8-2018

Demonstrating Functional Analytic Psychotherapy as an Independent Variable in Efficacy Research: A New Measure of Treatment Fidelity

Lindsey E. Knott
Western Michigan University, LindseyEKnott@gmail.com

DEMONSTRATING FUNCTIONAL ANALYTIC PSYCHOTHERAPY AS AN INDEPENDENT VARIABLE IN EFFICACY RESEARCH: A NEW MEASURE OF TREATMENT FIDELITY

by

Lindsey E. Knott

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Psychology
Western Michigan University
August 2018

Doctoral Committee:

Scott Gaynor, Ph.D.
Amy Damashek, Ph.D.
Amy Naugle, Ph.D.
Chad Wetterneck, Ph.D.

DEMONSTRATING FUNCTIONAL ANALYTIC PSYCHOTHERAPY AS AN INDEPENDENT VARIABLE IN EFFICACY RESEARCH: A NEW MEASURE OF TREATMENT FIDELITY

Lindsey E. Knott, Ph.D.

Western Michigan University, 2018

Functional Analytic Psychotherapy (FAP) is a contextual behavior therapy that takes an in session, in vivo focus to improve client outcomes. This in vivo (IV) focus is distinctive of FAP and involves the therapist utilization of contingencies in session to decrease problematic response classes and increase more adaptive response classes (i.e., a differential reinforcement procedure). This contingent responding is proposed to be FAP's mechanism of action leading to client change. FAP efficacy research mainly consists of small n single-case studies or group designs combing FAP with another CBT intervention. Maitland & Gaynor (2012) offered recommendations for increasing FAP efficacy research, including the development of a reliable measure to determine whether a therapist implemented the distinguishing features of FAP. Maitland and Gaynor (2016) developed a 10-item fidelity measure, called the FAP Adherence Form (FAP-AF), to assess for aspects of FAP across 5 items and distinguish it from supportive listening (SL) assessed over 4 items. The measure was used in two FAP outcome studies (Maitland & Gaynor, 2016; Maitland et al., 2016) and it distinguished FAP from SL sessions, suggesting potential utility in distinguishing FAP as an independent variable. The present study, using the data from Maitland et al. (2016), provides a more comprehensive analysis by assessing reliability between 3 independent coders, correlations with another FAP-specific fidelity metric, and investigating for possible mediator relationships between the FAP-AF and outcomes. The

alternative fidelity technique used for comparison was developed by Kanter, Schildcrout, & Kohlenberg (2005) and involves coding each turn of speech as in vivo (talk directed at the therapy process, therapy relationship, or in the moment interactions) or not. If the FAP-AF and in vivo counts capture the presence or absence of FAP elements, they should be reliable across coders. Likewise, the in vivo turns count should correlate with the 5-item FAP scale on the FAP-AF, but neither should correlate with the 4-item SL scale, distinguishing FAP from SL. Finally, if the FAP-AF acts as a proxy for FAP's mechanism of action it should mediate outcomes. Results demonstrated mostly excellent reliability between coders. Exact agreement on the FAP scale items ranged from 71% to 95%. Inter-method correlations indicated strong positive relationships between and within coders on the adherence metrics. Furthermore, the FAP scale and IV turns count significantly differed in FAP sessions compared to SL sessions suggesting the measures distinguished the presence of FAP from its absence. As such, the FAP scale may be useful in documenting FAP as an independent variable providing a reliable but less intensive method for assessing treatment fidelity. However, no adherence metrics replicated Maitland et al. (2016) in serving as a statistical mediator of outcomes. The individual level relationships observed between adherence and symptom outcomes suggested significant variability and an absence of an overall dose-response relationship. The lack of a clear dose-response relationship could indicate a limitation of the adherence measures used. Not all FAP interactions are equal in their potency an effect which is hard to quantify. Implications, future directions, and limitations are discussed.

ACKNOWLEDGEMENTS

I must first acknowledge my family—my parents, sister, and every single chosen family member I have acquired over the years. Your unwavering love and support, moments of selfless cheerleading, and encouraging words will forever be my light in the darkest of times. To my advisor, Dr. Scott Gaynor: I appreciate your depth of knowledge, compassion for others, fervor for teaching, and relentless adherence to the science in your psychology. Of the countless lessons I take from you, perhaps the most important is to always strive for integrity and growth as a professional, while never losing site of the values that provide us joy and connection. To my first and most enduring mentor, Dr. Chad Wetterneck: There are two truths related to your influence on my life. First, I would not be where I am today—on the cusp of achieving my dream—without your initial belief in me. Second, I would not be the person I am proud to call myself today – courageous, vulnerable, and authentic—without both your introduction to FAP into my life and your lived self as a model for those qualities. Endless thanks to you. To all my lab-mates, especially Rachel Petts and Rebecca Rausch: Thank you for your help with this project and for being a home base during my time at WMU. To my cohort (and honorary cohort-mate, McKenna) and all my friends: I could not have made it without the karaoke/game nights, breakfast dates, hikes, Zumba, and your friendship. Finally, to Rory: You have been on the front-line with me, in my heart, this whole time. I am immensely fortunate to have you by my side.

Lindsey E. Knott

TABLE OF CONTENTS

Table of Contents -- Continued

Table of Contents -- Continued

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Functional Analytic Psychotherapy (FAP) is a contextual behavior therapy with roots in radical behaviorism that targets client daily life problematic behaviors through in-session contingencies (Tsai et al., 2009). To understand the processes within FAP requires a review of the conceptual underpinnings of FAP, in particular its roots in radical behaviorism. Radical behaviorism is a theoretical framework that aims to get to the root of human behavior. According to Donohoe and Palmer (2004), behavior is defined as all of the activities in which an organism may engage, both those that are observable by others and those that are not. Therefore, radical behaviorism seeks to address not only overt behavior, but to address feelings and private experiences. To understand behavior requires knowledge of its reinforcement history and the current context (Kohlenberg & Tsai, 1991). Thus, behavior is said to be a function of one's prior unique experiences and life history in interaction with the current context (Tsai et al., 2009).

A radical behavioral approach assumes that *antecedents* in the environment evoke all behavior and that the *consequences* of the behavior impact the strength (increase or decrease) or maintenance of the behavior. Antecedents and consequences are considered controlling variables. The relationship between controlling variables and behavior is termed a*functional relationship* (Kohlenberg & Tsai, 1991) and a description of functional relationships is the goal of assessment in a radical behavioral and FAP framework. The reasoning is pragmatic. Once a functional relation is identified the behavior therapist can effect to alter antecedent control (attempting to evoke alternative or more adaptive behavior) an provide differential reinforcement (positive reinforcement for adaptive behavior and extinction or mild punishment for maladaptive behavior). The FAP therapist does so by focusing on in session antecedents and consequences.

All consequences that impact the strength of behavior (by increasing or decreasing) are considered *reinforcement*. Consequences that increase behavior are referred to as reinforcers,

while those that decrease behavior are called punishers. The removal of reinforcing consequences can also result in the decrease of behavior in a process called extinction. One important aspect is the proximity and immediacy of consequences. Generally, the most effective consequences are provided immediately after the behavior is emitted (Kohlenberg & Tsai, 1991), which is the basis for FAP's emphasis on in-session responding.

Principles of reinforcement and the perspective that behavior, both covert and overt, is controlled by environmental events, form the basic underpinnings of the techniques utilized in FAP. Following the pragmatic position that functional relationships exist for all behavior and that the context of reinforcement impacts the effectiveness of reinforcement, FAP hypothesizes that problematic behaviors emitted by the client in-session have a functional similarity to problematic behaviors in the client's daily life. The context of the therapy relationship must imply some contextual similarity to other contexts if previously reinforced behavior is emitted in the therapy setting. Thus, FAP presumes that the therapeutic relationship has some functional similarities to the client's out-of-session daily life relationships. FAP hypothesizes that many important functional relations for adult humans are interpersonal in nature, that is the relevant antecedents and consequences are social. Because therapy is an interpersonal interaction it is hypothesized that behaviors that are problematic (and adaptive) in daily life interactions will be evoked in therapy. Once evoked the FAP therapist is in a position to provide relatively immediate consequences for this in-session behavior to increase or decrease its frequency. Similarly, once problematic and adaptive interpersonal behaviors are identified the therapist can attempt to prompt the behavior allowing for additional opportunity to provide relevant consequences. Because the antecedents and consequences are those endemic to social encounters, FAP theorizes that altering reinforcement of in session client behavior can impact the

occurrence of behavior in functionally similar contexts outside of session. Within the therapy setting, the reinforcement of client's in-session behavior is done via the therapist's responses to the behavior emitted in session. Initially, FAP therapists attempt to reinforce any improvement but over time may change the threshold for reinforcement as the client's adaptive repertoire grows a process called shaping. Shaping is defined as the reinforcement of successive approximations producing gradual changes in behavior; thus, it is intended to strengthen a more adaptive repertoire of behavior through therapists' contingent responses to in-session behavior. When more adaptive interpersonal behavior is reinforced by therapist responses, client's out of session interpersonal behavior is predicted to improve to the extent that there are shared antecendents and consequences. Therefore, therapist contingent responding to client in-session behavior is theorized to be the mechanism of action in FAP.

FAP 5-rules and Shaping CRB

In order to facilitate FAP implementation with clients, researchers and FAP theorists developed 5 guidelines for shaping clinically relevant behavior (CRB). CRBs are those in-session behaviors that are considered functionally similar to out-of-session daily life problematic and adaptive behavior and are the central focus in FAP. The guidelines are referred to as the 5 FAP rules: 1) watch clinically relevant behavior (CRB), 2) evoke CRB, 3) reinforce CRB2, 4) observe potientially reinforcing effects on client CRB, and 5) summarize and generalize (Kohlenberg & Tsai, 1991). Weeks, Kanter, Bonow, Landes, & Busch (2012) out illustrate how the 5 FAP rules play out in what is called the logical FAP interaction: Identify a behavior in clients daily life that has an in-session parallel, observe or evoke the CRB, provide consequences, determine if the consequence had the intended function, and suggest how the CRB links to daily life behavior. The logical FAP interaction is comprised of the 5 rules, with

explicit emphasis on out-to-in and the in-to-out parallels that is designed to facilitate therapist implementation of FAP. The following section illustrates the logical FAP interaction with the FAP 5-rules.

The first rule is to watch for CRB. In order to effectively shape behavior, one must be able to identify, observe, and conduct functional analysis to understand what controlling variables are leading to the problematic behavior and how to respond to increase or decrease CRB. Those CRB that are problematic in the client's life are denoted as CRB1, while those CRB that are more adaptive or effective that a FAP therapist seeks to increase are denoted as CRB2 (Kohlenberg & Tsai, 1991). While the topographies of CRB1 and CRB2 may appear different (e.g., when given a compliment: CRB1 - bowing head, averting eye contact, and complimenting the other person instead. CRB2 - smiling and saying "thank you") the function of the behavior is of most concern. For example, in the compliment example, the CRB1 and CRB2 delineation could be reversed if someone were from a particular Eastern culture where humility is valued more than a heightened self, and smiling and saying "thank you" could seem boastful and function to alienate the person from their cultural group. The function, or what antecedents and consequences maintain the behavior, is of more importance than the topography, or how the behavior appears (Kohlenberg & Tsai, 1991).

The second rule is to evoke CRB. Therapist behavior and therapy setting events (e.g., homework or timeframe) may act as antecedents to elicit or evoke behavior of interest (Kohlenberg & Tsai, 1991). For a client who is late to work or struggles with responsibility, a discussion of timeliness when the client is late to session can evoke CRB that should parallel a confrontation similar to out-of-session behavior that may involve CRB, both internal (e.g., feelings of guilt) or external (e.g., avoidance), that influence one's CRB of timeliness. When the

client is discussing out-of-session problematic behavior, checking with the client as to whether a related CRB occurs in the room, or doing what is called an "out-to-in parallel", may also evoke CRB. A parallel statement like this can also be the act of Rule 1 as well, simply observing CRB (Weeks et al., 2012). Other techniques therapists may use to evoke CRB that relate to client daily life problems could be to do a mindfulness exercise, have the client share something vulnerable, check-in with the client about their in-session feelings and reactions, etc. These examples may not evoke CRB for every client, because each client has different repertoires and should be approached in an idiographic sense; however, these examples provide ideas for ways in which to evoke CRB from different response classes.

The third rule is to reinforce CRB2. In their continued functional analysis of in-session behavior, therapists should be mindful of the effect of the client's behavior and provide the appropriate consequences for behavior emitted. Natural reinforcement is considered the most appropriate consequence as it provides an approximation of the effect that would be seen in the client's external environment (Kohlenberg & Tsai, 1991). Natural reinforcement is distinguished from arbitrary reinforcement, which is reinforcement that does not naturally occur in the environment. For example, M&M's for answering questions is arbitrary while the natural reinforcer for answering questions might be the therapist saying "correct" or "incorrect" and/or through the internal state of "feeling heard", or the therapist asking a meaningful follow-up question or making a validating follow-up statement. So, for example, a natural reinforcer for someone who struggles with sharing information that makes them feel vulnerable might be a genuine acknowledgement by the therapist of that struggle, as well as the impact it has on the therapist, rather than a simple "thank you," or common arbitrary reinforcer that may not be in direct relationship to the behavior.

