sessions, there were problems with the homogeneity of variance in the data. In the Psychodynamic, Behavioral, and Client-centered subscales, the Levene statistic was significant at $p < .002$, suggesting that the comparison findings may have been contaminated by unequal variances. Nevertheless, CBCL Internalizing scores and therapist gender were entered in the primary analyses (HLM models) and will be discussed in more detail below. In general, most client and therapist characteristics could not be teased apart from the condition or variance discrepancies and will therefore have to be considered in the interpretation of the findings.

**Recentering predictors.** Before running the HLM analyses, condition was recentered. In general, the main reason for recentering is that it simplifies the interpretation of the findings (see
For continuous temporal variables, recentering refers to subtracting a constant from each observed value. Thus, if the constant selected represents the first timepoint of data collection (i.e., session one), the intercept in the level 1 model indicates the true value of Y (i.e., the TPOCS-RS extensiveness score), a client’s initial status. In the current study, weeks in treatment did not need to be centered because the first session occurred at week “0,” the first timepoint of data collection. For dichotomous time-invariant predictors, recentering the numbers to include “0” also improves interpretability. In the current study, condition had been arbitrarily numbered “1,” “2,” and “3,” before building the models, so the conditions were renumbered as ICBT Condition “0.5,” YAS-CC condition “0,” and YAS-UC condition “0.5,” to improve the statistical control and interpretation of higher level intercepts.

**Overview of hierarchical linear modeling analyses.** It was hypothesized that there are differences between the type and extensiveness of therapeutic interventions delivered by therapists in research and practice settings. Specifically, it was predicted that therapists in research settings deliver therapeutic interventions that adhere more to the protocol and implicitly violate the treatment protocol less than therapists in practice settings. No a priori prediction was made about the change in therapeutic intervention delivery over time since, to my knowledge, this has not been explored in previous studies and was therefore exploratory. The current study used hierarchical linear modeling (HLM) to compare the means of observer-rated scores of each treatment domain across the three conditions (ICBT, YAS-CC, and YAS-UC) over time (week-by-week). This method works best in accommodating missing data and does not necessitate that data collection points be the same for each participant (Singer & Willet, 2003). Moreover, this data analytic strategy enabled this study to test specific hypotheses 1, 2, 3, and 4 (described below) using models that followed a multi-level framework.
The means and trajectories of TPOCS-RS scores over time were examined using MIXED procedure in SPSS version 21 (IBM Corporation, 2012), with Maximum Likelihood (ML) estimation and an unstructured covariance structure, starting with an unconditional means model. ML estimation is advantageous for sample sizes above 30 and is a highly preferred technique in statistical estimation (Singer & Willett, 2003). Specifically, researchers have noted that ML estimates are superior to other estimation methods because they converge on the unknown values of population parameters, have approximately normal sample distributions with a set variance, and have relatively small standard errors in comparison to other estimation methods (Singer & Willett, 2003). ML focuses on the entire model (fixed and random effects), rather than just one piece of the model as in REML estimation, and therefore is suitable for the descriptive nature of the current study. The unstructured covariance structure was used because it is most frequently found in longitudinal data and often provides the best fit since it requires no assumption in the error structure (Shek & Ma, 2011). The unstructured covariance structure model assumes that variance between timepoints (i.e., “weeks in treatment”) is not constant and that the correlations between timepoints are discrepant across time.

**Construction of hierarchical linear models.** Each hypothesis was approached in a consistent way when fitting the data. First, the trajectory of each subscale was displayed using GRAPHDATASET in SPSS based on weeks in treatment, to enhance the interpretability of the models (see Figures 2 - 7). Second, an unconditional means model was fit to the data using an identity matrix structure (due to estimation of a single level). The unconditional means model lacks a slope parameter and therefore characterizes and partitions the outcome variation at each mean level. In this model, if a between-persons variance component is 0, there is little reason to predict outcome variation at that level and the model may not warrant further extension. The data
had three levels and I tested whether there was significant variance across the levels. First, the intraclass correlation coefficient (ICC) was computed for each subscale to assess proportion of total outcome variation that fell “between” people. If the subscale data demonstrated what some researchers have identified as sufficient between person variance (ICC > .25; Singer & Willet, 2003), the data were subsequently fit to an unconditional growth curve model using ML estimation and an unstructured covariance structure.

The unconditional growth curve model extends the model by including an assessment of the effects of time. The information criteria: -2Restricted Log Likelihood Deviance (-2LL), Akaike Information Criteria (AIC), and Bayesian Information Criteria (BIC), offer an estimate of information excluded from the model in a way that has been considered conservative (Singer & Willet, 2003). Models are preferred if they have lesser information criteria values than previous models. The chi square equation (difference between -2LL in subsequent models) indicates whether or not the subsequent model has a significantly better fit. Thus, for each subsequent extension of the model, the -2LL and degrees of freedom were compared to (i.e., subtracted from) the -2LL and degrees of freedom of the previous model. If the graph of the subscale supported the model choice and the chi square equation from the previous model was significant ($p < .05$), the extended model indicated incremental improvement from the previous model and the model extension was considered for optimal fit. In addition, if the variance components were significant ($p < .05$), subsequent model extensions were oftentimes tested because that indicated there was significant variation remaining in the data that was unexplained.

After the unconditional growth curve model was surpassed (due to meeting the aforementioned criteria), models progressed through the following stages: a linear growth curve model with intercept predictors, a linear growth curve model with intercept and slope predictors,
a quadratic growth curve model with intercept and slope predictors, a cubic growth curve model with intercept and slope predictors. In addition, conditions were dummy coded and entered into the HLM models to probe the overall mean contrasts between conditions within each subscale. Once two-level best-fit models were chosen for each subscale and (if) the model progressed to the unconditional growth curve model, two commonly used covariance structures (i.e., compound symmetric and first-order autoregressive) were substituted for the unstructured covariance structure to assess how the unstructured model compared to the other structures in regards to model fit. Compound symmetric and first-order autoregressive covariance structures are commonly used in HLM and a number of researchers have suggested comparing these three covariance structures (Shek & Ma, 2011; Singer & Willet, 2003). As expected, the unstructured covariance structure produced the lowest information criteria of all the covariance structures in each best-fit model, indicating that it was the most suitable covariance structure for the current study. The unstructured covariance structure was thus used throughout HLM model building. See Tables 5 – 16 for complete information about the HLM results in each subscale.

HLM construction also included an assessment of therapist main effects in intervention delivery. In the current study, therapist main effects refer to the potential dissimilarities among numerous therapists in regards to the mean scores for delivering certain types or extensiveness of therapeutic interventions (Crits-Cristoph & Mintz, 1991). After selecting the two-level model of best fit (described above), therapist was entered as a third level in the hierarchical model, between client and condition, to assess if clients were nested in therapists who were nested in condition. In particular, two criteria were used to assess whether hierarchical model construction of each subscale should be extended to include three levels. First, ICCs of the three-level model were calculated and compared to the ICCs of both the two-level best-fit and unconditional
models. If the ICC of the two-level best-fit model fell at or below 0.25, the recommended level for terminating model extension (Singer & Willet, 2003), there was little variation left to be explained in the HLM model so progressing to a three-level model (with therapist as a level) provided little added value. Following this criterion, models of the Coping Cat, Cognitive, Behavioral, and Family subscales were not extended.

Second, if the ICC of the two-level best-fit model fell above .25, the $Pseudo R^2$, an estimate of proportional reduction in level-1 variance components, was assessed relative to the best-fit and unconditional models. $Pseudo R^2$ is the equivalent of $R^2 \Delta$ in OLS regression, so Cohen’s (1988) guidelines for $R^2 \Delta$ were followed: .02 represented a small effect size, .13 represented a medium effect size, and .26 represented a large effect size. On the two subscales that had ICCs above .25 on the best-fit model, the Psychodynamic and Client-centered subscales, the three-level model did not significantly reduce the level 1 variance ($Pseudo R^2$s < .008) or level 2 variance (ICCs remained the same). In fact, the $Pseudo-R^2$s for all subscales when comparing the three-level model to best-fit models ranged from non-significant to small effects ($Pseudo R^2$s = 0 - .039). Thus, therapist did not appear to be a primary, systematic source of variance, did not greatly influence effect size, and was not included as a random effect in model construction (Kazdin, 2003).

**Hypothesis 1.** It was hypothesized that therapists in the CBT practice setting condition (YAS-CC) would deliver more proscribed therapeutic interventions (i.e., higher extensiveness of interventions that fall on the TPOCS-RS Family and Psychodynamic subscales) than therapists in the research setting condition (ICBT). If this hypothesis was supported, there would be a significant main effect of condition such that the mean extensiveness of interventions delivered in the separate TPOCS-RS Family and Psychodynamic subscales would be significantly higher
for the YAS-CC than the ICBT condition.

**Hypothesis 2.** It was hypothesized that therapists in the usual care practice setting condition (YAS-UC) would deliver significantly more family and psychodynamic therapeutic interventions than both the YAS-CC and ICBT conditions. If this hypothesis was supported, there would be a significant main effect of condition such that the mean extensiveness of interventions delivered in the TPOCS-RS Family and Psychodynamic subscales would be significantly higher for YAS-UC than both YAS-CC and ICBT.

**Evaluation of hypotheses 1 & 2.** To test the aforementioned hypotheses, a model for the Family subscale was constructed first. Although the unconditional means model suggested that both the variance within (\(\sigma^2 = 2.94\)) and between individuals (\(\sigma^2 = 0.72\)) were significant (\(p < .001\)), the ICC did not surpass the recommended 0.25 (ICC = .19). The graph of the full sample’s family extensiveness scores corroborated this finding, suggesting that there was a large amount of variance in extensiveness scores, with no singular line capturing the data. In each condition, the graph of the family extensiveness scores suggested that therapists in ICBT and UC conditions decreased their delivery over time whereas in YAS-CC, therapists increased their delivery over time (see Figure 2 below). However, there was no statistically significant change in the delivery of family interventions over time. Thus, there was low likelihood that building the model further (to include growth models) would enhance the analyses. Condition was entered into the unconditional means model to assess for mean differences in the delivery of family interventions across conditions. The fixed effects indicated that intercept and condition significantly positively predicted TPOCS-RS scores on the Family subscale, \(t(81.65) = 35.32, p < .001\) and \(t(87.12) = 9.87, p < .001\), respectively. Thus, the higher the assigned condition (i.e., YAS-UC was highest, followed by YAS-CC, and then ICBT), the more overall mean extensiveness of family
interventions delivered (See Table 5 for more details).