The fourth rule is to observe potentially reinforcing effects of therapist behavior on client CRB. This rule is simply to see what happens to CRB in response to the therapist's responses. If the CRB2 increases after the therapist response the effect is what would be expected based on the principle of reinforcement. If the behavior does not occur again even when prompted, more reinforced instances, shaping of partially improved instances or altering the reinforcement type or frequency may be necessary. Therapists are also encouraged to discuss with the client their experience with the responses of the therapist in order to gauge the reinforcing value of the therapist's responses (Kohlenberg & Tsai, 1991).

The fifth and final rule is to offer interpretations of variables impacting client behavior and to promoted generalization. Using this rule, the therapist may provide the client with hypotheses about the functional relationships between the client's behavior and contextual variables and explicitly discuss in-to-out parallels. Rule 5 may also include the assignment of homework in order to facilitate generalization into the client's daily life (Kohlenberg & Tsai, 1991). For example, a therapist might say "could you do what you did in here with me with a person of interest?" This example homework assignment also fits within the logical interaction framework as an "in-to-out parallel."

FAP Research

FAP represents a relatively straightforward application of behavior principles to the interactions occurring in psychotherapy. Thus, FAP rests on a very strong behavioral science foundation. However, there is a relative dearth of research in direct support of FAP, including a lack of clinical trials (Kanter et al., 2017; Mangabeira, Kanter, and Del Prette, 2012; Singh & O'Brien, 2018). At present, there is one comprehensive review of the FAP literature with a focus on single-case design research (Singh & O'Brien, 2018) and one reviewing all of the

various designs (Kanter et al., 2017). These reviews indicate that the foundation of FAP literature is comprised of predominantly single-subject A/B design research (Gaynor & Lawrence, 2002; Landes, Kanter, Weeks, & Busch, 2013; Singh & O'Brien, 2018; etc.,) one alternating time-series design (Maitland & Gaynor, 2016), several FAP-enhanced clinical trials (Gifford et al, 2012; Holman et al., 2012; Kohlenberg et al, 2002), and one stand-alone randomized controlled trial of FAP (Maitland et at., 2016). The twenty single-case studies reviewed by Singh and O'Brien (2018; 18 published and 2 unpublished) were found to have clinically significant post-treatment differences not due to chance or the passage of time; however, these outcomes could not be determined to be solely related to FAP due to these designs often lacking comparison groups among other factors. . Kanter and colleagues (2017) review consisted of 20 qualitative studies, 19 single-case designs, 6 uncontrolled and controlled clinical trials, and three studies of FAP therapist trainings. Overall, their review established similar effects as Singh and O'Brien (2018), arguing that the literature is promising but not sufficient to claim FAP as an evidence-based practice (Kanter et al., 2017).

Establishing a treatment as an evidence-based practice has shifted over the years. The primary designs historically deemed necessary for this process are Randomized Controlled Trials (RCTs) or carefully controlled single-case designs with comparison groups (Chambless & Hollon, 1998). In order to be considered the standard of quality, these designs must be comprised of well-defined treatment conditions, compared to control conditions with detailed manuals to both train therapists and track therapist adherence, and include reliable and valid process and outcome measures to track outcome significance. This process is coined the *gold standard* way for conducting treatment outcome research and most often specifically refers to RCTs designs colloquially. Ten years ago, Hayes and colleagues (2005) summarized the empirical status of

several contextual behavior therapies (i.e., DBT, ACT, and FAP) and further proposed that these treatments, due to their contextual nature, may be more difficult to study in the *gold standard* way. Principle-based treatments that take a contextual approach to treatment targets, such as ACT and FAP, have been historically difficult to manualize and therefore difficult to adhere to this *gold standard* approach. Although they utilize similar theoretical principles, ACT and DBT have made considerable gains in research efficacy. In particular, ACT now has well over 100 RCTs to substantiate its efficacy as a transdiagnostic treatment for a variety of disorders, while FAP has continued to lag behind. Given this limitation, the authors suggested that these growing research communities be given grace, if you will, for their slowly developing research base.

In 1998, Chambless and Hollon proposed what have become an influential set of criteria for defining empirically supported treatments. Based upon the authors' suggestions, a designation of "well established" would require at least two between-group design experiments or large series (10 participants or more) of single-case designs with compelling outcome evidence. Further, they suggested that the research be conducted by independent research teams, using treatment manuals, and include diverse participants. While these criteria have been the standard approach for establishing empirical support for the past two decades, recent recommendations for the synthesis and development of treatment guidelines have emerged (Tolin, McKay, Forman, Klonsky, & Thombs, 2015). Considering the recommendations of APA work groups, Tolin and colleagues proposed a system for establishing treatment guidelines that considers existing quantitative reviews for quality, relevance and generalizability, and assesses risk of bias. This system also involves the translation of these reviews into guidelines using the GRADE tool, which is a measure of quality of the evidence in the review as rated: *high, moderate,* or *low.* A highly confident review is evidenced by: 1) a wide range of studies with few

limitations, 2) little variation between studies, and 3) a narrow confidence interval. A *moderate* rating is evidenced by: 1) few studies presented, with some limitations but no major flaws, and 2) some variations between studies, or a wide confidence interval. A *low* rating is evidenced by: 1) major flaws throughout studies, 2) variation throughout studies, and 3) exceptionally wide confidence interval. At this time, only Kanter and colleagues' (2017) provide an up-to-date and comprehensive review of the literature; as previously mentioned they concluded that the evidence is insufficient to make claims for FAP as a front-line evidence-based practice at this time. While FAP has been examined as a stand-alone treatment in one controlled between-groups design (Maitland et al., 2016) and various single-case designs from independent research teams on a variety of participants (Kanter et al., 2017; Sing & O'Brien, 2018), these studies are associated with a variety of limitations including limited comparison conditions and small samples sizes. Thus, according to both the new criteria of Tolin and colleagues and former criteria of Chambless and Hollon's, the current research does not indicate strong empirical support. However, FAP appears promising, especially in the area of social functioning, with some suggestive support for its mechanism of action as the therapists as an in-session contingent responder (Kanter et al., 2017). Despite FAP's principle-based approach derived from behavior theory and the existing evidence, FAP itself does not have enough data from well-controlled studies to be considered a stand-alone, empirically supported treatment.

There appear to be several methodological issues specific to FAP that are significant barriers for researchers who may want to attempt FAP efficacy research. FAP can be challenging to study because of its functional, idiographic nature. FAP targets are often based on individual problematic behavior from client's unique contexts, rather than broad symptomatic functioning (Maitland & Gaynor, 2012). This is related to the aforementioned perspective that

behavior is only meaningful when understood functionally in context, the topography of behavior is of less importance (Kohlenberg & Tsai, 1991). Single-subject designs are applicable for evaluating changes in CRB given their idiographic nature. However, it would take at least two large-series single subject studies of at least 10 participants to adhere to efficacy standards set by Chambless and Hollon (1998). Furthermore, the *gold standard* approach for treatment efficacy utilizes RCTs, which rely on manuals for purposes of replication, extension, and defining the treatment condition (i.e., the independent variable). As FAP's independent variable is proposed to be therapist contingent responses to individualized in-session problematic and adaptive behavior that parallels out of session targets, this broad scope makes for difficulty in development of treatment manuals and standard measurement techniques intended to capture processes in FAP treatment. That is, how can FAP's broad reach is very promising for treating clients with any presenting problem; however, it does not specify a priori any particular DV (e.g., depression). FAP was not developed as a treatment for any particular DSM disorder, a deviation from the standard treatment development model (Rounsaville et al., 2001).

Evaluating Treatment Integrity

A methodological issue that is critical to treatment development is the assessment of treatment fidelity. Treatment fidelity or integrity refers to the degree to which the identified treatment (i.e., independent variable) has been implemented as intended (Waltz, Addis, Koerner, & Jacobson, 1993; Perepletchikova, Treat, & Kazdin, 2007). If the treatment has not been implemented as intended then the proposed IV has not been manipulated. Within integrity are 3 related concepts: 1) adherence, or the degree to which the therapist implements the treatment as specified; 2) competence, the level of therapist skill with which the treatment is implemented; and 3) differentiation, the degree or dimensions along which the treatment is distinct from other

approaches in its implementation (Waltz, Addis, Koerner & Jacobson, 1993; Perepletchikova, Treat & Kazdin, 2007). Inherent in measuring fidelity in FAP is the capability to distinguish or differentiate FAP from comparison treatments, to establish items that include descriptors of behavior relevant to FAP in order to specify the non-overlapping aspects between itself and other conditions. Therapist competence is directly related to adherence: if a therapist does not adhere to the core tenets of a treatment then the treatment could not have been delivered with competence. That is, in order to be considered "competent" it is required that adherence be met; however, the reverse is not required to be true (i.e., to be adherent one does not necessarily require full competence – one can engage a required protocol behavior but do so incompetently). Adherence is often more easily measured because it typically focuses on capturing whether a protocol provision was implemented (e.g., psychoeducation) or how many times particular procedures were implemented. While competence is often a more global rating of therapist skillfulness (e.g., how well a treatment is implemented). Rounsaville, Carroll, and Onken (2001) developed a stage model for behavioral therapies to address treatment outcome research. Within this model it is recommended that several requirements be met prior to undertaking efficacy research, including establishment of measures to evaluate therapist adherence to manualized guidelines. Utilizing these established measures (e.g., therapy checklists and/or therapy rating scales), trained adherence coders rate tapes of therapy sessions to capture aspects of treatment integrity. These measures essentially serve as the manipulation check in the between-group (RCT) design study.

As a behavior analytic approach, FAP has often tested utilizing single-subject analyses to identify and code CRB, the conceptually specified dependent variable in FAP (Maitland & Gaynor, 2012). At present the primary coding system for this task is the FAP Rating Scale

(FAPRS; Callaghan & Follete, 2008), which relies on independent observers coding turn-by-turn interactions of FAP sessions. Theoretically, this method should be adequate to track both the dependent and independent variable (e.g., FAP's mechanism of action). That is, by collecting time series data on CRB, and therapist responses showing systematic changes in the frequency of the former based on introduction of the latter. However, the FAPRS rarely achieves reliability among observers (Follete & Bonow, 2009). Furthermore, when FAPRS has shown utility in previous studies it was implemented by researchers who: 1) had familiarity with the participants' case conceptualizations and therefore may be subject to experimenter bias, or 2) required significant, extensive training in the use of the FAPRS with the added training in the case conceptualization (Kanter et al., 2017). Additionally, this protocol itself is meticulously time-consuming as it requires the observation of both therapist and participant turn-by-turn behavior from start to finish of each session in the hopes of capturing the hypothesized mechanism of FAP, or the therapist contingent responses to desired in-session approximations toward target behavior and the subsequent increases in those CRB2 (Kanter et al., 2017). If reliability is established, the FAPRS may be beneficial as it offers evidence of FAP's mechanism, as well as providing information regarding therapist adherence and competence. However, the studies in which the FAPRS has been successfully utilized within the evidence-base were focused on single-case designs led by expert FAP therapists, limiting the generalizability of the outcomes, particularly for therapists of average to little FAP experience. Furthermore, the FAPRS does not include a metric to distinguish from other mechanisms (e.g., cognitive change) therefore missing an imperative component of treatment fidelity (Kanter et al., 2017). Often this strenuous and unreliable process deters researchers from attempting efficacy research, or at the very least limits

the ability to make inferences about the data collected. Without an adequate behavior analytic method to assess CRB, FAP single-subject research may remain limited.

Given these limitations with evaluating FAP's dependent and independent variable using the FAPRS, alternative techniques for measuring FAP adherence have been attempted. Kanter, Schildscrout, & Kohlenberg (2005) introduced an alternative technique to capture the in vivo nature of FAP by counting the frequency of in-session statements identified as "in vivo (IV) statements." IV statements were defined as "talk aimed at working on client problems that occurred in therapy in relation to the therapy process, the therapy relationship, and anything having to do with therapy." This approach was used to analyze data from Kohlenberg et al., (2002) who evaluated FAP-enhanced Cognitive Therapy (FECT) versus Cognitive Therapy (CT) alone finding that while clients responded well to both treatment approaches FECT was superior (remission rates of 79% and 60%, respectively). Kanter and colleagues evaluated whether FAP offered greater in vivo focus by examining the frequency of IV statements in both conditions. Trained undergraduate coders rated the frequency of IV statements and "other" statements in 59 randomly selected sessions from either the FECT or CT condition; to calculate reliability, coders observations were calculated into a proportion score of IV turns/total number of turns for each session and intraclass correlations were run. In addition to excellent reliability between coders (r = .97), results showed the FECT conditioned contained an average of 2.8 times the amount of IV statements than the CT condition. Consistent with the theory of FAP, an in vivo focus to the here-and-now of in session therapeutic relationship was more present within the FAP condition and may have contributed to differential outcomes; these frequency counts may then be capturing an important aspect of FAP (Kanter, Schildcrout, & Kohlenberg, 2005).