Condition was dummy coded and entered into the HLM model to further probe the mean comparison on the Family subscale (See Table 6). Specifically, condition contrasts (with ICBT and YAS-UC as reference groups) were entered into the unconditional means model in place of the condition variable. As predicted, therapists in the YAS-CC condition delivered significantly more family interventions than therapists in the ICBT condition ($t(63.64) = 2.68, p = .009$), and therapists in the YAS-UC condition delivered significantly more family interventions than the other two conditions (vs. ICBT $t(85.40) = 10.85, p < .001$ and vs. YAS-CC $t(72.07) = 7.01, p < .001$). Therapists in the YAS-UC condition delivered family interventions most extensively ($M = 3.88, SD = 1.95$) followed by YAS-CC therapists ($M = 2.43, SD = 1.76$), with ICBT therapists delivering the least extensive family interventions ($M = 2.00, SD = 1.68$).

![Figure 2. TPOCS-RS Family extensiveness scores by condition. ICBT = CBT in research setting, YAS-CC = CBT in practice setting, YAS-UC= usual care](image-url)
Next, the Psychodynamic subscale model was constructed. On this subscale, the linear growth curve model with condition entered as an intercept predictor appeared to be the best fitting model for a number of reasons. For one, the graphs of the psychodynamic extensiveness ratings in each condition all depicted a line that was linear (positive or negative) over sessions (See Figure 3). Second, the chi square equation of the subsequent model (i.e., a linear growth curve model with intercept and slope predictors) was not significant, indicating that extending the model further would provide little additional information. Finally, the variance in slopes was not significant at this level, suggesting that there was little variation left to be explained in the slope of time (See Table 7).

The fixed effects of the linear growth curve model with intercepts indicated that intercept, weeks in treatment, and condition were significant positive predictors of the extensiveness of psychodynamic scores ($t(73.61) = 27.04, p < .001, t(78.85) = 2.42, p = .018$, and $t(77.34) = 7.58, p < .001$, respectively). In other words, the more time in treatment, the more psychodynamic interventions delivered across the full sample. The higher the assigned condition (i.e., YAS-UC), the more overall psychodynamic interventions delivered. However, the interaction between condition and time in treatment was not significant; the conditions differed in overall delivery of psychodynamic interventions, but not in their rate of change in psychodynamic delivery.

Condition was dummy coded and entered into the linear growth curve model to compare intervention extensiveness across conditions (See Table 8). Consistent with hypothesis 1, therapists in the YAS-CC condition delivered psychodynamic interventions to a more extensive degree ($M = 1.56, SD = 0.83$) than therapists in the ICBT condition, ($M = 1.23, SD = 0.54$; $t(72.86) = 2.96, p = .004$). In addition, consistent with hypothesis 2, therapists in the YAS-UC condition delivered a higher mean extensiveness of psychodynamic interventions ($M = 1.94, SD$
than therapists in both the ICBT and YAS-CC conditions ($t(77.91) = 7.61, p < .001$; $t(77.01) = 3.77, p < .001$, respectively). The line on the graph of the YAS-UC condition showed considerable variability across time, but a higher overall extensiveness than either of the other two conditions.

![Graph showing TPOCS-RS Psychodynamic extensiveness scores by condition](image)

**Figure 3.** TPOCS-RS Psychodynamic extensiveness scores by condition. ICBT = CBT in research setting, YAS-CC = CBT in practice setting, YAS-UC = usual care.

Taken together, the HLM models of both the Family and Psychodynamic subscales provided sufficient evidence supporting hypotheses 1 and 2. As predicted, therapists in the ICBT condition delivered the least therapeutic interventions proscribed by the Coping Cat program (significantly less than therapists in the YAS-CC condition) whereas therapists in the YAS-UC condition delivered the most therapeutic interventions proscribed by the Coping Cat program.
### Table 5

Results of Hierarchical Model of Change for the TPOCS-RS Family Subscale

<table>
<thead>
<tr>
<th></th>
<th>Unconditional Means Model</th>
<th>Means Model w/Condition</th>
<th>Unconditional Growth Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed Effects (t-statistic)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>23.72**</td>
<td>35.32**</td>
<td>17.32**</td>
</tr>
<tr>
<td>Slope</td>
<td>Time</td>
<td></td>
<td>-.45</td>
</tr>
<tr>
<td>Condition</td>
<td></td>
<td>9.87**</td>
<td></td>
</tr>
<tr>
<td><strong>Variance Components (Wald-statistic)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Person Variance</td>
<td>20.68**</td>
<td>20.75**</td>
<td>20.26**</td>
</tr>
<tr>
<td>Variance in Level 1 Initial Status</td>
<td>4.48**</td>
<td>2.55*</td>
<td>3.20**</td>
</tr>
<tr>
<td>Variance in Level 2 Slope</td>
<td></td>
<td></td>
<td>-1.11</td>
</tr>
<tr>
<td>Covariance of Level 2 Intercept and Slope</td>
<td></td>
<td></td>
<td>1.01</td>
</tr>
<tr>
<td><strong>Goodness of Fit Statistics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
<td>3839.07</td>
<td>3770.48</td>
<td>3837.35</td>
</tr>
</tbody>
</table>
AIC | 3845.07 | 3778.48 | 3849.35  
BIC | 3859.65 | 3797.91 | 3878.50  
Pseudo R² | 0 | .01  

*Note.** = \( p < .001 \), * = \( p < .05 \)

**Table 6**

*Results of Condition Contrasts in the Means Model of the TPOCS-RS Family Subscale*

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope YAS-UC vs. ICBT</td>
<td>85.40</td>
<td>10.85**</td>
</tr>
<tr>
<td>YAS-CC vs. ICBT</td>
<td>63.63</td>
<td>2.68*</td>
</tr>
<tr>
<td>YAS-UC vs. YAS-CC</td>
<td>72.07</td>
<td>7.01**</td>
</tr>
</tbody>
</table>

*Note.** = \( p < .001 \), * = \( p < .05 \)
Table 7

*Results of Hierarchical Model of Change for the TPOCS-RS Psychodynamic Subscale*

<table>
<thead>
<tr>
<th>Model</th>
<th>Unconditional Means Model</th>
<th>Unconditional Growth Model</th>
<th><strong>Linear Growth Model</strong></th>
<th>Linear Growth Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictors</td>
<td>Intercept</td>
<td>Slope</td>
<td>Intercept</td>
<td>Slope</td>
</tr>
<tr>
<td>Fixed Effects (t-statistic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>24.91**</td>
<td>20.54**</td>
<td>27.04**</td>
<td>26.69**</td>
</tr>
<tr>
<td>Slope</td>
<td></td>
<td>2.36*</td>
<td>2.42*</td>
<td>1.83</td>
</tr>
<tr>
<td>Condition</td>
<td></td>
<td></td>
<td>7.58**</td>
<td>7.07**</td>
</tr>
<tr>
<td>Time*Condition</td>
<td></td>
<td></td>
<td></td>
<td>-1.43</td>
</tr>
<tr>
<td>Variance Components (Wald-statistic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Person Variance</td>
<td>20.63**</td>
<td>19.82**</td>
<td>19.78**</td>
<td>19.79**</td>
</tr>
<tr>
<td>Variance in Level 1 Initial Status</td>
<td>5.57**</td>
<td>4.16**</td>
<td>3.10*</td>
<td>3.16*</td>
</tr>
<tr>
<td>Variance in Level 2 Slope</td>
<td>-1.01</td>
<td>-0.57</td>
<td>-0.67</td>
<td></td>
</tr>
<tr>
<td>Covariance of Level 2 Intercept and Slope</td>
<td>2.93*</td>
<td>2.95</td>
<td>2.88*</td>
<td></td>
</tr>
</tbody>
</table>

Goodness of Fit Statistics

78
<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YAS-UC vs. ICBT</td>
<td>77.91</td>
<td>7.61**</td>
</tr>
<tr>
<td>YAS-CC vs. ICBT</td>
<td>72.86</td>
<td>2.96*</td>
</tr>
<tr>
<td>YAS-UC vs. YAS-CC</td>
<td>77.01</td>
<td>3.77**</td>
</tr>
</tbody>
</table>

Note. ** = p < .001, * = p < .05

Table 8

Results of Condition Contrasts in the Linear Growth Curve Model of the TPOCS-RS Psychodynamic Subscale
**Hypothesis 3.** It was hypothesized that therapists in the YAS-CC condition would deliver less prescribed therapeutic interventions (i.e., lower frequency and thoroughness of interventions that fall on the TPOCS-RS CBT-Coping Cat subscale) than therapists in the ICBT condition. If this hypothesis were supported, there would be a significant main effect of condition such that the overall mean extensiveness of interventions delivered in the TPOCS-RS CBT-Coping Cat subscale would be significantly lower for YAS-CC than ICBT.

**Hypothesis 4.** It was hypothesized that therapists in the YAS-UC condition would deliver prescribed interventions significantly less extensively (i.e., lower frequency and thoroughness of interventions that fall on the TPOCS-RS CBT-Coping Cat subscale) than therapists in both the YAS-CC and ICBT conditions. This hypothesis would be supported if there was a significant main effect of condition such that mean extensiveness of interventions delivered in the TPOCS-RS CBT-Coping Cat subscale would be significantly lower for YAS-UC than both the YAS-CC and ICBT conditions.