Although the method by Kanter and colleagues (2005) is less intensive, in terms of both training and coding, and likely captures the here-and-now (e.g., in vivo) aspect of FAP, there is little additional data on the use of this technique to substantiate treatment fidelity in FAP. To our knowledge this method has not been used in other studies and replication would be valuable. While the in vivo nature of FAP may be cornerstone for implementing techniques within FAP, it is unclear whether in vivo focus alone distinguishes FAP or is efficient for evaluating FAP's mechanism of action. While the in vivo focus is theoretically infused throughout FAP's fundamental techniques, such as the logical FAP interaction and the FAP 5-rules, there may be certain aspects of these techniques (e.g., specific rules) that demonstrate differential effects on outcomes, which may be valuable information for refinement of our conceptualization of FAP's mechanism of action. The frequency of in vivo counts lacks the specificity to provide information of this kind. For instance, the therapist may prompt discussion of the therapy relationship or attempt to evoke CRB but if the respective responses are not reinforced, FAP's mechanism was not engaged. As such, a fidelity measure that contains information relevant to different aspects of FAP would appear valuable for both fidelity assessment as well as for use in component, process, and mediational analyses. The current study proposes a measure that may have utility in both fidelity assessment and process analyses. Furthermore, as the current state of FAP literature lacks comparison of available adherence metrics, this study seeks to fill this gap via comparing the utility of various FAP treatment fidelity techniques.

Measurement Development

In prior work comparing FAP to SL in an alternating treatments design (Maitland & Gaynor, 2016) and RCT (Maitland et al., 2016) an adherence form was developed to measure whether elements of FAP and supportive listening occurred in session. As such this 10-item

measure was designed to try and capture both the occurrence of FAP and SL through 2 scales: the FAP scale and SL scale. The SL scale comprises the first four questions of the measure and was aimed at assessing for therapist behaviors associated with supportive listening, such as reflective and empathic responses, inquiring about daily social relationships, inquiring about feelings related to daily social relations, and attempts to understand the client's daily social behavior from the client's vantage point. An example item from the SL scale is: "Did the therapist engage in reflective and empathic listening in reaction to the client?" The FAP scale comprises the latter 5 questions and were designed based on the FAP 5-rules and the idea of a logical FAP interaction. An example item from the FAP scale is: "Did the therapist check with the client to see his/her response to the therapist sharing his/her reaction?" All items in both scales are rated by frequency between 0 and 3, where 0 = "zero occurrences of behavior" and 3 = "three or more occurrences of behavior." Coder instructions are also listed at the top of the form, and are summarized in the following statement: "read over every question before watching the session video, making notes as needed, rewinding as needed, and rate and score the adherence form at the end." The initial scale used in the studies of Maitland and colleagues had a rating maximum of 3+ occurrences. In the present study, coders complete the measure according to the instructions used in the Maitland studies but also using frequency counts of the number of observed instances for each FAP item. Thus, we could examine the values of the added information from frequency counts over the Likert-type rating in the original FAP scale. Also, because the frequency information was linked to specific FAP items it would offer added detail over the total IV frequency count method. Full measure is represented in Appendix A.

Preliminary Data from Two Studies

Maitland and Gaynor (2016) initially utilized the adherence measure for an alternating treatments design comparing up to 5 FAP sessions with up to 5 supportive listening sessions. Thirteen participants reporting difficulties in interpersonal relating received a total of 107 sessions and 80 were coded for elements of FAP and SL using the adherence measure. Results indicated that the SL scale average in SL and FAP sessions was significantly different, suggesting higher SL scores in SL sessions. Likewise, the difference in FAP scale average between SL and FAP sessions was highly statistically significant, suggesting higher FAP scores in FAP sessions. These data indicated that according to the adherence measure FAP sessions were distinctive from SL sessions from the perspective of a coder looking at the application of FAP rules focusing the therapeutic interaction on in session contingencies. That is, FAP sessions were unique from SL in offering the distinctive elements of FAP, and the critical independent variable was manipulated.

Similar data were found in a sample of 22 participants randomized to either six (45-60 minute) sessions of FAP or six (15 minute) sessions of watchful waiting (WW; Maitland et al., 2016). The WW condition comprised the same principles of empathic responding and a focus on interpersonal relationship without a directive change component. Utilizing the same adherence measure, we evaluated FAP as an IV for 33 randomly selected sessions. Twenty-two of these sessions were FAP, and half of the coded FAP sessions were coded for the first 15 minutes to account for time effects with the WW condition. The remaining 11 were watchful waiting sessions. Results indicated that a trend toward significance emerged for the difference between the WW mean and the first 15 minutes of FAP sessions on the SL scale. In regard to the FAP scale there was a statistically significant mean difference between WW and the first 15 minutes

of FAP sessions favoring FAP. Thus, the adherence measure suggested that while both conditions provided client-centered engagement, the FAP sessions were unique in their in vivo focus, the distinctive element of FAP. Thus, across two studies subsets of sessions the adherence measure supported the assertion that FAP was implemented and was different from the comparison condition. Since this measure was valuable for distinguishing FAP from SL conditions in prior studies, it may be a valuable tool for use in further efficacy research.

Hypotheses

The present work seeks to further examine this conclusion by applying the adherence measure to a larger sample of FAP sessions, collecting data using both Likert-type and frequency count ratings of FAP items, and comparing the total scores to those from the IV count method. Specifically proposed were three main hypotheses, containing sub-hypotheses within each. 1) The FAP Adherence Form (FAP-AF) will be a reliable measure of adherence and will demonstrate strong inter-observer agreement between multiple observers: a) according to total FAP Scale Likert-type ratings, b) frequency of FAP instances ratings, and c) on the individual items of the FAP scale. 2) The FAP scale of the FAP-AF will correlate: a) with an alternative measure of fidelity, specifically the Kanter et al., (2005) method of rate of In-vivo Turns counts; however, b) the FAP scale and IV Turns counts will not correlate with the SL scale. 3) Finally, it was hypothesized that this measure will correlate with and mediate outcomes found in Maitland et al., (2016) data, specifically the PDSQ as this relationship was previously found by the aforementioned researchers using preliminary coding data with one coder. If these mediational relationships are replicated, then this would suggest the FAP-AF might stand as a proxy for FAP's mechanism of action.

METHOD

Participants

Twenty-two participants were randomized to and completed six sessions of either watchful waiting (SL) or FAP. Eleven participants comprised the FAP condition with an average age of 20 (M = 20.27, SD = 2.90). Participants racial identity comprised eight White identifying, two African American, and one identifying as multi-racial. Of those, two indicated have a Hispanic background. Six of the eleven participants were male, with the remaining identifying as female. All participants were full-time students 5 of which were freshman, 5 sophomores, and 1 graduate student. The average GPA was a 3.26 (SD = .56) and ranged from 2.5-3.8. All participants denied history of substance use treatment. Participants in the FAP condition met a variety of DSM-4 diagnoses, including 9 for Social Anxiety Disorder, 8 for Avoidant Personality Disorder, 6 for Generalized Anxiety Disorder (GAD), and 3 for Major Depressive Disorder (MDD).

Eleven participants comprised the SL condition with an average age of 21 (M = 21.45, SD = 3.73). Five participants identified as male, and the remaining 6 as female. Participants' racial/ethnic identities were as follows: 10 identified as White, 1 was multi-racial, and 1 participant identified having a Hispanic identity. All but one participant was a full-time student, with the one participant indicating part-time status. Participants' average GPA was 3.08 (SD = .72) ranging from 2.00 to 3.99. The SL participants met criteria for a variety of DSM-IV conditions, including 8 for Social Anxiety Disorder, 5 for GAD, 2 for Avoidant PD, 1 for Panic Disorder, 1 for Binge Eating Disorder, and 1 for Agoraphobia.

Inclusion and exclusion criteria were the same criteria from the original RCT (Maitland et al., 2016). Inclusion criteria was based on two criteria. The first inclusion criterion required

that participant scores on the Fear of Intimacy Scale (FIS; Descutner & Thelen, 1991) were one standard deviation below their gender's mean score. The second inclusion criterion required that participants demonstrate a diagnosis of either Major Depressive Disorder, Social Anxiety Disorder, or Generalized Anxiety Disorder as determined by clinical interview and responses on the Psychiatric Diagnostic Screening Questionnaire (PDSQ; Zimmerman & Mattia, 1990), or demonstrate a diagnosis of Dependent Personality Disorder or Avoidant Personality Disorder, which was determined by the Structured Clinical Interview for DSM-IV TR (SCID-II; First, Gibbon, Spitzer, Williams, & Benjamin, 1997) (Maitland et al., 2016). Participants were excluded from the study based on three criteria: 1) if they met diagnostic criteria based on clinical interview and the PDSQ for PTSD, OCD, substance dependence, or psychosis; 2) if acute suicidal ideation was endorsed in the PDSQ and interview; and 3) if individuals interested in the study were determined to be already enrolled in psychotherapy services or had initiated psychotropic medication treatment in the previous 6-month period.

Therapists

The first author of Maitland et al., (2016) served as the primary therapist for the study. The therapist received extensive training in FAP prior to the start of the study, including an advanced 8-week training, and 4 two-day workshops led by FAP researchers/clinicians. In addition to completing a 200-hour doctoral practicum, the therapist served as a co-trainer for a two-day FAP training, and had previously served as the protocol therapist and principle researcher for a treatment outcome study in FAP. The secondary therapist in Maitland et al. (2016) was also a graduate student, completing a 200-hour practicum, and received training in the protocol by the primary therapist. The secondary author reviewed Tsai et al., (2009) and had previously attended a two-day FAP workshop.

Measurement Tools

Functional Analytic Psychotherapy Adherence Form (FAP-AF; Appendix A). This is a 10-item measure developed to evaluate the presence of statements consistent with the logical FAP interaction and FAP 5-rules. It is intended to evaluate FAP as an independent variable, to assess treatment fidelity, and as a possible process measure for mechanism research. Each item is rated between 0 and 3, where 0 indicates "zero occurrences of behavior," 1 is "one occurrence," 2 is "two occurrences," and 3 indicates "three or more occurrences of behavior." These ratings were initially chosen as arbitrary benchmarks. In this study and in previous studies (Maitland et al., 2016), item 10 is excluded from analysis due to experimenter error in the SL condition (item 10: "Did the therapist assign the client for homework to engage in specific out of session behaviors that followed from in-session interactions?"). Instructions for administration are provided at the top of the measure and are: "read over every question before watching the session video, making notes as needed, rewinding as needed, and rate and score the adherence form at the end." In this study, coders were asked to continue to count the frequency of each item beyond the 4-point Likert scaling in order to collect data for another metric of interest (i.e., *FAP Instances*) and to reserve the ability to expand or collapse the scale after empirical study. The measure is represented in Appendix A.

SL Scale. The supportive listening scale consists of the first 4-items of the FAP-AF. These items are intended to capture the basic aspects of supportive therapy with a focus on the client's interpersonal relationships. Items are rated as indicated above. Item 1 reads: "To what extent was the therapist's behavior mainly directed toward attempts to understand the daily life social relationships from the client's vantage point?" Item 2 reads: "Did the therapist engage in reflective and empathic listening in reaction to the client?" Item 3 reads: "Did the therapist

prompt/encourage the client to discuss daily life social relations?" Item 4 reads: "Did the therapist turn the focus of the session on the client's feelings/emotional reactions to events in his/her daily life social relations?" The scale was calculated by calculating a total score for the 4-items in each session and obtaining the average across the total number of sessions coded per participant. This was then completed for each coder.

FAP Scale. The FAP scale consists of the last 5-items of the FAP-AF, beginning with item 5 and ending with item 9 (excluding item 10). Each item was derived from one of the FAP 5-rules, with the presumption that a FAP-consistent therapist would follow a format or style similar to the logical FAP interaction as proposed by Weeks et al., (2012). The items were designed to focus solely on therapist behavior and broad enough to capture FAP principles regardless of functional assessment. Items are rated as indicated above. Item 5 was intended to map onto rule 1 (e.g., watch CRB) and reads: "Did the therapist turn the focus of the session on the clients in-session behavior?" Item 6 was intended to map on to both rules 1 and 2 (e.g., evoke CRB) and perform as a measure of an out-to-in parallel. Item 6 reads: "Did the therapist compare in-session events to the participant's daily life?" Item 7 was intended to map on to rule 2; it reads: "Did the therapist prompt/encourage the client to engage in particular responses in the session?" Item 8 was intended to map on to rule 3 (e.g., reinforce CRB2), and it reads: "Did the therapist share his/her reaction to the client's in session behavior?" Item 9 was intended to map on to rule 4 (e.g., check for therapist effects); it reads: "Did the therapist check with the participant to see his/her response to the therapist sharing his/her reaction?" The scale was calculated by summing a total score for the 5-items for each session and obtaining the average across the total number of sessions coded per participant. This was then completed for each coder.