**Evaluation of hypotheses 3 & 4.** To test hypotheses regarding prescribed therapeutic interventions, a HLM model for the Coping Cat subscale was constructed. On this subscale, the linear growth curve model with condition as an intercept and slope predictor appeared to be the best fitting model for a few reasons. First, the graph of Coping Cat extensiveness ratings in the full sample and each condition displayed linear lines, see Figure 4. For the ICBT condition, the graph demonstrated a slightly positive linear line whereas for both YAS-CC and YAS-UC conditions, the graph demonstrated a slight negative linear line (with the exception of YAS-CC displaying a burst of Coping Cat extensiveness at the end of treatment; See Figure 4). Second, although the chi-square equation between the linear and subsequent, quadratic model was significant ($p < .001$), suggesting that the model could be extended further, only the graph of
the ICBT condition displayed what could be considered a visibly negative quadratic trend (i.e., starting treatment delivering interventions at low extensiveness, increasing extensiveness as time progresses, delivering peak extensiveness mid-way through treatment, and decreasing extensiveness towards the end of treatment). Therefore, the ICBT condition may have driven the findings supporting a quadratic model fit (Singer & Willet, 2003). Given the pertinence of choosing a model that best captured the data in the entire sample (rather than one condition), the linear model was chosen. Lastly, the variance of slopes was not significant at the linear growth curve level, suggesting that there was little variation left to be explained in the change in extensiveness over time. Overall, the linear growth curve model suggested that therapists’ delivery of core Coping Cat interventions changed in a linear manner (i.e., either increased or decreased) throughout treatment (See Table 9).

The fixed effects of the linear growth curve model of the Coping Cat subscale indicated that intercept \((t(93.28) = 32.31, p < .001)\), condition \((t(102.64) = -9.79, p < .001)\), and the interaction between condition and weeks in treatment \((t(66.28) = -3.60, p = .001)\), significantly predicted TPOCS-RS Coping Cat extensiveness. In other words, condition negatively predicted Coping Cat extensiveness such that the condition assigned to the lowest ranking (i.e., ICBT) delivered the most Coping Cat interventions. Over time, conditions increasingly differed in the extensiveness of Coping Cat interventions delivered. Weeks in treatment alone did not significantly predict TPOCS-RS Coping Cat extensiveness, likely because intervention extensiveness had different slopes depending on the condition.

Next, conditions were dummy coded and contrasted against one another (with ICBT and YAS-UC as the reference groups) in the linear growth curve model of the Coping Cat subscale. As predicted in hypothesis 3, therapists in the YAS-CC condition delivered interventions
significantly less extensively within the Coping Cat subscale \( (M = 4.35, SD = 1.77) \), than therapists in the ICBT condition \( (M = 5.66, SD = 1.54; t(952) = -2.00 \ p = .046) \). In other words, therapists in the Coping Cat condition in a practice setting delivered prescribed therapeutic interventions less extensively than therapists in the Coping Cat condition of the research setting. In addition, therapists in the ICBT condition delivered prescribed therapeutic interventions increasingly more extensively over time than therapists in the YAS-CC condition, \( (t(952) = -5.15, p < .001) \). Further, as predicted in hypothesis 4, therapists in the YAS-UC condition delivered interventions significantly less extensively within the Coping Cat Subscale \( (M = 2.05, SD = 1.12) \) than therapists in both the ICBT and YAS-CC conditions \( (t(126.20) = -11.69, p < .001 \) and \( t(105.49) = -8.64, p < .001 \), respectively). Thus, therapists in UC delivered interventions prescribed by the Coping Cat program at the lowest extensiveness of all the conditions, as expected. Exploratory analyses indicated that therapists in YAS-UC condition delivered increasingly less interventions prescribed by the Coping Cat program than therapists in the ICBT condition as time progressed \( (t(731.03) = -3.53, p < .001) \), but the slope of their extensiveness delivery did not significantly differ from those delivered by therapists in the YAS-CC condition \( (t(606.26) = -1.55, p = .122; \) See Table 10).
Figure 4. TPOCS-RS Coping Cat extensiveness scores by condition. ICBT = CBT in research setting, YAS-CC = CBT in practice setting, YAS-UC= usual care.
Table 9

Results of Hierarchical Model of Change for the TPOCS-RS Coping Cat Subscale

<table>
<thead>
<tr>
<th>Model</th>
<th>Unconditional Means Model</th>
<th>Unconditional Growth Model</th>
<th>Linear Growth Model</th>
<th>Linear Growth Model</th>
<th>Quadratic Growth Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictors</td>
<td>Fixed Effects (t-statistic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intercept</td>
<td>26.04**</td>
<td>26.02**</td>
<td>30.81**</td>
<td>32.31**</td>
</tr>
<tr>
<td></td>
<td>Slope</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>1.56</td>
<td>1.52</td>
<td>0.34</td>
<td>4.48**</td>
</tr>
<tr>
<td></td>
<td>Condition</td>
<td>-17.30**</td>
<td>-9.79**</td>
<td>-7.20**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time*Condition</td>
<td>-3.60*</td>
<td>-3.44*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time*Time</td>
<td>-5.04**</td>
<td>-3.44*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time<em>Time</em>Condition</td>
<td>2.70*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance Components (Wald-statistic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Person Variance</td>
<td>20.72**</td>
<td>19.89**</td>
<td>19.80**</td>
<td>19.92**</td>
<td>20.63**</td>
</tr>
<tr>
<td>Variance in Level 1 Initial Status</td>
<td>6.08**</td>
<td>4.81**</td>
<td>3.17*</td>
<td>3.06*</td>
<td>3.78**</td>
</tr>
<tr>
<td>Variance in Level 2 Slope</td>
<td>1.16</td>
<td>-1.76</td>
<td>-1.24</td>
<td>-3.44*</td>
<td></td>
</tr>
</tbody>
</table>
Covariance of Level 2 Intercept and Slope  
2.39*  
2.08  
1.56  
c

Goodness of Fit Statistics

-2 Log Likelihood  
3587.20  
3566.22  
3437.62  
3425.81  
3397.91

AIC  
3593.20  
3578.22  
3451.62  
3441.81  
3417.91

BIC  
3607.78  
3607.37  
3485.63  
3480.68  
3466.50

Pseudo R^2 within person  
.06  
0  
0  
0  
.03

Note. Model of best fit is bolded, ** = p < .001, * = p < .05, c = could not be computed because of redundancy in the data

Table 10

Results of Condition Contrasts in the Linear Growth Curve Model of the TPOCS-RS Coping Cat Subscale

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>df</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YAS-UC vs. ICBT</td>
<td>126.20</td>
<td>-11.69**</td>
</tr>
<tr>
<td>YAS-CC vs. ICBT</td>
<td>952</td>
<td>-2.00*</td>
</tr>
<tr>
<td>YAS-UC vs. YAS-CC</td>
<td>105.49</td>
<td>-8.64**</td>
</tr>
<tr>
<td>YAS-UC vs. ICBT*TIME</td>
<td>731.03</td>
<td>-3.53**</td>
</tr>
<tr>
<td>Comparison</td>
<td>F-value</td>
<td>p-value</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>YAS-CC vs. ICBT*TIME</td>
<td>952.00</td>
<td>-5.15**</td>
</tr>
<tr>
<td>YAS-UC vs. YAS-CC*TIME</td>
<td>606.26</td>
<td>-1.55</td>
</tr>
</tbody>
</table>

*Note.* ** = $p < .001$, * = $p < .05$
**Exploratory analyses.** The means and trajectories of extensiveness scores within the Cognitive, Behavioral, and Client-centered subscales were investigated for exploratory purposes. For the Cognitive subscale, the unconditional means model was the best fit. Although the variance components were significant, the ICC was acceptable (ICC = .45), and the graph of the full sample’s cognitive extensiveness scores displayed a slightly negative linear line (see Figure 5), the chi square equation indicated that the subsequent model (the unconditional growth curve model) was not an incremental improvement in model fit (See Table 11). Thus, the model was not extended beyond the unconditional means model. Condition was entered into the unconditional means model to assess for mean differences in the delivery of cognitive interventions across conditions. Fixed effects of the unconditional means model of the Cognitive subscale indicated that intercept positively predicted TPOCS-RS scores, \( t(93.61) = 33.80, p < .001 \), and condition negatively predicted TPOCS-RS scores, \( t(95.65) = -12.25, p < .001 \). That is, the lower the assigned condition (i.e., ICBT), the more the mean cognitive interventions delivered throughout treatment.
For the Behavioral subscale, the linear growth curve model with condition as both an intercept and slope predictor appeared to be the best fit. The line on the graph of the full sample’s behavioral extensiveness scores appeared negative and quadratic. However, similar to the Coping Cat subscale, upon inspection of the graphs of each condition, the ICBT condition graph demonstrated a slightly positive linear line whereas the graphs for both YAS-CC and YAS-UC conditions demonstrated slight negative linear lines. Therefore, although the chi square equation suggested that the subsequent model, the quadratic growth curve model, was a significantly better fit ($p < .05$), the converse positive and negative linear lines in each condition over time appeared to be driving this full sample result.

The fixed effects of the linear growth curve model of the Behavioral subscale indicated that intercept and time in treatment positively predicted TPOCS-RS extensiveness ratings,
(t(80.88) = 25.66, \( p < .001 \) and \( t(50.27) = 5.58, \ p < .001 \), respectively) and condition and the interaction between condition and time in treatment negatively predicted TPOCS-RS extensiveness ratings within the Behavioral subscale (\( t(92.59) = -5.75, \ p < .001 \) and \( t(69.18) = -8.10, \ p < .001 \), respectively). For the full sample, the more time in treatment, the more extensive the behavioral interventions delivered. Further, the lower the assigned condition (i.e., ICBT condition), the more behavioral interventions delivered. Over time, the conditions increasingly differed in extensiveness levels (See Figure 6 and Table 13).

![Figure 6. TPOCS-RS Behavioral extensiveness scores by condition. ICBT = CBT in research setting, YAS-CC = CBT in practice setting, YAS-UC= usual care.](image)

Finally, for the Client-centered subscale, the unconditional means model was the best fit. The lines on the graph of the full sample’s client-centered extensiveness ratings indicated an approximately horizontal line (i.e., marginal rate of change) and this trend persisted in the graphs
of each condition (see Figure 7), suggesting that therapists in each condition delivered a consistent extensiveness of client-centered interventions throughout treatment. Moreover, the -2LL in the subsequent model (the unconditional growth curve model) was greater than the -2LL in the unconditional means model, suggesting that extending the model resulted in a poorer model fit (See Table 15). As was done with the Family and Cognitive subscales, condition was entered into the unconditional means model to evaluate mean differences in the delivery of client-centered interventions across conditions. Intercept significantly positively predicted TPOCS-RS client-centered extensiveness \( t(93.13) = 42.75, p < .001 \) whereas condition significantly negatively predicted TPOCS-RS client-centered extensiveness, \( t(94.01) = -4.54, p < .001 \). That is, the lower the assigned condition (i.e., ICBT), the more extensive the overall mean delivery of client-centered interventions.

![Graph showing TPOCS-RS Client-centered extensiveness scores by condition. ICBT = CBT in research setting, YAS-CC = CBT in practice setting, YAS-UC = usual care.](image)

*Figure 7. TPOCS-RS Client-centered extensiveness scores by condition. ICBT = CBT in research setting, YAS-CC = CBT in practice setting, YAS-UC = usual care.*
For the exploratory subscales, as with the other subscales, conditions were dummy coded and contrasted against one another (both ICBT and YAS-UC served as the reference groups) in the HLM model that was deemed the best fit (See Tables 12, 14, and 16). On the Cognitive subscale, therapists in the ICBT condition delivered significantly more therapeutic interventions ($M = 4.83, SD = 1.81$) than therapists in the YAS-CC condition ($M = 4.07, SD = 1.77; t(81.40) = 3.92, p < .001$). In addition, therapists in the YAS-UC condition delivered the least cognitive therapeutic interventions ($M = 1.85, SD = 1.00$), significantly less than both ICBT ($t(93.83) = -13.12, p < .001$) and YAS-CC conditions ($t(87.52) = 7.61, p < .001$).

On the Behavioral subscale, therapists in the ICBT condition did not differ from therapists in the YAS-CC condition in their average delivery of therapeutic interventions (ICBT $M = 4.72, SD = 1.83$ vs. YAS-CC $M = 3.11, SD = 1.58; t(1210.96) = -0.96, p = .339$). However, therapists in both the ICBT and YAS-CC conditions delivered significantly more behavioral therapeutic interventions than therapists in the YAS-UC condition (YAS-UC $M = 1.63, SD = 0.88; t(1190.73) = -6.74, p < .001$ and $t(1215.96) = 5.13, p < .001$, respectively). Interestingly, therapists in the ICBT condition delivered behavioral interventions increasingly more extensively than therapists in the YAS-CC and YAS-UC conditions ($t(1034.47) = 8.64, p < .001$ and $t(1119.55) = 9.27, p < .001$, respectively). Conversely, therapists in the YAS-UC and YAS-CC conditions did not differ in slope of behavioral extensiveness, $t(1045.33) = 0.46, p = .645$.

Finally, on the Client-centered subscale, therapists in the ICBT condition delivered interventions significantly more extensively ($M = 4.68, SD = 1.44$) than therapists in the YAS-CC ($M = 3.84, SD = 1.28$) and YAS-UC ($M = 3.74, SD = 1.12$) conditions ($t(87.63) = 3.58, p = .001$ and $t(94.89) = 4.15, p < .001$, respectively). However, therapists in the YAS-CC and YAS-UC conditions did not significantly differ in their extensiveness of client-centered therapeutic
interventions delivered, \( t(91.88) = 0.41, \ p = .680 \). Thus, therapists delivering CBT in research settings delivered more client-centered interventions than therapists in both CBT and UC conditions in practice settings. Therapists in practice settings did not differ across conditions in their delivery of client-centered interventions.
Table 11

*Results of Hierarchical Model of Change for the TPOCS-RS Cognitive Subscale*

<table>
<thead>
<tr>
<th></th>
<th>Unconditional Means Model</th>
<th>Means Model w/Condition</th>
<th>Unconditional Growth Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed Effects (t-statistic)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>25.54**</td>
<td>33.80**</td>
<td>24.80**</td>
</tr>
<tr>
<td>Slope</td>
<td></td>
<td>-1.58</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td></td>
<td>-12.25**</td>
<td></td>
</tr>
<tr>
<td>Condition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Variance Components (Wald-statistic)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Person Variance</td>
<td>20.72**</td>
<td>20.80**</td>
<td>19.81**</td>
</tr>
<tr>
<td>Variance in Level 1 Initial Status</td>
<td>5.83**</td>
<td>4.77**</td>
<td>4.49**</td>
</tr>
<tr>
<td>Variance in Level 2 Slope</td>
<td></td>
<td></td>
<td>.70</td>
</tr>
<tr>
<td>Covariance of Level 2 Intercept and Slope</td>
<td></td>
<td></td>
<td>.70</td>
</tr>
<tr>
<td><strong>Goodness of Fit Statistics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
<td>3694.91</td>
<td>3603.90</td>
<td>3687.66</td>
</tr>
</tbody>
</table>
AIC | 3700.91 | 3611.90 | 3699.66
BIC | 3715.49 | 3631.34 | 3728.81
Pseudo $R^2$ | 0 | .01

*Note.* Model of best fit in bold, ** = $p < .001$, * = $p < .05$

Table 12

*Results of Condition Contrasts in the Means Model of the TPOCS-RS Cognitive Subscale*

<table>
<thead>
<tr>
<th>Statistic</th>
<th>$df$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed Effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>YAS-UC vs. ICBT</td>
<td>93.83</td>
</tr>
<tr>
<td></td>
<td>YAS-CC vs. ICBT</td>
<td>81.40</td>
</tr>
<tr>
<td></td>
<td>YAS-UC vs. YAS-CC</td>
<td>87.52</td>
</tr>
</tbody>
</table>

*Note.* ** = $p < .001$, * = $p < .05$
Table 13

Results of Hierarchical Model of Change for the TPOCS-RS Behavioral Subscale

<table>
<thead>
<tr>
<th>Model</th>
<th>Unconditional Means Model</th>
<th>Unconditional Growth Model</th>
<th>Linear Growth Model</th>
<th>Linear Growth Model</th>
<th>Quadratic Growth Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictors</td>
<td>Intercept</td>
<td>Slope</td>
<td>Intercept</td>
<td>Slope</td>
<td></td>
</tr>
<tr>
<td>Fixed Effects (t-statistic)</td>
<td>24.33**</td>
<td>25.14**</td>
<td>19.83**</td>
<td>25.66**</td>
<td>17.91**</td>
</tr>
<tr>
<td></td>
<td>7.07**</td>
<td>6.70**</td>
<td>5.58**</td>
<td>7.97**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-16.65**</td>
<td>-5.75**</td>
<td>-2.58*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-8.10**</td>
<td>-7.69**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-6.65**</td>
<td>5.21**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance Components (Wald-statistic)</td>
<td>20.74**</td>
<td>19.86**</td>
<td>19.83**</td>
<td>20.02**</td>
<td>20.84**</td>
</tr>
<tr>
<td></td>
<td>5.77**</td>
<td>2.42*</td>
<td>2.38*</td>
<td>1.12</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td>2.79*</td>
<td>-2.64*</td>
<td>-0.45</td>
<td>.81</td>
<td></td>
</tr>
</tbody>
</table>

95
Covariance of Level 2 Intercept and Slope

<table>
<thead>
<tr>
<th></th>
<th>3.73**</th>
<th>3.62**</th>
<th>1.89</th>
<th>c</th>
</tr>
</thead>
</table>

Goodness of Fit Statistics

<table>
<thead>
<tr>
<th></th>
<th>3749.59</th>
<th>3593.63</th>
<th>3498.32</th>
<th>3450.49</th>
<th>3397.81</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 Log Likelihood</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>3755.59</td>
<td>3605.63</td>
<td>3512.32</td>
<td>3466.49</td>
<td>3417.81</td>
</tr>
<tr>
<td>BIC</td>
<td>3770.16</td>
<td>3634.78</td>
<td>3546.33</td>
<td>3505.36</td>
<td>3466.40</td>
</tr>
<tr>
<td>Pseudo R²</td>
<td>.19</td>
<td>0.01</td>
<td>0</td>
<td>.03</td>
<td></td>
</tr>
</tbody>
</table>

Note. Model of best fit is bolded, ** = p < .001, * = p < .05, c = could not be computed due to redundancy in data

Table 14

Results of Condition Contrasts in the Linear Growth Curve Model of the TPOCS-RS Behavioral Subscale

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope YAS-UC vs. ICBT</td>
<td>1190.73</td>
<td>-6.74**</td>
</tr>
<tr>
<td>YAS-CC vs. ICBT</td>
<td>1210.96</td>
<td>-.96</td>
</tr>
<tr>
<td>YAS-UC vs. YAS-CC</td>
<td>1215.96</td>
<td>-5.13**</td>
</tr>
<tr>
<td>YAS-UC vs. ICBT*TIME</td>
<td>1034.48</td>
<td>-8.64**</td>
</tr>
</tbody>
</table>
YAS-CC vs. ICBT*TIME  1119.56  -9.27**
YAS-UC vs. YAS-CC*TIME  1045.33  .46

*Note. ** = p < .001, * = p < .05

Table 15

Results of Hierarchical Model of Change for the TPOCS-RS Client-Centered Subscale

<table>
<thead>
<tr>
<th></th>
<th>Unconditional Means Model</th>
<th>Means Model w/Condition</th>
<th>Unconditional Growth Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effects (t-statistic)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>43.61**</td>
<td>42.75**</td>
<td>36.51**</td>
</tr>
<tr>
<td>Slope Time</td>
<td></td>
<td></td>
<td>-.50</td>
</tr>
<tr>
<td>Condition</td>
<td></td>
<td>-4.54**</td>
<td></td>
</tr>
<tr>
<td>Variance Components (Wald-statistic)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Person Variance</td>
<td>20.78**</td>
<td>20.77**</td>
<td>20.84**</td>
</tr>
<tr>
<td>Variance in Level 1 Initial Status</td>
<td>5.72**</td>
<td>5.47**</td>
<td>5.21**</td>
</tr>
<tr>
<td>Variance in Level 2 Slope</td>
<td></td>
<td></td>
<td>-3.11*</td>
</tr>
</tbody>
</table>
Covariance of Level 2 Intercept and Slope

Goodness of Fit Statistics

<table>
<thead>
<tr>
<th></th>
<th>-2 Log Likelihood</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st model</td>
<td>3082.75</td>
<td>3088.75</td>
<td>3103.32</td>
</tr>
<tr>
<td>2nd model</td>
<td>3064.25</td>
<td>3072.25</td>
<td>3091.68</td>
</tr>
<tr>
<td>3rd model</td>
<td>3088.34</td>
<td>3100.34</td>
<td>3129.49</td>
</tr>
</tbody>
</table>

Pseudo R\(^2\)        0 .02

Note. Model of best fit is bolded, ** = \(p < .001\), * = \(p < .05\), c = could not be computed due to redundancy in covariance parameter.