FAP Instances (Appendix A). This approach involved coding of the FAP-AF beyond the limit of item scaling (i.e., 0-3). Coders were asked to continue monitoring therapist behavior for items 5-9 on the FAP-AF for the duration of each session until session termination (approximately 15 for the WW condition or 60 minutes for the FAP condition). FAP instances were calculated by acquiring sums for each item, then taking the item sums of items 5-9 for each session, and finally obtaining the average across the total number of sessions coded for each participant. This was done for each coder.

IV Turns Count (Kanter, Schildcrout, & Kohlenberg, 2005; Appendix B). This approach involves coding the frequency of in vivo turns or statements made by the therapist in 5-minute intervals. In vivo turns are defined as "talk aimed at working on client problems that occurred in therapy in relation to the therapy process, the therapy relationship, and anything having to do with therapy." IV turns count were calculated in IV rate/min. This rate was compared with ratings on the 4 SL and 5 FAP items of the FAP-AF. This method is referred to as the IV turns count throughout the paper.

Psychiatric Diagnostic Screening Questionnaire (PDSQ; Zimmerman & Mattia, 2001; Appendix C). This is a 125-item self-report measure that was administered during Maitland et al., (2016) original study to screen for DSM-IV-TR Axis I disorders and to examine symptomatic changes following treatment. The measure evaluates one's endorsement of psychiatric symptoms across 13 diagnostic categories, including MDD, PTSD, GAD, OCD, Bulimia, Panic Disorder, Agoraphobia, Social Anxiety Disorder, Alcohol Abuse, Drug Abuse, Somatization, Hypochondriasis, and Psychosis. The subscales representing these categories have good to excellent internal consistency, ranging from Cronbach's alpha of .66 to .94. The PDSQ

was administered at pre-, post-, and follow-up. Data from the PDSQ was utilized from pre- and post-treatment only.

PDSQ residualized change scores. Residualized change scores from the PDSQ were used to examine the fitness of the mediational model relevant to the adherence metrics. Residualized change scores were calculated by regressing post-treatment PDSQ scores on pre-treatment PDSQ scores.

PDSQ difference scores. The difference of post-treatment PDSQ scores and pre-treatment PDSQ scores was calculated. Scores were utilized to visually represent the relationship between treatment outcome (PDSQ difference scores) and the hypothesized mediators (i.e., FAP scale, FAP instances, and IV turns count).

Fear of Intimacy Scale (FIS; Descutner, C. J., & Thelen, M. H. 1991; Appendix D). The FIS is a 35-item scale administered in Maitland et al., (2016) study as a measure of one's fear and experience with interpersonal intimacy, self-disclosure, and social desirability. The FIS has high internal consistency with a Cronbach's alpha of .93 and good test-retest reliability (r = .89, $p < .001$). This measure was administered as a primary outcome measure of social intimacy and was examined at pre-, post-treatment, and follow-up. FIS residualized change scores were calculated in the current study to evaluate the fitness of the mediational model relevant to the adherence metrics.

Data Collection

Data was collected from Maitland et al., (2016) from every session provided to all 22 participants in both the FAP and WW condition. Sessions were randomized first by participant then by order of each session. Three coders were trained in the use of the FAP-AF, FAP

instances, and IV turns count. All three coders evaluated 32 (25%) sessions, two evaluated 64 (50%) sessions, and one coder (the author) evaluated 132 (100%) sessions. Coders observed each session within 5-minute intervals for the duration of the session (15 minutes for WW condition and 45-60 minutes for FAP).

Coders and Training

A total of three coders were trained in the use of the three metrics of adherence: the FAP-AF, FAP instances total, and the IV turns count. Coders were graduate students in clinical psychology with at least a basic introduction to FAP in both coursework and clinical application. Coders were trained prior to data collection. Training consisted of reading and discussing several articles pertaining to FAP therapy implementation (Weeks and colleagues, 2012; Tsai et al., 2009). Coders then watched sessions from Maitland & Gaynor (2016), the alternating treatment design study, and practiced completing all three metrics of adherence. The first 2 sessions (1 FAP and 1 SL) were viewed as a group and included discussion on observations and scoring. The next 6 sessions (4 FAP and 2 SL) were coded independently and later items were discussed when significant disagreement was found. Then the three coders independently coded the same randomly selected 32 sessions of either FAP or WW. Two of the three coders coded an additional 32 randomly selected sessions. Finally, one coder coded all sessions in random order.

Data Analytic Plan

To assess interobserver agreement, multiple methods were used. Intraclass correlations (ICC) were used as a first pass at examining agreement between coders, followed by percent agreement. When a study design calls for multiple coders to rate data from multiple subjects, oftentimes ICCs are used to evaluate agreement between coders. There are several important

considerations when conducting ICC analyses (Hallgren, 2012). First, it is important to specify whether the model will be one-way or two-way based on the selection of coders. If coders are randomly selected from a larger population of coders than the model would be designated as a "one-way" model; whereas, if coders are selected as the sole represented coders then the model is termed "two-way". The second step is to decide whether agreement is best captured by absolute agreement or based on consistency of coder ratings. Third, the quantification of reliability of coder ratings must be specified. Typically, this is either simply the average of all coder ratings (specified by the term "average") or the ratings provided by a single coder (specified by the term "single"). Finally, it is important to specify whether the outcomes from the analyses are meant to generalize to a larger population that it was drawn from or if it is meant simply to speak to the effect of the coders of the actual study. These effects would be identified as either random or fixed effects, respectively (Hallgren, 2012).

In the present study, a two-way random effects, absolute agreement, averaged ICC was selected to evaluate the reliability of coders' observations of the study therapist's behavior across sessions. Utilizing a research randomizer tool, we determined the order of coding by first randomizing the participant order, then randomizing the sessions. The first 24% of randomly selected sessions (n = 32) were assigned to be coded by all three coders. Of the remaining 76% of sessions, 2 of the 3 coders then continued to code a randomly selected set of sessions (n = 36, 27%), totaling to 51% of sessions for the two of three coders. Finally, the study author continued to code the remaining sessions in the original randomly selected order until all sessions were rated (n = 132, 100%). The following scaling was used to determine strength of ICC reliability: poor → 0 – 0.4; fair → 0.4 – 0.59; good → 0.6 – 0.74; excellent → 0.75 – 1.00 (Cicchetti & Sparrow, 1981).

For individual item agreement, both ICC and percent agreement methods were employed in order to evaluate inter-observer agreement. Percent agreement was calculated by taking the total number of coder agreements for a given item and dividing it by the number of total observations. This was utilized for coders 1 and 2 who for 38 FAP sessions. This method was a traditional method for determining inter-observer reliability and was employed in this study in an effort to sidestep the potential obstacles to agreement related to the scaling of our measure (e.g., ceiling effect and truncated range).

Pearson's correlations were employed to evaluate the relationship between coders (e.g., coder 1 scores and coder 2 scores) and between the measures (i.e., the scale scores of the FAP-AF, FAP instances, and the FAP IV turns count). Hayes' (2013) PROCESS procedure was utilized to test simple mediation models between the FAP scale, FAP instances, and FAP IV turns counts and outcome measures from Maitland et al., (2016). Prior to this procedure, a three-step process was employed to prepare and evaluate the mediational model. Step one, utilized t-tests to distinguish our IV (FAP scale) from the comparison group (SL). Step two, engaged a Pearson's correlation to examine the relationships between the outcome measures of interest (i.e., FIS and PDSQ) with the adherence metrics (FAP scale, FAP instances, and IV turns count). Finally, step three engaged the Hayes' (2013) PROCESS procedure to evaluate the indirect path (e.g., treatment → mediator → change) compared to the direct path (treatment → change) as a test of mediation between the two outcome measures and the adherence metrics (the proposed mediators of outcome). Mediation is specified when a significant indirect effect ($p < .05$) is indicated by a point estimate with a 95% bias corrected confidence interval that did not include zero (Hayes, 2013; Preacher & Hayes, 2008).

RESULTS

Interrater Reliability

Inter-rater reliability was assessed using Intraclass Correlation Coefficients (ICCs). ICCs are applicable with data coming from two or more raters each providing ratings on multiple participants. A two-way random model, specifying absolute agreement (so to be sensitive to systematic differences between raters), was selected for each ICC. ICCs examined agreement between the (2 or 3) coder ratings on each of the three adherence metrics - the FAP scale, FAP instances, and/or frequency of In Vivo turns count. The first set of analyses examined agreement across the full sample (i.e., ratings of sessions from those receiving either FAP or SL, the FAP+SL sample) and the second set of analyses examined agreement only among sessions of participants receiving FAP. It was important to conduct both sets of analyses because little or no FAP was expected to occur in SL sessions (Maitland et al., 2016); thereby potentially producing high agreement (on the absence of FAP), that might lead to the first set of ICCs being inflated. Therefore, the second set of ICCs included only on the sample receiving FAP to explicitly examine agreement on the presence of FAP. Table 1 presents the sample sizes, ICCs, corresponding 95% confidence intervals, means, and standard deviations from each of the relevant variables across both sets of analyses.

Set 1 Analyses: Reliability of the FAP scale, FAP instances total, and frequency of IV turns count in the FAP+SL sample. Three coders rated 32 sessions using the FAP scale yielding an *ICC* of .98 ($p < .001$) and suggesting excellent agreement, a result which was replicated using data from two coders rating 68 sessions, *ICC* = .98, $p < .001$). When FAP instances were recorded, rather than a Likert scale rating, the resulting *ICCs* were .88, $p < .001$ for the 3 coders and .90, $p < .001$ for the two coders, numerically smaller than the FAP scale

Table 1

Intra-Class Correlations, Confidence Intervals, and Coder Means of FAP Scale Scores, Total Instances Scores, and IV Turns Count Method.

	ICC	R1 mean/sd	R2 mean/sd	R3 mean/sd
FAP+SL Sample				
FAP Scale 3 coders (n = 32)	.98 [a]	6.19 (4.90)	6.06 (4.85)	6.53 (5.04)
FAP Scale 2 coders (n = 68)	.98 [a]	5.10 (4.82)	4.99 (4.77)	
FAP Instances 3 coders (n = 32)	.88 [a]	11.67 (11.13)	15.5 (13.42)	17.56 (16.18)
FAP Instances 2 coders (n = 68)	.90 [a]	9.44 (10.41)	12.40 (12.75)	
IV Turns Count 2 coders (n = 68)	.99 [a]	13.76 (15.22)	13.03 (14.29)	
FAP Sample				
FAP Scale 3 coders (n = 21)	.82 [a]	9.43 (2.25)	9.24 (2.34)	9.95 (1.91)
FAP Scale 2 coders (n = 38)	.82 [a]	9.11 (2.19)	8.92 (2.83)	
FAP Instances 3 coders (n = 21)	.70 [b]	17.76 (8.84)	23.62 (8.81)	26.76 (12.13)
FAP Instances 2 coders (n = 38)	.71 [b]	16.87 (8.23)	22.18 (8.46)	
IV Turns Count 2 coders (n = 38)	.99 [a]	24.61 (12.08)	23.32 (11.12)	

[a] = excellent
[b] = good
[c] = fair

(Cicchetti & Sparrow, 1981) ICC: 0.00–0.40 = poor, 0.40–0.59 = fair, 0.60–0.74 = good, and 0.75–1.00= excellent."

results, but also suggestive of excellent agreement. Finally, when the frequency of in vivo turns was counted, the ICC for the two coders was also excellent, *ICC* = .99, $p < .001$).

Set 2 Analyses: Reliability of the FAP scale and FAP instances total in the FAP subsample. Three coders rated 21 sessions using the FAP scale yielding an ICC of .82 ($p <$ .001) and indicating excellent agreement, a result which was replicated using data from two coders rating 38 sessions, *ICC* = .82, $p < .001$). While the FAP scale ICCs were in the excellent range according to commonly used interpretive standards (e.g., Cicchetti, 1994) their strength was reduced from .98 in the FAP+SL sample to .82 in the FAP sample. When FAP instances were recorded, rather than a Likert scale rating, the resulting *ICCs* were .70, $p < .001$ for the 3 coders and .71, $p < .001$ for the two coders and suggestive of good agreement. Thus, the FAP

instances ICCs were reduced from the excellent range when using the FAP+SL sample to the good range in the FAP subsample.

Total scores from session ratings of coders using the FAP scale, tallies of FAP instances, or frequency counts of in vivo turns, all indicated excellent-good (and mostly excellent) inter-rater agreement and did so regardless of whether data from two or three coders were included in the analyses. In short, the overall impressions of the coders were in agreement as to what they saw in the session tapes.

Item Analyses ICC. In order to examine inter-observer reliability on each item, ICCs were run on each of the FAP scale items completed by the two coders rating the 38 FAP sessions. The ICCs here were variable and several were much worse than when examining the overall scale. For instance, items 5 (Did the therapist turn the focus of the session on the clients in-session behavior?) and 6 (Did the therapist compare in-session events to the participants daily life?) fell within the poor range (*ICC*s = .00 and .35, respectively), while item 8 (Did the therapist share his/her reaction to the clients behavior?) was within the good range (*ICC* = .66), and items 7 (Did the therapist prompt/encourage the client to engage in particular responses in the session?) and 9 (Did the therapist check with the participant to see his/her response to the therapist sharing his/her reaction?) were within the excellent range (*ICC*s = .90 and .75, respectively). Examination of Table 2 illustrates why these results appeared to occur. The FAP scale was coded on a 0-3 scale. When the means were high and the standard deviation was low (items 5 and 6), the resulting ceiling effect and truncation of range left no variability to be identified in the ICC analyses and they were poor. When the means were lower and the standard deviations higher (items 7 and 9) the ICCs were excellent. The means and standard deviations for item 8 fell in between and the corresponding ICC was good.

Table 2
FAP Scale Item Score Intra-Class Correlations, Confidence Intervals, and Rater Means for Coders 1 and 2 (n = 38)

	ICC	R1 mean/sd	R2 mean/sd
Item 5	.00	2.92 (.36)	3.00 (.00)
Item 6	.35	2.53 (.89)	2.74 (.55)
Item 7	.90 [a]	0.58 (.98)	0.68 (1.04)
Item 8	.66 [b]	2.50 (.98)	2.55 (.86)
Item 9	.75 [a]	0.29 (.52)	0.37 (.67)

[a] = excellent
[b] = good
[c] = fair

The FAP instances data were not constrained at the upper limit and, hence, allowed for more variability in the data such that (if the aforementioned interpretation in terms of ceiling effects and truncation of range was at least partially correct) the ICCs should be higher in these analyses. Indeed, this was generally the case as agreement was good for 3 items (items 6,8,9), excellent for 1 (item 7), and poor for only 1 (item 5; see Table 3). Most of the items again were lower than of the ICCs established when looking at scale totals rather than items. It seems important that the item mean scores of each FAP instances item followed the exact same pattern as was found when using the FAP scale Likert ratings. For instance, item 5, the item that had a poor ICC in both analyses, had the highest mean score each time.

To further examine the suggestion that the variable ICC results were influenced, at least in part, by ceiling effects and range restriction, exact, $\pm$ 1, and $\pm$ 2 agreement percentages were calculated for each item. Item percent agreement are represented in figures 1 – 5. Percent agreement was obtained by dividing the total number of coder agreements by the number of total

Table 3

FAP Scale Item Instances Data Intra-Class Correlations, Confidence Intervals, and Coder Means for Coders 1 and 2 (n = 38)

	ICC	R1 mean/sd	R2 mean/sd
Item 5	.36	7.00 (3.86)	11.79 (4.84)
Item 6	.71 [b]	4.58 (2.40)	4.71 (2.93)
Item 7	.93 [a]	1.26 (3.46)	0.87 (2.55)
Item 8	.65 [b]	3.74 (2.51)	4.47 (2.71)
Item 9	.70 [b]	0.37 (.67)	0.34 (.58)

[a] = excellent
[b] = good
[c] = fair

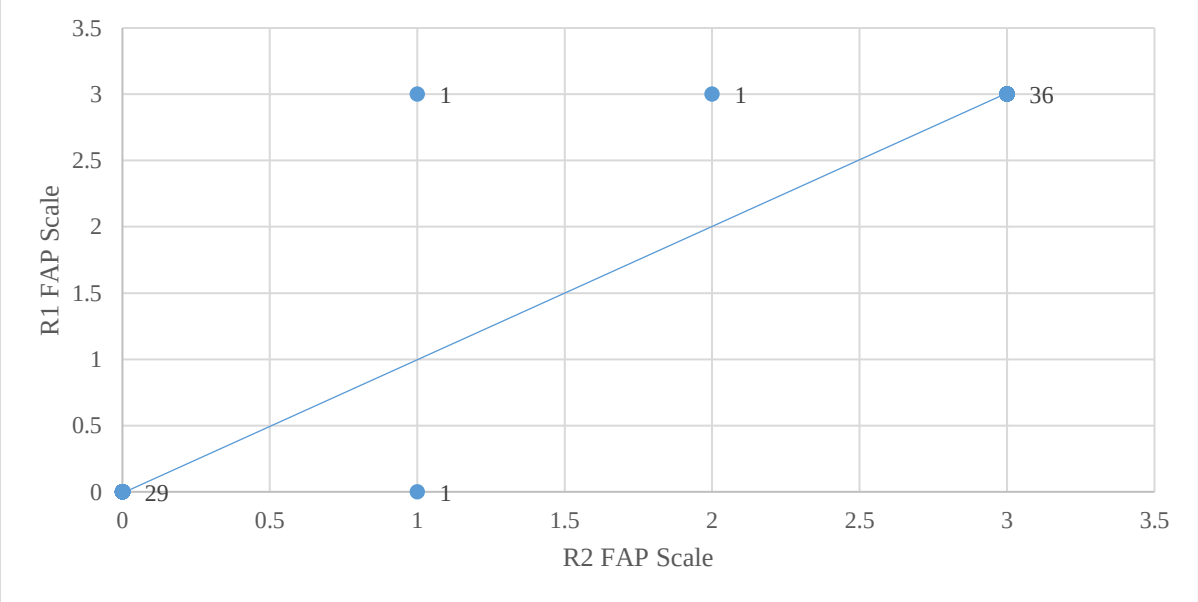

Figure 1.[1]Item 5 Percent Agreement

[1] For figures 1 – 5
* Blue line represents the line of perfect agreement
** Labels on each scatterplot point represent the total number of observations which fit this particular point of agreement

observations using each of the FAP scale items completed by the two coders rating 38 FAP sessions. Exact agreement was 95%, 79%, 87%, 71%, and 90% for items 5-9, respectively, indicating strong agreement for each item. Notice particularly that for item 5, the item with the