Table 16

Results of Condition Contrasts in the Means Model of the TPOCS-RS Client-centered Subscale

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YAS-UC vs. ICBT</td>
<td>94.89</td>
<td>-4.15**</td>
</tr>
<tr>
<td>YAS-CC vs. ICBT</td>
<td>87.63</td>
<td>-3.58*</td>
</tr>
<tr>
<td>YAS-UC vs. YAS-CC</td>
<td>91.88</td>
<td>.41</td>
</tr>
</tbody>
</table>

Note. ** = \(p < .001\), * = \(p < .05\)
**Therapist gender and CBCL analyses.** Given that therapist gender and client’s CBCL Internalizing $T$ scores significantly predicted TPOCS-RS scores in the preliminary ANOVA analyses, therapist gender and CBCL Internalizing scores were separately entered as fixed effects into the HLM models of best-fit to assess whether they (a) improved the model fit (i.e., significantly reduced information criterion), and (b) had a significant fixed effect on TPOCS-RS scores. In each of the best-fit models, the addition of therapist gender and CBCL internalizing scores improved the model fit. However, CBCL Internalizing scores did not have significant fixed effects in any of the best-fit models, even on the Psychodynamic subscale. Therapist gender, on the other hand, displayed a significant fixed effect in the Client-centered subscale best-fit model, such that female therapists delivered significantly more client-centered interventions than male therapists. Unlike the ANOVA analyses, therapist gender did not have a significant fixed effect in the Behavioral, Coping Cat, and Psychodynamic subscales, and similar to the ANOVA analyses, therapist gender did not have a significant fixed effect in the Family and Cognitive subscales. Overall, these findings indicate that CBCL Internalizing scores did not appear to significantly influence the TPOCS-RS findings, but therapist gender may have had an effect. Female therapists may deliver more client-centered interventions than male therapists, but these findings are difficult to interpret due to the disproportionate number of female therapists in the study (i.e., 83.7% female).

**Discussion**

Although EBTs are generally efficacious when delivered in a lab setting, when certain EBTs are transported and delivered in practice settings, they sometimes perform equal to or worse than psychosocial treatment typically delivered in practice settings (Guarton, 1992; Southam-Gerow et al., 2010; Weisz et al., 2006). EBTs may sometimes fail to perform as
expected in practice settings because practice settings differ from research settings in a variety of ways (e.g., client, therapist, contextual characteristics; Southam-Gerow et al., 2003), and researchers speculate that these differences may influence EBT implementation and client outcomes (Garland et al., 2010). However, little research has been conducted to begin to understand this issue; up until this point, researchers have not had all the information requisite to pinpoint ways to improve EBT implementation, and subsequently begin to close the science-practice performance gap (Bickman, 2000).

The current study aimed to take researchers a step toward closing the science-practice performance gap, by offering them information about the ways in which in-session processes, or implementation integrity, differ in the two settings. The study fulfilled this goal by choosing one EBT: CBT for youth anxiety. The study characterized CBT implementation in a practice setting and compared it to CBT delivery in a research setting using a treatment integrity measure that focused on therapists’ delivery of interventions (i.e., treatment adherence and treatment differentiation). In particular, the TPOCS-RS was used to identify the type of therapeutic interventions that therapists delivered in each setting, the extensiveness with which they delivered these interventions (on average and over time), and the degree to which these interventions fit or did not fit the prescribed treatment domain. Study findings demonstrate that D&I researchers can use treatment integrity measures to effectively operationalize implementation integrity (McLeod et al., 2013). The present study depicts what CBT implementation actually looks like so that future researchers can go on to (a) investigate why CBT implementation may meet or fall short of expectations in certain areas, and (b) brainstorm how CBT protocols or implementation procedures can be modified appropriately.
With these overarching study methods and purposes in sight, I will begin by reviewing support for the hypotheses and implications of the findings. Next, I will draw on the conceptual model of therapeutic change to present some speculations as to how pre-treatment contextual differences may have influenced the different delivery patterns across settings. I will then review the exploratory findings and implications and follow with a discussion of factors that strengthened and weakened the study conclusions. Finally, I will end by offering recommendations for how to further these findings in future research.

**Was There a Difference in CBT Delivery in Research and Practice Settings?**

Altogether, current study hypotheses regarding CBT implementation were supported. Therapists delivering CBT in the research setting delivered proscribed interventions (i.e., interventions considered implied violations of the CBT protocol) significantly less extensively (i.e., less frequently and thoroughly) than therapists delivering CBT in the practice setting and (2) prescribed interventions (i.e., core interventions from the CBT protocol) significantly more extensively than therapists delivering CBT in the practice setting. In addition, therapists in UC (1) delivered interventions prescribed by the CBT protocol least extensively and (2) interventions proscribed by the protocol most extensively of all the conditions.

**Was there a difference across settings in the delivery of proscribed interventions?**

The first hypothesis of the current study proposed that therapists in the CBT condition of the practice setting would deliver interventions proscribed by the Coping Cat program more extensively than therapists in the CBT condition of the research setting. This hypothesis was supported. In other words, therapists delivering CBT in the practice setting departed more from the treatment protocol than therapists delivering CBT in the research setting, consistent with past researchers’ speculations (Fixsen et al., 2005; Weisz et al., 2009). This finding holds importance
for the field of D&I research because it demonstrates that therapists delivering CBT in practice settings, in fact, may be implementing CBT in a *different way* than therapists in research settings. In the past, researchers have found that therapists in practice settings deliver a high frequency of interventions outside the protocol (e.g., Barber et al., 2006), but the degree of proscribed interventions delivered in practice settings was never compared to that of research settings. This study is novel in indicating that therapists in practice settings may be adding more outside interventions than therapists in research settings, and these delivery differences may be occurring in systematic ways.

This finding is also significant because these implementation integrity differences may, in part, contribute to the science-performance gap (McLeod et al., 2013). For example, it could be that delivering certain amounts of family or psychodynamic interventions predicts poorer client outcomes, sustaining the science-performance gap and indicating that researchers may need to alter implementation procedures. Alternatively, it could also be that the delivery of a certain level of proscribed interventions predicts better client outcomes, indicating that researchers may need to consider incorporating certain family and psychodynamic interventions into the CBT program itself (i.e., practice-based evidence; Garland et al., 2006), to adapt to practice-based needs. Given that the current study had similar client outcomes in the research and practice setting, the latter interpretation is not unfounded. Future D&I researchers should assess if and how the delivery of certain amounts of proscribed interventions influences client outcomes so that, eventually, this information could be used to alter CBT implementation methods or protocols.

In the second hypothesis, it was proposed that therapists in the control, UC condition would deliver proscribed interventions (i.e., implicit violations according to CBT protocol, not
according to UC) the most extensively of all conditions. This hypothesis was supported, and served as another way to assess CBT implementation. This finding demonstrates that although therapists delivering CBT in a practice setting may deliver more proscribed therapeutic interventions than therapists delivering CBT in a research setting, they still deliver less proscribed interventions than therapists in UC. In fact, therapists in the practice setting condition delivered proscribed interventions in a way that was more comparable to therapists in the research setting (i.e., the “gold standard” benchmark) than UC. This finding was particularly pronounced for family interventions; therapists delivering CBT in the practice setting delivered family interventions slightly more than therapists in the research setting (i.e., a small effect size), but markedly less than therapists in the UC condition (i.e., a large effect size). As such, it appears that some CBT implementation procedures effectively reduced therapists’ delivery of interventions that are proscribed by the Coping Cat program. Implementation researchers may therefore already have strategies in place to encourage therapists to refrain from delivering family and psychodynamic interventions in training, but it is unclear what these strategies are or whether they are even necessary.

Researchers could use this descriptive comparison to guide future research efforts. For example, if future research finds that the delivery of a moderate dose of proscribed interventions (i.e., the level delivered in the CBT practice setting condition) predicts positive client outcomes, they might consider incorporating certain proscribed interventions into the CBT protocol itself. On the other hand, if delivery of a moderate level of proscribed interventions is associated with poorer client outcomes, researchers could explore which implementation procedures in place already promote reduction in proscribed intervention delivery, as these strategies could then be expanded or emphasized in future CBT implementation studies. Overall, findings from
hypotheses one and two indicate that there is a difference between the average proscribed interventions delivered in research vs. practice settings, and these differences can inform implementation research.

**Was there a difference across settings in the delivery of prescribed interventions?**

The third study hypothesis was that therapists delivering CBT in the research setting would deliver prescribed therapeutic interventions (i.e., core Coping Cat interventions) more extensively than therapists delivering CBT in the practice setting. This finding was supported, indicating that therapists delivering CBT in research settings may adhere to the CBT protocols more than therapists delivering CBT in practice settings, as some researchers have hypothesized (McLeod et al., 2013; Weisz et al., 2009). Although past studies have indicated that therapists’ delivery of core therapeutic interventions may be notably low in practice settings (Southam-Gerow et al., 2010), the current study was the first to directly contrast delivery of prescribed interventions in a practice setting with that of a research setting. This study was the first to indicate a significant adherence discrepancy across the settings.

Understanding the ways in which treatment adherence differs across settings could be valuable to the field of implementation research. For instance, if future researchers found a positive adherence-outcome association, they could use this characterization of prescribed interventions for benchmarking purposes. They could train therapists in practice settings to reach a minimum prescribed extensiveness benchmark (e.g., using the TPOCS-RS) before they were certificated to participate in the study. In the current study, therapists delivering CBT in practice settings fell a point and a half below that of therapists in the “gold standard” research condition (i.e., 5.65). This relatively modest discrepancy suggests that helping therapists in practice settings boost their extensiveness level would not be an impractical goal. A few researchers have
already begun to use the TPOCS-RS for similar purposes and will soon be able to provide the field with more information about its effectiveness (e.g., Nakamura and colleagues at University of Hawaii).