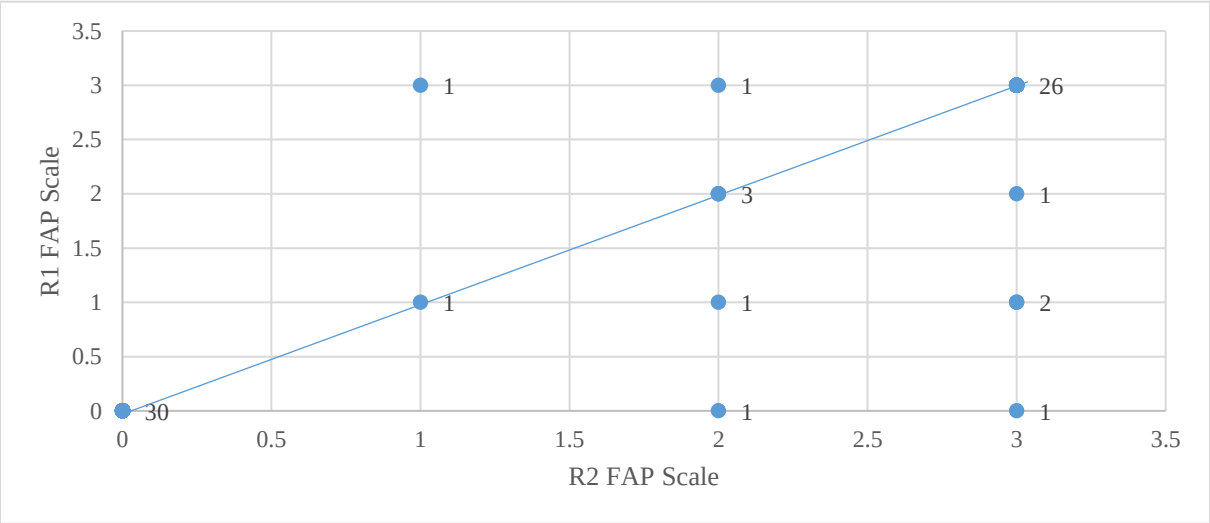

Figure 2. Item 6 Percent Agreement

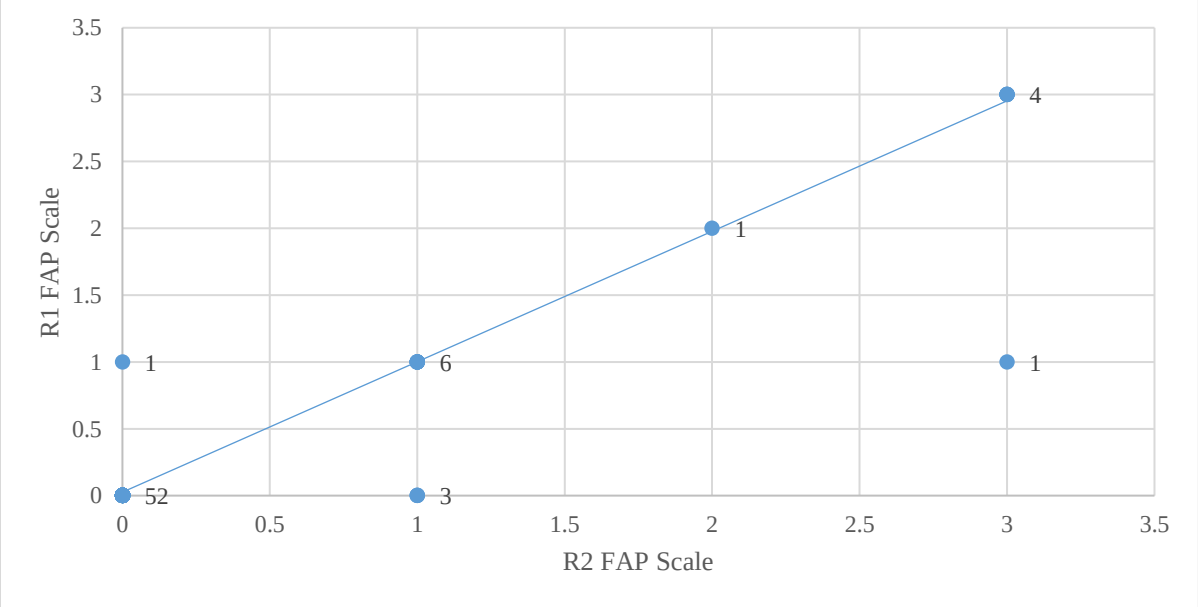

Figure 3. Item 7 Percent Agreement

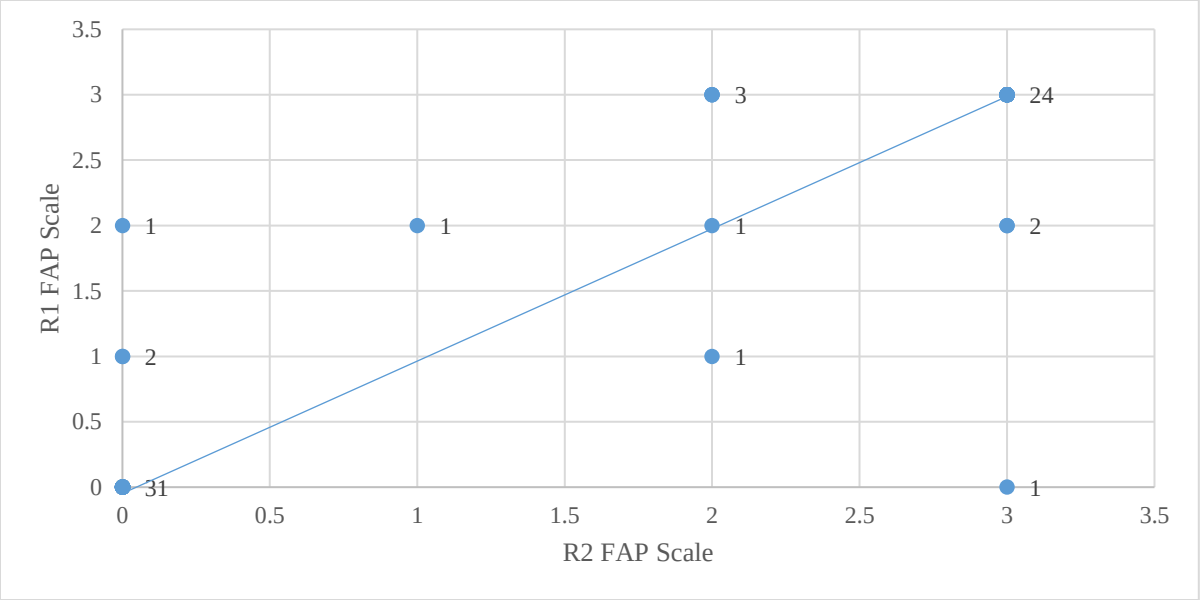

Figure 4. Item 8 Percent Agreement

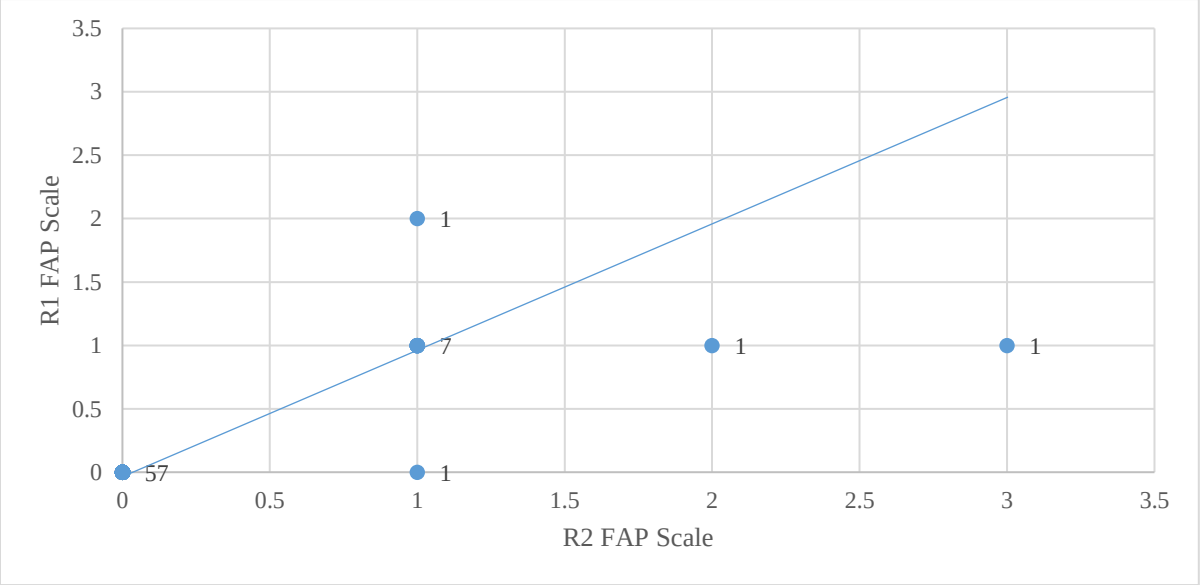

Figure 5. Item 9 Percent Agreement

lowest ICCs, the exact agreement was the highest (95%). When agreement was calculated in a manner that allowed coders to vary by one point from each other agreement ranged from 87-97% across items and was 97-100% when coder ratings were allowed to fall within 2 points of each

other (see Table 4). These data suggest that coders strongly agreed on what they saw in the tapes at the level of the individual FAP scale item.

Inter-method Reliability. Intraclass correlations comparing coder total scores on the FAP scale, of FAP instances, and the frequency of IV turns count, all showed good to excellent (and mostly excellent) agreement *within* scales (e.g., coder 1s FAP scale with coder 2s FAP scale). The next series of analyses examined the relationships *between* the methods of assessing adherence (e.g., coder 1s FAP scale with coder 1s and 2s IV turns count). If all of these are

Table 4
FAP Scale Item Percent Agreement (n = 38)

	Perfect	**± 1**	**±2**	**±3**
Item 5	36/38 = 94.74%	37/38 = 97.37%		
Item 6	30/38 = 78.95%	33/38 = 86.84%	37/38 = 97.37%	
Item 7	33/38 = 86.84%	37/38 = 97.37%	38/38 = 100%	
Item 8	27/38 = 71.05%	36/38 = 94.74%	37/38 = 97.37%	
Item 9	34/38 = 89.47%	37/38 = 97.37%		

roughly equal measures of adherence to FAP they would be expected to correlate highly. This should occur within coders (i.e., coder 1s FAP scale score should be strongly related to her FAP instances, and frequency of IV turns count) and between coders (i.e., coder 1s FAP scale score should be strongly related to the other coders FAP instances, and frequency of IV turns count.)

***Within coders relationship between coding methods*.** The within coder correlations between the FAP scale, FAP Instances, and IV turns counts were all strong and positive, ranging from $r = .73$ to $r = .96$. (see Tables 5, 6, & 7).

Table 5

Correlations for 3 Coders Across the 3 Methods: FAP Scale Scores, FAP Instances Scores, and IV Turns Count. (FAP+SL n = 32 sessions)

	FS1	FS2	FS3	FI1	FI2	FI3	IV1	IV2	IV3
FAP R1 Scale (FS1)	1								
FAP R2 Scale (FS2)	.99**								
FAP R3 Scale (FS3)	.99**	.96**							
FAP R1 Instances (FI1)	.92**	.92**	.87**						
FAP R2 Instances (FI2)	.94**	.96**	.92**	.94**					
FAP R3 Instances (FI3)	.93**	.91**	.90**	.95**	.96**				
IV Turns Count R1 (IV1)	.93**	.91**	.89**	.94**	.95**	.98**			
IV Turns Count R2 (IV2)	.92**	.91**	.88**	.95**	.94**	.98**	.99**		
IV Turns Count R3 (IV3)	.93**	.91**	.89**	.94**	.95**	.99**	.99**	.99**	1

* $p < .05$

** $p < .01$

Table 6

Correlations for 3 Coders Across the 3 Methods: FAP Scale Scores, FAP Instances Scores, and IV Turns Count. (FAP Sample n = 21 sessions)

	FS1	FS2	FS3	FI1	FI2	FI3	IV1	IV2	IV3
FAP R1 Scale (FS1)	1								
FAP R2 Scale (FS2)	.90**								
FAP R3 Scale (FS3)	.89**	.74**							
FAP R1 Instances (FI1)	.85**	.83**	.71**						
FAP R2 Instances (FI2)	.78**	.80**	.66**	.85**					
FAP R3 Instances (FI3)	.82**	.76**	.75**	.87**	.88**				
IV Turns Count R1 (IV1)	.87**	.80**	.78**	.86**	.88**	.95**			
IV Turns Count R2 (IV2)	.86**	.81**	.77**	.87**	.86**	.95**	.99**		
IV Turns Count R3 (IV3)	.85**	.78**	.77**	.86**	.88**	.98**	.99**	.99**	1

* $p < .05$

** $p < .01$

Consistent with the logic explicated above, coder 1's FAP scale score significantly correlated to other rating methods made by coder 1, such as the FAP instances scores ($r = .92$, $p < .001$) and IV turns count ($r = .93$, $p < .001$). Coder 2's FAP scale data also significantly correlated with coder 2's own observations on FAP instances ($r = .96$, $p < .001$) and on IV turns

Table 7

Correlations for 2 Coders across the 3 methods: FAP scale scores, FAP instances scores, and IV turns count. (FAP Sample n = 38 sessions)

	FS1	FS2	FI1	FI2	IV1	IV2
FAP R1 Scale (FS1)	1					
FAP R2 Scale (FS2)	.82**					
FAP R1 Instances (FI1)	.82**	.72**				
FAP R2 Instances (FI2)	.79**	.78**	.85**			
IV Turns Count R1 (IV1)	.80**	.69**	.88**	.89**		
IV Turns Count R2 (IV2)	.80**	.73**	.85**	.89**	.97**	1

* $p < .05$

** $p < .01$

count ($r = .91$, $p < .001$). Finally, coder 3's FAP scale significantly predicted strong positive relationships with her FAP instances ($r = .90$, $p < .01$) and her IV turns count ($r = .89$, $p < .01$).

When the data was isolated to the FAP sample alone, coder 1's FAP scale score significantly predicted her FAP instances scores ($r = .82$, $p < .001$) and IV turns count ($r = .80$, $p < .001$). Similarly, Coder 2's FAP scale data also significantly correlated with coder 2's own observations on FAP instances ($r = .78$, $p < .001$) and on IV turns count ($r = .73$, $p < .001$), as did coder 3's FAP scale with her FAP instances ($r = .75$, $p < .01$) and her IV turns count ($r = .77$, $p < .01$). While these results all continued to be positive correlations, the strength of the relationships of the within coder relationships in the FAP condition alone were lower than for the FAP+SL subsample. As the FAP sessions were longer in duration than the SL sessions (60 mins vs. 15 mins) which may have contributed to the greater variability in observations.

Overall, a coder's score on one of the adherence metrics (e.g., the FAP scale) correlated highly with scores on the other two metrics (e.g., FAP instances and IV turn counts).

Between coders relationship between coding methods. Results for correlations between methods were all strong, positive correlations ranging from $r = .87$, $p < .001$ to $r = .99$, $p < .001$ for both samples (see Table 5). Coder 1's FAP scale predicted both Coder 2 and Coder 3's FAP

instances (r = .94, p < .001 and r = .93, p < .001, respectively) and IV turns count (r = .92, p < .001 and r = .93, p < .001, respectively). Coder 2's FAP scale also predicted both Coder 1 and 3's FAP instances (r = .92, p < .001 and r = .91, p < .001, respectively) and IV turns count (r = .91, p < .001 and r = .91, p < .001, respectively) in a strong positive direction. Finally, Coder 3's FAP scale, while still strong positive correlations was the least predictive of Coder 1 and 2's FAP instances (r = .87, p < .001 and r = .92, p < .001, respectively) and IV turns count (r = .89, p < .001 and r = .88, p < .001, respectively).

In the FAP sample, the range of correlations fell between moderate to strong, positive correlates of r = .66 to r = .98 (see Table 6). Coder 1's FAP scale predicted both Coder 2 and Coder 3's FAP instances (r = .78, p < .01 and r = .82, p < .01, respectively) and IV turns count (r = .86, p < .001 and r = .85, p < .01, respectively). Coder 2's FAP scale also predicted both Coder 1 and 3's FAP instances (r = .83, p < .01 and r = .76, p < .01, respectively) and IV turns count (r = .80, p < .01 and r = .78, p < .01, respectively) in a strong positive direction. Finally, Coder 3's FAP scale, while still moderate positive correlations was the least predictive of Coder 1 and 2's FAP instances (r = .71, p < .01 and r = .66, p < .01, respectively) and IV turns count (r = .78, p < .01 and r = .77, p < .01, respectively).

When comparing Coders 1 and 2 extended number of sessions (n sessions = 38), Coder 1's FAP scale total score was significantly correlated with Coder 2's FAP scale score (r = .83, p < .001), instances scores (r = .79, p < .001), and IV turns count (r = .80, p < .001). Similarly, Coder 2's FAP scale score significantly correlated with Coder 1's instances score (r = .72, p < .001) and IV turns count (r = .69, p < .001). See Table 7 for correlations between Coder 1 and 2.

Regardless of the coder, the FAP scale consistently predicted the other coder's FAP scale, IV turns count, and FAP instances scores.

Process of Change

Having established that the coders were in agreement within and across adherence coding metrics, an examination of whether the adherence coding metrics could contribute to establishing FAP as an independent variable in efficacy research and our understanding of the change process in FAP was examined.

In Maitland et al. (2016), 11 WW and 11 FAP sessions, one session for each participant chosen at random, were coded on both the FAP and SL scales. The FAP scale scores were significantly correlated with residualized change scores on both primary outcome measures: the FIS (r = .47) and the PDSQ (r = .70). Moreover, using Hayes' (2013) PROCESS procedure for SPSS, the FAP scale score served as cross-sectional statistical mediator of change on the PDSQ, but not the FIS. Maitland et al. (2016) suggested the possibility of coding all the sessions, with the prediction that the FAP scale might then be expected to serve as an even more robust mediator. The following results examined that prediction using data from Coder 1 who coded all the sessions.

The first step was to document that the FAP codes readily differentiated the two treatments from one another. That is, that the FAP scale, FAP instances total, and frequency of IV turns count would verify the manipulation of the independent variable. A t-test comparing the FAP scale score means of those who received FAP (M = 9.06, SD = 1.13) to those who received SL (M = 0.05, SD = 0.08) were statistically significantly different, t = 26.31, p < .001. Likewise, t-tests comparing the FAP instances total means of those who received FAP (M = 16.58, SD =5.44) to those who received SL (M = 0.05, SD = 0.08) were also statistically significantly different, t = 10.07, p < .001, as were those comparing the frequency of IV turns means of those who received FAP (M = 23.69, SD = 7.89) to those who received SL (M = 0.05, SD =0.08), t =

9.93, $p < .001$. All three of the variables of interest -- the FAP scale, FAP instances, and/or frequency of IV turns count – readily distinguished the presence of FAP from its absence.

The second step was to examine the relationships between the residualized change scores on the FIS and PDSQ and the various FAP adherence metrics across the full sample (N = 22). Significant correlations were found between residualized change scores on the FIS and the FAP scale (r = -.44, p = .04), FAP instances total (r = -.44, p = .04), and frequency of IV turns count (r = -.46, p = .03). Importantly, FIS change was not significantly related to either the SL scale (r = -.18, p = .42) or the SL instances total (r = -.32, p = .15). The pattern of results was the same when residualized change scores on the PDSQ was used. Significant correlations were found between PDSQ change and the FAP scale (r = -.52, p = .01), FAP instances total (r = -.52, p = .01), and frequency of IV turns count (r = -.58, p = .01), while PDSQ change was not significantly related to either the SL scale (r = -.16, p = .47) or the SL instances total (r = -.37, p = .09). All three of the metrics of adherence to FAP -- the FAP scale, FAP instances, and/or frequency of IV turns count – were significantly associated with clinical improvements.

The third step was to determine if the FAP adherence metrics would serve as statistical mediators of change. Six individual analyses examined whether any of the FAP adherence metrics served as statistical mediators of FIS or PDSQ change. The test of mediation examined the significance of the indirect path (e.g., treatment → mediator → change) compared to the direct path (treatment → change). A point estimate, based on 10,000 bootstrapped samples, with a 95% bias corrected confidence interval that did not include zero, was used to indicate a significant indirect effect ($p < .05$) suggesting mediation (Hayes, 2013; Preacher & Hayes, 2008). Initial mediator analyses were run examining the residualized FIS change scores. The indirect effect of condition on residualized FIS change through the FAP scale was not

statistically significant (point estimate = 0.06, $p > .05$; 95% confidence interval [CI] [-4.34, 4.58]). See Figure 6 for an example of the mediational model. Similarly, the indirect effect of condition on residualized FIS change through the FAP instances total was not statistically significant (point estimate = -0.35, $p > .05$; 95% confidence interval [CI] [-2.41, 1.32]). Finally, the indirect effect of condition on residualized FIS change through the IV turns count was not statistically significant (point estimate = -0.56, $p > .05$; 95% confidence interval [CI] [-2.30, 1.29]). Next, mediator analyses were run examining the residualized PDSQ change scores. The indirect effect of condition on residualized PDSQ change through the FAP scale was not statistically significant (point estimate = -0.56, $p > .05$; 95% confidence interval [CI] [-6.56, 3.63]; see Figure 7). Furthermore, the indirect effect of condition on residualized PDSQ change through the FAP instances score was not statistically significant (point estimate = -0.51, $p > .05$; 95% confidence interval [CI] [-2.49, 1.24]; see Figure 8). Finally, the indirect effect of condition on residualized PDSQ change through the IV turns count was not statistically significant (point estimate = -1.06, $p > .05$; 95% confidence interval [CI] [-3.19, .56]; see Figure 9).

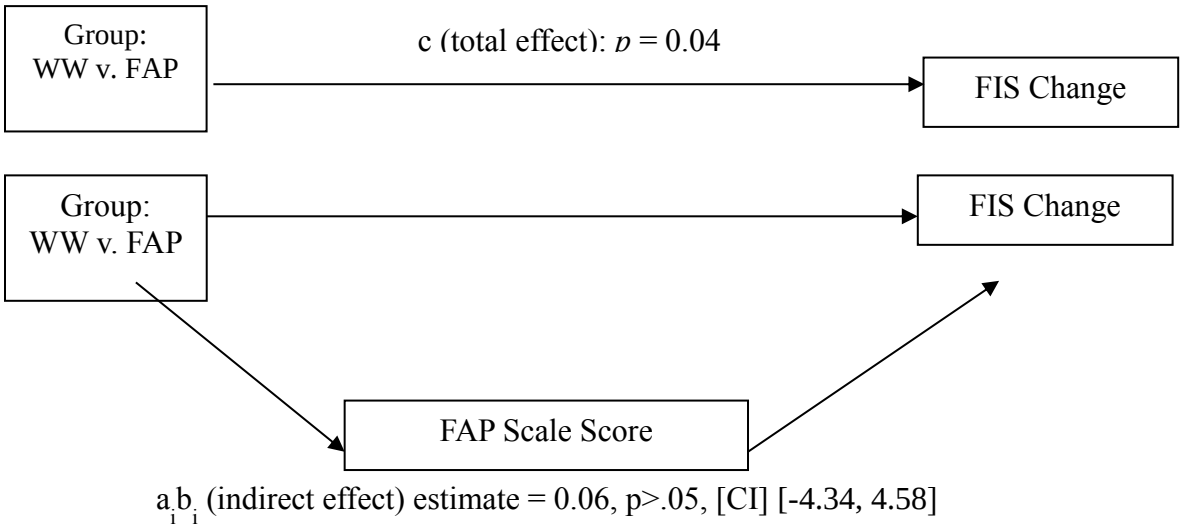

Figure 6. Mediational Model

Scatterplots

Scatterplots showing the relationship between the mean FAP scale, FAP instances, an IV turns scores and the amount of symptomatic change on the PDSQ are presented. See Figures 7, 8, and 9[2]. FAP condition participants are represented by black circles and SL participants are represented by grey circles.

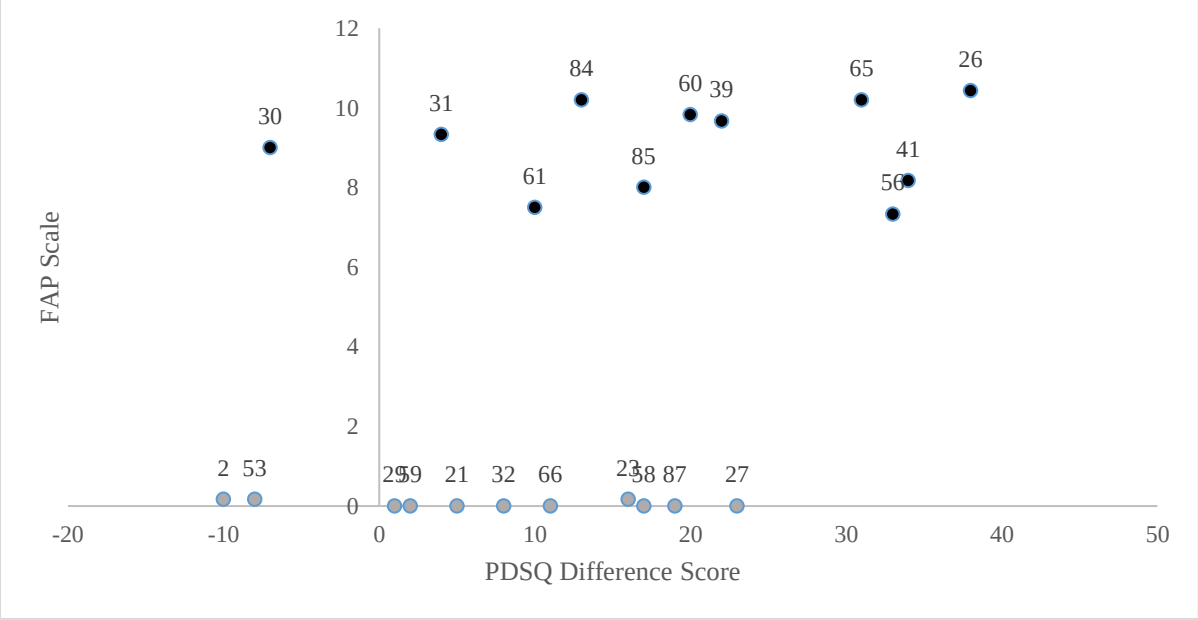

Figure 7. FAP Scale Average and PDSQ Difference Score

Baron and Kenny (1986) described the logic and established the basic procedure for models of mediation, which is comprised of 4 steps: 1) The IV (FAP vs. WW) correlates with the DV (i.e., PDSQ change), 2) the IV correlates with the proposed mediator(s) (i.e., FAP scale, FAP instance,

[2] For Figures 7, 8, & 9 data labels represent the participant number

- FAP

- SL

IV turns count), 3) the mediator correlates with the DV, and 4) after controlling for the mediator, the relationship between IV and the DV is now significantly reduced. Having established the relationship between treatment condition (FAP vs WW) and PDSQ outcome and the relationship between treatment condition and FAP scale, FAP instances, and IV turns, but having failed to find any of the FAP adherence measures as statistical mediators scatterplots were generated to examine individual relationships between the proposed mediators and the DV (PDSQ change).

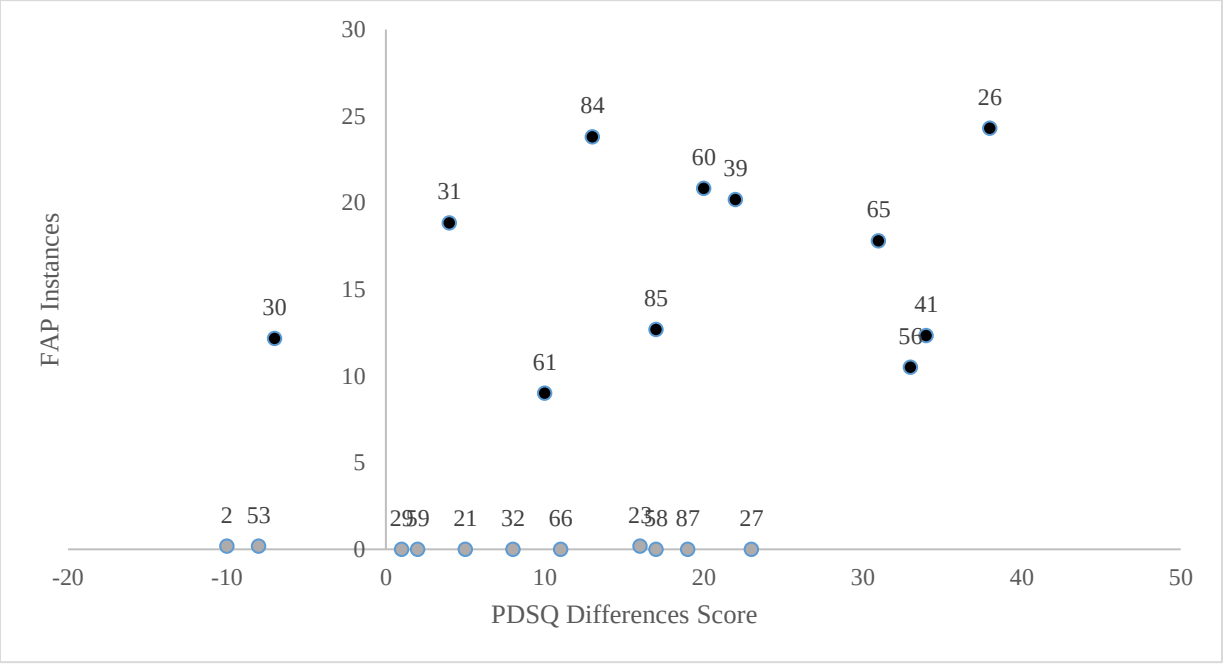

Figure 8. Instances Average and PDSQ Difference Score

Figure 7 represents the relationship between the FAP scale and the pre-post PDSQ difference scores for each participant. Consistent with previously reported analyses (Maitland et al, 2016), there is evidence that the FAP condition performed better than the SL condition as the filled circles are distributed to the right on the X axis, indicating greater PDSQ change.

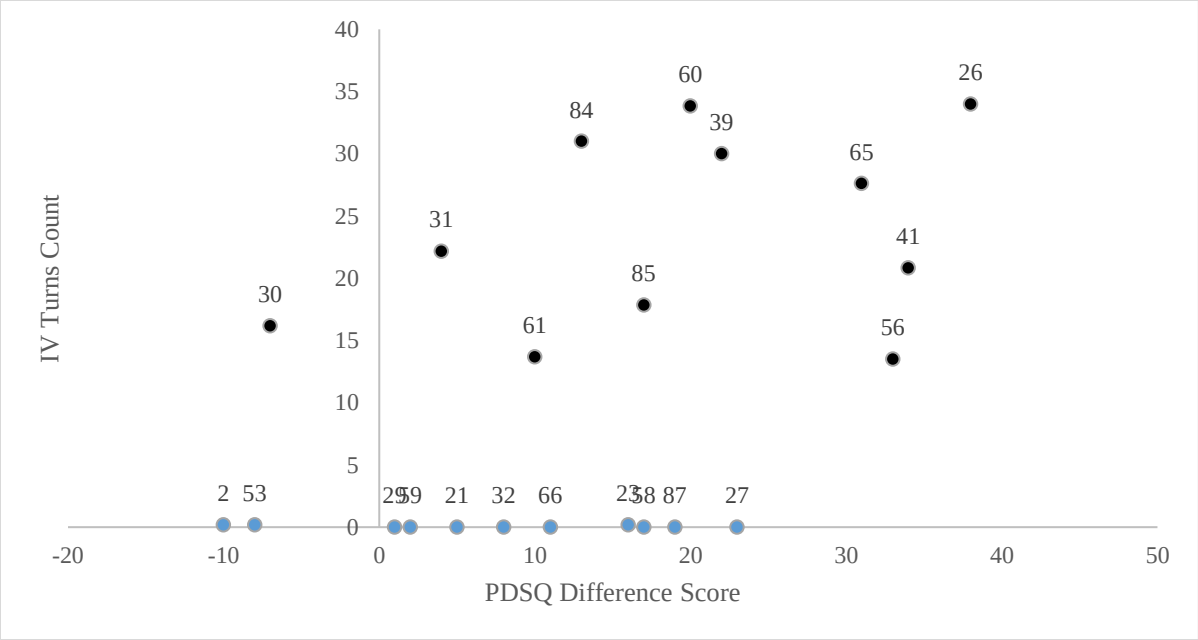

Figure 9. IV Turns Count Average and PDSQ Difference Score

Participants' FAP scale scores were measured along the Y axis. Those in the FAP condition clustered tightly between a vertical range of 7 and 11, whereas those in the SL condition clustered tightly around 0; therefore, simple visual inspection further demonstrates the distinction between the two conditions and FAP as an independent variable. However, little difference exists for scores on the FAP scale between those who improved and those who made little to no improvement on the PDSQ. To be consistent with the logic of mediation, the expected relationship would be for those who received less FAP to also demonstrate less symptom improvement and those who receive more FAP to demonstrate more symptom improvement (Baron & Kenny, 1986). Examination of several specific cases is illustrative. Participant 26 received the most FAP of any participant according to the FAP scale (FAP scale average score = 10.43) and saw the most change on the PDSQ (PDSQ change score = 38). The data from participant 26 fits well with the mediational model as evidenced by a high dose of FAP and strong symptom improvement. However, consider participants 41 and 56 who reported the next

largest changes on the PDSQ (PDSQ change scores = 34 and 33, respectively). These two participants received some of the lowest quantities of FAP according to the FAP scale (8.17 and 7.33, respectively). Therefore, participants 41 and 56 data are inconsistent with the dose-response relationship expected in statistical mediator models. Similarly, consider participants 30 and 31 whose scores on the FAP scale were average for the group; however, their change on the PDSQ was minimal (31) or in the direction of worsening (30), despite receiving an average amount of FAP.