The fourth hypothesis proposed that therapists who were told to deliver the interventions they routinely used and believed to be effective, those in the UC condition, would deliver the least amount of prescribed therapeutic interventions (i.e., interventions that aligned with the CBT protocol, not UC). This hypothesis was supported and was consistent with past descriptive studies of UC. For example, one past study indicated that therapists in UC delivered notably low doses of core or “active” EBT components such as homework (Garland et al., 2010). More pertinently, this control comparison indicated that therapists delivering CBT in a practice setting delivered prescribed interventions at a level more similar to the “gold standard” than the control condition. This finding suggests that although therapists delivering CBT in practice settings may not meet “gold standard” expectations, they may still deliver doses that are higher than those delivered in UC.

Thus, it may be that some CBT implementation procedures (i.e., instructions that delineate the appropriate training and supervision practices of the EBT) adequately promoted treatment adherence. In the future, researchers could identify the adherence-promoting procedures and make appropriate implementation alterations depending on process-outcome findings. For example, if future researchers found that a moderate level of adherence (i.e., that of CBT in the practice setting), rather than a high level of adherence (i.e., that of the CBT in the research setting) predicted positive client outcomes, they may investigate what types of CBT implementation procedures foster flexibility in adherence, so that those strategies could be incorporated in future training sessions and protocols. Taken together, these findings offer future
researchers an incentive to explore how variation in therapists’ delivery of interventions outside the protocol may influence client outcomes (e.g., amount of symptom reduction), so that they can ultimately shed light on ways to improve CBT implementation.

**How Do the Primary Findings Relate to the Conceptual Model?**

On the whole, these findings suggest that when therapists in practice settings are instructed to use and adhere to a protocol, they tend to incorporate more proscribed and less prescribed interventions in each session than therapists in research settings (McLeod et al., 2013). The Conceptual Model of Therapeutic Change in Therapy would suggest that the differences in CBT delivery across settings may have been due to discrepant therapy inputs (e.g., client characteristics) or implementation procedures which may in turn, have led to differences in change mechanisms (e.g., identifying automatic thoughts) and client outcomes (e.g., reduction in symptoms; see Figure 1 above; McLeod, Islam, & Wheat, 2013).

In the current study, client characteristics are one therapy input that may have led to implementation differences. Therapists in the practice setting may have delivered more proscribed therapeutic interventions due to the significant client differences in the two settings. For example, clients in the practice setting were more ethnically diverse, less financially fortunate, and had more externalizing disorders than clients in the research setting. Thus, therapists in the CBT condition of the practice setting may have believed it appropriate to deliver family and psychodynamic interventions (i.e., interventions focused on childhood/familial relationships) more extensively given that their clients may have had more crises (e.g., financial, relational) or different impairing problem areas (e.g., conduct) than clients in research settings (Southam-Gerow et al., 2003; Weisz et al., 2009). For example, therapists in practice settings may have felt compelled to respond to clients’ unique problem area needs when delivering
proscribed therapeutic interventions because CBT did not sufficiently enable them to address certain issues (Southam-Gerow et al., 2003). In this way, client characteristics may have influenced CBT implementation, and according to the model, may have influenced certain change mechanisms and outcomes.

However, the current study did not investigate how specific implementation differences influenced certain outcomes. Given that client outcomes (i.e., CBCL \( T \) scores, diagnostic remission rates) were generally the same across conditions, these findings may indicate that differences in the delivery of prescribed and proscribed interventions may have mixed, complex, or isolated effects on certain change mechanisms and client outcomes. That is, delivering certain types and doses of therapeutic interventions in each condition may influence change mechanisms and client outcomes in unique and potentially equally effective ways. Given that the current study aimed to describe the intervention extensiveness discrepancies, rather than explain them, future researchers will have to explore these process-outcome associations and their implications.

In general, future researchers should more deeply probe how therapy inputs and implementation procedures influence implementation integrity and subsequent client outcomes. For example, if they found that the delivery of a moderate amount of proscribed interventions actually predicted better client outcomes than a lower amount of proscribed interventions, future researchers could consider making appropriate adaptations to the CBT program itself, to better align with contextual characteristics of practice settings (Hogue et al., 2008). Perhaps the delivery of certain family interventions (e.g., recruits other family members) is necessary in practice settings due to the unique client characteristics (e.g., ethnicity, family income; Garland et al., 2010). The ultimate goal of implementation research is to modify the EBT or improve implementation procedures so that optimal performance is achieved in a multitude of practice
settings (Southam-Gerow et al., 2008). Future implementation researchers should capitalize on current study findings to bring the field a step closer to achieving that goal.

**What Did CBT Delivery Look Like in Research and Practice Settings Over Time?**

The way therapists delivered therapeutic interventions sometimes changed over time, depending on the type of interventions delivered. Over the course of treatment, all therapists delivered either increasing or decreasing levels of prescribed, behavioral, and psychodynamic interventions, depending on their condition. Therapist’s delivery of cognitive, family, and client-centered interventions, on the other hand, did not significantly change over the course of treatment. Given the lack of literature on implementation integrity trajectories, no a priori hypotheses were formed regarding these analyses. However, the delivery of CBT was characterized, compared, and contrasted in both settings to give researchers ideas of how they could test certain delivery differences in future studies and eventually enhance future implementation efforts.

*What did the delivery of proscribed interventions look like over time?* Therapists in all conditions delivered more items on the Family subscale consistently over the course of treatment whereas they delivered increasing or decreasing doses of interventions from the Psychodynamic subscale over the course of treatment. Interestingly, trajectories of extensiveness did not significantly differ across conditions for both treatment domains. In other words, therapists in both research and practice settings delivered proscribed interventions at the same rate of change over the course of treatment. Overall, these findings highlight the advantage of assessing treatment differentiation over time in implementation research; past researchers have found that therapists depart from the protocol (Barber et al., 2006; Hogue et al., 2008), but this
study demonstrates that therapists, regardless of treatment setting, may diverge from the protocol in specific, systematic ways over time.

Understanding the trajectory of proscribed intervention delivery may help move the field a step closer to solving issues in the transportation and implementation of CBT. For instance, if future researchers found that the delivery of psychodynamic interventions predicted poorer client outcomes, the trajectory of psychodynamic interventions could pinpoint one area of CBT implementation that requires development: quality control efforts in treatment monitoring. All therapists delivering CBT, but especially those delivering CBT in the practice settings, delivered more psychodynamic interventions as treatment progressed. Perhaps more time passed since their initial CBT training and they began to naturally deliver interventions they once used and believed to be effective. In this scenario, future implementation coordinators could consider teaching therapists in practice settings more explicitly about the type of interventions that implicitly violate CBT protocol (i.e., family and psychodynamic interventions), and revisiting this topic biweekly to prevent therapist drift. Throughout the study, supervisors could watch therapist tapes and use the TPOCS-RS to encourage therapists to stay below the proscribed benchmark (i.e., maximum extensiveness level expected; 1.23 for psychodynamic). Conversely, if future researchers found that increased delivery of psychodynamic interventions over time actually predicted better client outcomes, future researchers could assess which specific psychodynamic interventions (e.g., explores past) influenced outcomes so that these interventions could be incorporated in the treatment in some way.

**What did the delivery of prescribed interventions look like over time?** As time in treatment progressed, the extensiveness of interventions prescribed by the Coping Cat program grew increasingly different across conditions. Therapists delivering CBT in the research setting
delivered increasing levels of prescribed interventions as time passed whereas therapists in the practice setting (i.e., YAS-CC and YAS-UC) delivered decreasing levels of prescribed interventions as time passed. Therapists in the ICBT condition delivered increasingly more Coping Cat interventions as time progressed than both the YAS-CC and YAS-UC conditions. Therapists in YAS-CC and YAS-UC did not differ in their change of Coping Cat intervention delivery. The trajectory of prescribed intervention extensiveness could have important implications for CBT implementation. For example, if future researchers found that decreased CBT adherence in the second half of treatment predicted positive client outcomes, implementation researchers could consider adapting CBT protocols or training sessions to encourage therapists to hold less structured sessions as treatment progressed (i.e., practice-based evidence). On the other hand, if increased adherence in the second half of treatment predicted more positive client outcomes, future implementation researchers might consider incorporating a booster CBT training session or brief consultation for therapists mid-treatment to raise adherence levels. These proscribed and prescribed trajectories can give future researchers ideas about ways they can comprehensively evaluate CBT implementation and eventually modify CBT implementation or protocols.

**How Do the Exploratory Trajectory Findings Relate to the Conceptual Model?**

The current study indicated that depending on the type of intervention, extensiveness differences between research and practice settings may get more pronounced over the course of treatment (McLeod et al., 2013). Given that these trajectories have not been assessed in comparable studies, no literature, to my knowledge, offers reasons for why therapists might deliver interventions along certain extensiveness paths or how these differences might inform future research. Using the Conceptual Model of Therapeutic Change in Therapy, I will speculate
as to why delivery may have changed over time as a result of therapy input and implementation differences (McLeod, Islam, Wheat, 2013; Southam-Gerow et al., 2008).

When examining the delivery of proscribed interventions, all therapists (regardless of condition) delivered consistent doses of family interventions and increasing doses of psychodynamic interventions over time. Certain contextual differences may have contributed to this discrepancy. For example, therapists may have delivered consistent doses of family interventions throughout treatment because the client’s family was either heavily or minimally included in treatment; perhaps in the practice setting, families were more involved in treatment than they were in the research setting and therapists naturally responded by consistently delivering more extensive family interventions (i.e., a client characteristic). The proscribed trajectories could also be related to the perceived nature of the interventions themselves (i.e., a therapist characteristic; Garland et al., 2010). Therapists in practice settings may have believed family interventions were highly relevant and applicable to therapy, regardless of treatment phase. Conversely, they may have believed that the delivery of psychodynamic interventions was more appropriate or applicable as treatment progressed. Finally, it could be that therapists in both CBT conditions delivered family interventions more regularly than psychodynamic interventions because CBT supervisors were generally more accepting of the use of family interventions given their common occurrence in routine practice (Garland et al., 2010). However, these interpretations of proscribed intervention trajectories are mere speculations and will need to be explored in future research.

Unlike the proscribed trajectories, the prescribed intervention trajectory differed according to condition. Therapists in the research setting delivered increasing amounts of prescribed interventions whereas therapists in the practice setting delivered decreasing amounts
of prescribed interventions. One interpretation of this finding is that the Coping Cat protocol encourages therapists to deliver increasing amounts of prescribed interventions over the course of treatment and therapists in the research setting adhere to this structure better than therapists in the practice setting. For example, at the beginning of treatment, the Coping Cat protocol instructs therapists to spend time engaging in assessment and rapport building exercises, interventions that are not considered core CBT interventions. Conversely, the Coping Cat protocol instructs therapists to engage in exposure interventions in the second half of treatment; exposure is considered a core CBT intervention, and may thus boost the delivery of prescribed interventions as treatment progresses. Using this same reasoning, it could be that therapists in the practice setting decline in treatment adherence as treatment progresses because they deliver exposures less extensively than therapists in the research setting. Finally, it could be that different therapy inputs (i.e., client, therapist, setting characteristics) lead to the differences in trajectories across settings. For example, it could be that the organization policies in practice settings did not accommodate exposures (e.g., off-site therapy sessions) as well as research settings, leading to the implementation discrepancy. Again, these rationales for trajectory differences across settings and over time are mere speculations, as testing them was outside the scope of the current study. However, these points aim to pave the way for future implementation research.

Future researchers should assess if there is a significant difference in the delivery of exposure interventions across settings and investigate how these differences predict outcomes, as this could be important for implementation studies. For example, if researchers found that higher adherence at the end of treatment predicted positive client outcomes, training coordinators could consider spending more time teaching therapists how to deliver exposures (e.g., role-play delivering exposures) in practice settings. If researchers found that lower adherence at the end of
treatment predicted more positive client outcomes, they could consider adjusting CBT protocol procedures for exposures, to be more in line with practice setting needs. Identifying how implementation integrity differs in systematic ways, on average and across time, is an initial step toward helping researchers understand why CBT sometimes fails to produce expected client outcomes in practice settings (e.g., Guarton et al., 1992).

**Other Exploratory Findings and Implications**

Exploratory analyses indicated that there were significant differences across conditions in the delivery of cognitive, behavioral, and client-centered interventions. Therapists delivering CBT in the research setting delivered cognitive interventions significantly more extensively than therapists delivering CBT in the practice setting, and both of these conditions delivered cognitive interventions more than therapists in UC. Given that certain interventions from within the Cognitive subscale were selected to construct the Coping Cat subscale, the similar pattern of overall extensiveness levels across conditions made conceptual sense.

Interestingly, findings related to behavioral interventions differed from those of cognitive interventions. Therapists delivering CBT in the research setting delivered more behavioral interventions than therapists in UC, but their delivery of behavioral interventions did not significantly differ from that of therapists delivering CBT in practice settings. This finding indicates that therapists delivering CBT across settings may differ mostly in their delivery of cognitive (e.g., cognitive distortion), rather than behavioral interventions (e.g., relaxation). Exploratory analyses also indicated that therapists in research settings tended to increase their delivery of behavioral interventions over time (similar to the Coping Cat subscale), whereas therapists in practice settings did not increase their behavioral intervention delivery. It could be that therapists in practice settings deliver specific behavioral interventions early in treatment
(e.g., relaxation) at similar levels to research settings whereas their delivery extensiveness of other behavioral interventions (e.g., exposure) differs.

In addition, in the “gold standard” research setting condition, behavioral and Coping Cat interventions were delivered at increasing extensiveness levels over treatment while extensiveness levels of cognitive interventions did not change. Perhaps therapists focus on clients’ inner processes (e.g., cognitions) throughout treatment because these strategies are easier for clients to understand or are more consistently suitable whereas behavioral interventions require prior skill acquisition and are thus reserved for later modules (e.g., behavioral activation). Another possibility is that the structure of the Coping Cat protocol itself may account for this finding. The Coping Cat protocol instructs therapists to engage in assessment and rapport building exercises at the beginning of treatment, and these interventions would not be represented in the way the TPOCS-RS was scored. Conversely, the protocol instructs therapist to begin exposure sessions in the second half of treatment, interventions that would mostly be captured by codes within the Behavioral subscale of the TPOCS-RS. Future researchers should probe the differences between cognitive, behavioral, and Coping Cat trajectories to see if these findings can be replicated or meaningfully applied to implementation studies.

Exploratory analyses also indicated that therapists in the research setting delivered client-centered interventions (e.g., extensively validated the client, took the client’s perspective) more extensively than therapists in both conditions of the practice setting, and that therapists in both conditions in the practice setting did not differ in their delivery of client-centered interventions. Importantly, this finding differs from the parent study, which found that therapists in UC delivered more client-centered interventions than therapists delivering CBT in the practice setting (Southam-Gerow et al., 2010). When the delivery of client-centered interventions was
compared across the practice setting UC and CBT conditions using the same scoring strategy as the parent study, this finding was replicated. Given the larger sample of sessions in the current study, and thus its increased power, current study findings likely more accurately depict the data (Ioannidis, 2005; Kazdin, 2003).

These client-centered findings could be interpreted in a multitude of ways. It is possible that because client-centered interventions are integral to building rapport with the client, the rapport building exercises (e.g., “get to know you” activities, games at the end of session) built into the Coping Cat protocol tend to bolster the delivery of client-centered interventions. Given that client-centered interventions are oftentimes delivered with the aim of building a strong client-therapist relationship, this finding offered some support for literature suggesting that protocols do not hinder the development of a strong alliance (Langer, McLeod, & Weisz, 2011), an argument that has been debated in past literature (e.g., Addis & Krasnow, 2000). In fact, the relationship component of implementation integrity may be equally as important as the therapeutic intervention component (i.e., adherence and treatment differentiation), in altering clients’ change mechanisms (e.g., increase change in cognitions), and subsequent outcomes (e.g., reducing anxiety symptoms; McLeod, Islam, & Wheat, 2013). Another interpretation is that therapists in the research setting also deviate from the protocol, but perhaps in more expected or “acceptable” ways (Waltz et al., 1993). Lastly, the delivery of client-centered interventions did not change over the course of treatment, regardless of condition. Similar to the Family subscale, it could be that therapists feel more comfortable delivering client-centered interventions in tandem with core interventions, in all stages of treatment. In general, these findings illustrate that CBT delivery differs across research and practice settings, even when assessing interventions that are not considered integral to or implicit violations of CBT protocol.
Finally, it is worth noting that the two types of proscribed therapeutic interventions appeared to be delivered at different average levels, across conditions. Family interventions were delivered more frequently and thoroughly than psychodynamic interventions. Some evidence suggests that Family therapy may be more widely practiced than Psychodynamic therapy in practice-settings (Baumann, Kolko, Collins, & Herschell, 2006; Garland et al., 2010); this may have been reflected in the higher overall extensiveness levels of family versus psychodynamic interventions delivered. There could also be a methodological explanation of this finding. Given the presence of parent sessions in the CBT conditions of both research and practice settings, the delivery of family interventions may have simply been more appropriate than the delivery of psychodynamic interventions. Future researchers could test if results differ without the inclusion of parent sessions. Overall, researchers should examine how these two types of interventions differentially influence outcomes (with and without parent sessions) and make necessary CBT implementation alterations as seen fit.

Strengths and Limitations

The current study benefitted from a number of strengths. For one, an observational instrument was used to measure therapist performance, providing less biased data than data gathered from therapists’ self-reports of their own performance (DiMatteo, 2004; Hill, 1991; Mowbray et al., 2003). Second, HLM analyses were conducted which allowed the current study to (a) account for missing data, (b) describe the change in the delivery of certain therapeutic interventions over time, and (c) depict detailed graphs of the course of intervention delivery in each treatment domain subscale (Singer & Willet, 2003). Combined, these factors increased the internal validity of the findings and will provide researchers with new ways of conceptualizing implementation integrity over the course of treatment.
In addition, the current study extended the findings from the YAS parent study (i.e., Southam-Gerow et al., 2010). The current study was conducted with a much greater number of sessions than the YAS study and when the same subscale scoring strategy used in the YAS study was applied to the current sample, there were slightly different findings. Unlike the YAS study, the current study found that therapists in the UC condition did not differ from therapists in the CBT condition on the extensiveness of psychodynamic and client-centered interventions delivered. This was likely a result of the current study’s significantly larger sample of sessions, which worked to increase its power (Ioannidis, 2005; Kazdin, 2003). On the whole, the disparate parent and current study findings underscore the importance for future researchers to replicate this study using the aforementioned methods that fortified the study’s design.

Despite the advantages of the current study’s design, there were some methodological problems that limited the conclusions. For one, self-report measures of intervention delivery were not integrated into the design, limiting the study to an observational measure only. Although this issue was addressed by ensuring that the average item inter-rater reliability was above a “good” level (ICC > .60; Shrout, Spitzer, & Fleiss, 1987), there may have been some remaining observer errors or biases that influenced the findings. Observational measures also did not assess therapists’ cognitions, intentions, or feelings, nor did it assess clients’ responses to certain interventions (Garland et al., 2010). Future researchers should strive to include multi-method (i.e., both self-report and observational) measures into their implementation studies to enhance the internal validity and informational value of the findings. For instance, researchers could use self-report measures to understand why therapists diverge from the protocol in particular areas. Second, all therapy sessions held during the study were not available for coding or deemed codable (e.g., damaged, inaudible). Though the current study analyses suggested that
the majority of sessions were coded (i.e., greater than two-thirds) and equally sampled from all phases of treatment, sessions held past 40 weeks in treatment were excluded from the analyses due to falling more than two and a half standard deviations above the mean. Thus, for two clients, the findings did not capture the full course of intervention delivery. Lastly, the two parent studies did not administer the same diagnostic instruments; one used a semi-structured interview (ADIS-IV) and one used a highly structured, computer-administered interview (DISC 4.0). Participant inclusion in the study as well as rates of primary diagnoses and externalizing disorders therefore differed across studies, decreasing the interpretability of the findings.

Researchers attempting to conduct similar studies in the future should prioritize the protection and organization of session video recordings and attempt to ensure that conditions in research and lab settings administer the same diagnostic and assessment measures.

There were also some data and analytic issues that limited the study conclusions. For one, therapists did not deliver therapeutic interventions from all treatment domain subscales to equal degrees. Specifically, few therapists delivered therapeutic interventions that fell within the Psychodynamic subscale of the TPOCS-RS, resulting in a non-normal distribution. Implications related to the Psychodynamic subscale should therefore be considered with caution; only a handful of therapists delivered psychodynamic interventions at high extensiveness levels in practice setting conditions, suggesting that these findings may be due to therapist characteristics and are not generalizable to other practice settings.

Additionally, although the current study’s heterogeneous sample increased external validity, it also increased error variance and thus decreased internal validity (Southam-Gerow et al., 2010). More specifically, conditions were composed of recordings with significantly different client and therapist characteristics with the exception of therapist gender and CBCL
Internalizing $T$ scores. Unfortunately, even the findings related to therapist gender must be interpreted with caution because there were significantly more female than male therapists in the study. Moreover, the Kendall et al. study did not collect as much data on therapist characteristics (e.g., age, professional experience) as the YAS study, further restricting the possibility of conducting sample comparisons across conditions. In general, the study had an unbalanced design, thus I could not accurately model therapist factors. Although clients were nested within therapists, therapist was not included as a third level in the HLM analyses because therapist did not account for significant variation in model construction (likely because a number of therapists only had one client). Taken together, these issues with client and therapist factors limited the interpretability of the findings.

Finally, the current study was merely descriptive in nature. No causal inferences could be drawn from the findings, and no specific, conclusive solutions could be offered to mental health professionals. However, researchers should use the descriptive results of the current study to further future implementation research.

**Future Directions**

The current study’s comparative characterization can inform future implementation research in a number of ways. For one, future researchers could assess how delivering certain types of interventions (i.e., prescribed, proscribed) at certain extensiveness levels differentially affect clients’ session-by-session outcomes (e.g., symptom reduction, satisfaction with treatment) across settings; this would allow them to make necessary alterations to either CBT implementation or the CBT protocol itself (Fixsen et al., 2005). Second, depending on process-outcome findings, researchers could consider using the mean extensiveness levels delivered in the present study as a benchmark for future implementation studies (McLeod et al., 2013). Third,
researchers could conduct studies in which they modify one aspect of implementation integrity to assess how specific implementation procedures in isolation (e.g., more or less therapist training) influence implementation integrity and outcomes. Fourth, researchers could use descriptive information from this study to conduct more qualitative or mixed methods studies aimed at improving CBT protocol or implementation. Finally, researchers could conduct studies in which they match clients across research and practice settings, to better control for client and therapist characteristics.

In regards to assessing implementation-outcome links, if researchers find that areas where CBT implementation falls short of expectations (i.e., significantly less prescribed and/or more proscribed therapeutic interventions are delivered) correspond with a gap in symptom reduction or even an increase in negative symptoms, they could brainstorm ways to improve quality control efforts, such as modifying CBT training sessions in practice settings (Fixsen et al., 2005; Frank et al., 1991; McLeod & Southam-Gerow, 2012). Given that some research has suggested a positive adherence-outcome relationship, this implementation plan is likely (Hogue et al., 2008). For example, therapists in practice settings may not understand the importance of delivering one or two core interventions to a highly extensive degree (i.e., frequently and thoroughly over the course of a session; Garland et al., 2010); future researchers may consider emphasizing this principle. In particular, researchers could modify training sessions to include more “active” learning techniques found to be effective in therapist training, such as using role-rehearsal with feedback rather than didactic teaching (Beidas & Kendall, 2010; Edmunds, Beidas, and Kendall, 2013). Process-outcome research could thus spur efforts to improve therapist training or supervision practices in EBT implementation studies.
On the other hand, if future researchers find that areas where CBT implementation falls short of expectations actually predict better client outcomes (i.e., outcome-augmenters; McLeod et al., 2013) than the areas where CBT implementation meets expectations, they could take a different approach. This type of adherence-outcome relationship finding may not be as unlikely as some might suspect, as some researchers have found a negative curvilinear relation to EBT adherence and outcomes such that therapists produce optimal outcomes when they moderately adhere to the EBT protocol (e.g., deliver a rating of a 4 on the TPOCS-RS; Barber et al., 2006). Further, proscribed interventions might produce positive client outcomes; one study using the TPOCS-RS found that high extensiveness of two items on the Psychodynamic subscale predicted reduced parent-reported depression (Weisz et al., 2009). In this scenario, it would be important for future researchers to further investigate the positive outcomes of these proscribed therapeutic interventions, disseminate replicable findings, and where appropriate, incorporate them into implementation procedures or treatment protocols.

In addition, as mentioned above, depending on process-outcome findings, future researchers could use benchmarks to ascertain that therapists in practice settings have reached a level where they are capable of delivering core CBT interventions to certain extensiveness levels. This quality control method could be used across implementation studies, allowing for cross-study comparisons. This could also enable process-outcome researchers to control for two aspects of implementation integrity (i.e., treatment adherence and treatment differentiation) when testing whether other implementation integrity components (i.e., relational factors, therapist competence) or therapy inputs (e.g., client characteristics) explain outcome differences. For example, if outcomes still differed across research and practice settings when the delivery of prescribed and proscribed interventions were the same, researchers could focus on investigating
how therapist competence or client characteristics (e.g., comorbidity) influence outcomes (e.g., less symptom reduction). Treatment integrity benchmarks could allow researchers to understand and control aspects of implementation integrity in a way that would further enhance their studies.

Third, researchers could conduct studies in which they modify one implementation procedure in isolation (e.g., active learning techniques in supervision) and assess its influence on implementation integrity (i.e., levels of treatment adherence and differentiation) and client outcomes (Edmunds at al., 2013). For example, if higher treatment adherence is associated with positive client outcomes, researchers could use extensiveness benchmarks to pinpoint the types of implementation procedures that enable therapists in practice settings to meet adherence benchmarks and the types of procedures that hinder therapists from meeting the adherence benchmarks. Understanding the effectiveness of certain implementation procedures on process-level outcomes would undoubtedly enhance CBT implementation efforts.

Fourth, future researchers may consider conducting more qualitative or mixed studies assessing CBT implementation. Qualitative studies allow for a more holistic understanding of the contexts under investigation, and thus would provide comprehensive, descriptive information of what exactly occurs in implementation studies (Kazdin, 2003). Therapists in practice settings will likely be willing to participate in these kinds of studies (e.g., interviews), as they have done so in past studies and many therapists agree that the use of such data can improve their services (e.g., Baumann et al., 2006; Bickman et al., 2000). In particular, it would be helpful to ascertain whether therapists in practice settings recognized that they were deviating from the manual at the same time that observers identified the divergence, to ascertain if the divergence was due to deficits in skill or difference of opinion/attitude. It would also be useful to know why therapists typically decide to deliver family and psychodynamic interventions at particular stages in therapy.
(e.g., due to certain client characteristics or responses), so these decisions could either be promoted or resolved in training and supervision sessions (depending on process-outcome findings). Additionally, researchers could ask therapists questions about what they found helpful or unhelpful from the initial CBT training (Beidas & Kendall, 2010). Finally, researchers could request that therapists, supervisors, and stakeholders complete both qualitative and quantitative measures that would allow them to assess the influence of contextual factors such as attitudes toward protocols or perceptions of organizational climate (Baumann et al., 2006; Southam-Gerow et al., 2008). Answers to these questions may help researchers tailor training, supervision, and other implementation procedures in ways that boost therapists’ level of confidence, adherence, and competence in delivering CBT interventions.

Finally, given the different client and therapist characteristics in research and practice settings, future researchers could conduct analyses in which they control for demographic variables across settings by matching, or grouping participants with similar demographics together before analyzing the data (Kazdin, 2003). For instance, using current study data, participants could be matched on characteristics such as primary diagnoses, ethnicity, and rate or severity of externalizing disorders. This would control for some of the confounding effects of client and therapist characteristics, and would enhance the interpretability of future findings.

Conclusion

Implementation researchers are attempting to change the fact that 90 percent of youths who receive public mental health services are in programs with no evidence supporting their impact (Bickman, 2008). By bringing EBTs into practice settings, implementation researchers anticipated that youth would finally receive the quality of care they ought to receive. However, even when EBTs are transported to practice settings, therapists may not be delivering them in the
way they were designed (Weisz et al., 2006). The current study characterized the delivery of
CBT for youth with anxiety to exemplify ways in which the within session processes of EBTs
(i.e., implementation integrity) may differ across settings. Even though client outcomes were the
same across the research and practice setting in this study, there were key differences in the type
and extensiveness of EBT interventions delivered in both settings, on average and over time.
Through combined D&I and treatment integrity frameworks, I presented future researchers with
a picture of what was actually occurring in both contexts with the hope that future researchers
would explore how or why certain EBT implementation difficulties or client outcomes occur
(Doss, 2004; Kazdin, 2000). By doing so, the current study moves the field a step closer to
resolving EBT transportation and implementation issues.

In sum, EBT delivery differences warrant the field’s attention because they suggest that
there are in-session processes that may have hindered EBT adoption in past implementation
studies (especially those that had less favorable client outcomes; e.g., Guarton, 1992; Hawkins et
al., 1991), and continue to be problematic in current implementation studies. These difficulties at
the implementation level have the potential to halt EBT dissemination at a large-scale level
(Southam-Gerow et al., 2008), and prevent a large number of youth from receiving treatments
that have been proven to be effective in lab settings (Kazdin, 2000). As a field, we need to
continue to methodically assess and adjust the strategies we use to transport EBTs to practice
settings so that we can begin to close the science-practice performance gap, and ultimately
provide the millions of youths suffering from mental health disorders the quality of care they
deserve.
List of References


Barkley (Eds.), *Treatment of childhood disorders* (pp. 271–335). New York, NY: Guilford.


National Institute of Mental Health. Release of landmark and collaborative study conducted by Harvard University, the University of Michigan and the NIMH Intramural Research Program (release dated June 6, 2005 and accessed at www.nimh.nih.gov).


Adolescent Psychiatry, 49(10), 1043-52. doi:10.1016/j.jaac.2010.06.009


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 before entering the Clinical Psychology doctoral program at Virginia Commonwealth University in Richmond, Virginia.