Figure 8 represents the relationship between the FAP instances score and the PDSQ difference scores. Participants' FAP instances scores are measured along the Y axis. Those in the FAP condition cluster in ranges between 9 to 24.9 and the SL condition ranges approximately around zero. Similar results were found as in Figure 7. Participant 26 was rated with the highest average FAP instances and greatest PDSQ change score, while participants 41 and 56 had the second and third highest PDSQ change scores, respectively, but with some of the lowest FAP instances scores. Participants 31, and 30 demonstrated the least PDSQ change; however, participant 31 (the second lowest PDSQ score) rated just above the median in FAP instances, and participant 30, while having one of the lowest 3 FAP instances scores demonstrated a notable amount of FAP instances at 12.17 for having a clinical depreciation of -10.

Figure 9 represents the relationship between the IV turns count and the difference in PDSQ scores. Participants IV turns count scores are measured along the Y axis. Similar to both Figures 7 and 8 those in the FAP condition fall between the 13 and 34 range, whereas the SL condition clusters around zero. Participant 26 maintains the highest number of IV turns count with highest PDSQ change score. Participants 41 and 56, with the second and third highest PDSQ changes, while having among the lowest IV turn counts. Finally, participants 30 and 31

again appeared to be exposed to a significant quantity of IV work but failed to demonstrate robust PDSQ change.

Secondary Analysis

Additional analyses were run in an attempt to further explore the possibility of a mediational relationship between the fidelity metrics and outcome measures. The proposed mechanism of action in FAP is therapist provided consequences (reinforcement) contingent on CRB. While all five items are relevant to implementation of FAP's mechanism, item 8 ("Did the therapist share his/her reaction to the client's in session behavior?") specifically attempts to capture the provision of interpersonal consequences as conceptualized in FAP. " As such, it was hypothesized that item 8 (as measured by both the FAP Scale and the FAP Instances scores) might best capture the mechanism of action in FAP and thereby mediate outcomes on the PDSQ and FIS.

Four additional mediator analyses were run to evaluate this hypothesis, specifically, whether Item 8 on the FAP Scale and FAP Instances were significant statistical mediators of PDSQ and FIS change. A point estimate, based on 10,000 bootstrapped samples, with a 95% bias corrected confidence interval that did not include zero, was used to indicate a significant indirect effect ($p < .05$) suggesting mediation (Hayes, 2013; Preacher & Hayes, 2008). Item 8 FAP Scale average was calculated by obtaining the average scaled rating for item 8 for each participant observed by coder 1. Item 8 FAP Instances average was calculated by obtaining the average number of overall observations of Item 8 for each participant observed by coder 1. Initial mediator analyses were run examining the residualized FIS change scores. The indirect effect of condition on residualized FIS change through Item 8 FAP scale average was not statistically significant (point estimate = -.04, $p > .05$; 95% confidence interval [CI] [-1.91, 4.16]). Similarly,

the indirect effect of condition on residualized FIS change through Item 8 FAP instances average was not statistically significant (point estimate = -0.28, $p > .05$; 95% confidence interval [CI] [-1.67, 1.42]). Next, mediator analyses were run examining the residualized PDSQ change scores. The indirect effect of condition on residualized PDSQ change through Item 8 FAP scale average was not statistically significant (point estimate = -0.31, $p > .05$; 95% confidence interval [CI] [-2.67, 3.08]). Furthermore, the indirect effect of condition on residualized PDSQ change through Item 8 FAP instances average was not statistically significant (point estimate = -.15, $p > .05$; 95% confidence interval [CI] [-2.52, 1.58]). These data do not support the hypothesis of our secondary analyses that Item 8 as the most direct measure of FAP's mechanism of action would mediate change.

DISCUSSION

This study evaluated a proposed measurement of FAP treatment fidelity, the FAP-AF. The FAP-AF was examined for its inter-rater reliability; correlations with other established fidelity metrics (both within and between coders) and its ability to serve as a potential mediator of change.

For the overall measure, inter-rater reliability was found to be excellent between 3 coders from both conditions. Further, this finding was sustained when 2 of those 3 coders extended from the 24% of sessions to evaluating 52% of sessions at random. When isolated to the FAP condition alone, and regardless of the number of coders, intra-class correlations were within the excellent range, suggesting excellent reliability. Therefore, these data suggest the FAP-AF successfully distinguishes FAP from SL. As such, this measure has excellent reliability and may be useful in a replication study for FAP versus SL or other treatment designs with similar components.

Regarding inter-method reliability, examining the relationship between the FAP scale, the FAP instances score, and the IV turns counts, both between coders and within coders strong, positive correlations were obtained. When isolated to the FAP condition, these strong, positive relationships generally sustained and only dipped to moderate positive relationships for the between coder relationships. These data suggest that our method may be as good as other pre-established, previously examined (Kanter et al., 2005), and more tedious methods of fidelity. Furthermore, these data demonstrate that multiple coders at various levels of FAP expertise are able to reliably distinguish FAP from SL using the FAP-AF.

Based upon the promising preliminary results found in Maitland et al., (2016)'s RCT with the FAP scale of the FAP-AF acting as a mediator for one of the main outcomes (i.e., PDSQ), this study explored the whether this finding held when the FAP Scale was coded for the entire dataset. However, the data demonstrated no significant indirect effect for either FIS or PDSQ residualized change scores for the FAP scale of the FAP-AF. These data cannot therefore suggest our measure to be a proxy for FAP's mechanism of action.

Mediator analyses were run on the other fidelity metrics and similar non-significant findings were sustained with these data. However, the indirect effect of condition on residualized PDSQ change through the IV turns count was trending toward significance, as seen by a point estimate of -1.06, $p > .05$; 95% and confidence interval [CI] [-2.80, .23]. Despite considering all sessions for these mediator relationships, it is important to consider the small sample size of the original RCT ($n = 22$). In addition to reduced power, a small sample size has a limited data pool and therefore a limited breadth of observations of the population it intends to represent. Perhaps a larger sample size would allow for more variability and enhance our findings.

Scatterplots were run to examine the relationships of the PDSQ difference scores with the FAP scale, FAP instances, and IV turns count. Through visual inspection several notable findings elucidated the lack of findings in the mediational models while offering visual support for aspects of adherence. The clustering of FAP condition and WW condition around separate portions of the vertical axis further demonstrates the distinction between the two conditions and FAP as an independent variable. Applying the logic of mediation, those who received less FAP should also demonstrate less symptom improvement while those who receive more FAP should demonstrate more symptom improvement. The scatterplots offer visual insight into this lack of relationship. Several participants, observed in all three adherence metrics, had some of the highest PDSQ difference scores (indicative of clinical improvement) but also some of the lowest FAP means. These participants illustrate potential model failures. These findings may suggest particular implications regarding the mechanism of FAP and the frequency of FAP-consistent interactions. For instance, an important empirical inquiry is whether the amount of FAP-consistent statements (proxy for in-session shaping and contingent responding) impacts outcomes. Researchers have yet to explicate the lower and upper limits for the number of "FAP-consistent statements" and/or IV statements necessary to see treatment effects. As in the current study, even those who received the "least amount of FAP" had varying outcomes from fair to excellent in terms of treatment response. Several participants received some of the highest doses of FAP and varied along symptom outcome in a similar fashion. This provokes the question of whether dosing of FAP matters and to what degree is necessary to see treatment effect. In a small n RCT these questions may be difficult to examine, and larger samples with more available observations might provide more trends from which to draw inferences. While it is expected to have variability in outcomes across participants due to variety of circumstances (e.g., different

life habits, presenting concerns, daily life stressors, measurement error, etc.), larger sample size and subsequent increase in power may mitigate some of the problems in variability and sampling error. Future studies should examine these questions regarding dosing in order to further elucidate the breadth and depth of FAP's proposed mechanisms.

The measure is sensitive enough to capture the provision of FAP compared to an alternative treatment. That is to establish FAP as an IV for efficacy research. The measure did not function as a statistical mediator of outcome. One reason may be that the scoring of items is based on quantity of occurrences, which treats each occurrence as equally important, introducing some potential insensitivity. For instances, mediational analyses expect a dose-response relationship between treatment, mediator, and outcome. That is, a client who perceived 100 reinforcers should have better outcome than one receiving 10. However, this assumes, as our measure does, that each reinforcer is of equal magnitude. If may well be that a client who received fewer total reinforcers that were of higher magnitude (e.g., more interpersonally meaningful) would have a better outcome than on who receive more total reinforcers each of which was a smaller magnitude. Quantifying the potency of therapist contingent response is challenging. The FAP-AF emphasizes frequency, but future work could seek to address this limitation.

A second potential reason for failure of the FAP-AF to serve as a mediator of outcome is that the FAP treatment provided had its effects for reasons other than those captured by the FAP-AF. That is, the measure captured what it was designed to capture, but what it captured was not significantly linked to outcome. It might be that FAP outperformed WW for reasons other than those conceptually underpinning FAP.

Future Directions

While this study found relatively little difference between the different fidelity techniques in terms of reliability, the FAP-AF may have some logical utility and convenience that the other metrics (i.e., the FAP instances and the IV turns counts) do not have. For instance, the FAP-AF has the convenience of Likert scaling and has been used in time-limited samplings (15-minute samples of sessions) and still been effective in distinguishing FAP (Maitland et al., 2016; Knott et al., 2016 *conference presentation*) as an independent variable. There is no data to show what the limits are for distinguishing FAP with IV turns or FAP instances, other than whole 60-minute sessions demonstrated in this study. Future studies could examine the strengths and limits of these other metrics. Future research should focus on examining the fit of the Likert scaling. The decision to scale the items from 0 – 3 was done based on intuitions about reasonable range. It may be possible that the current item scaling resulted in ceiling effects for some of the items; leading to a failure to capture variation necessary for examining indirect (mediator) effects. The data from this study demonstrate that the 4-point Likert shows sound psychometrics thus far, but these concerns merit further investigation. Furthermore, the FAP-AF is designed after the FAP 5-rules. These items may potentially be useful for examining further mechanism or predictive relationships regarding therapist in-session use of the FAP 5-rules. The IV turns count represents here-and-now statements regarding the relating between therapist and client, which may not necessarily allow for examination into each of the 5-rules with such precision. Future research might delve further into the empirical question of how different these two metrics are and the breadth of the utility of either of them.

Regarding the lack of findings in mediator relationship between the treatment fidelity metrics and outcomes, several implications may be indicated. First, it is possible the treatment

fidelity techniques, in particular the FAP scale, do no serve as a proxy for FAP's mechanism of action, therapist contingent responding. It may also be surmised that the mechanism of our outcomes is different than what FAP proposes, therefore our metric is not designed to capture the mechanism. Further, perhaps limitations related to measurement and measurement error contributed to that failure to identify that FAP scale as a mediator. For instance, the dependent variables, the FIS and PDSQ, were global compared to in-session, individual FAP targets. It is also surely the case that our dependent variables were influenced not only by events occurring in therapy but also daily life events experienced by our participants. To the extent these daily life events were distributed across the sample they would add sample variation unrelated to treatment making it more difficult to find mediator relationship, with such a relatively small sample.

Another consideration relates to quantity versus quality (potency/magnitude) of consequences. Consider two common FAP session experiences: 1) the client requires several interventions of blocking CRB1 and iterations of the logical FAP interaction in which the therapist moves through the 5-rules within session perhaps four to five times across one or more response classes, or 2) the therapist moves through a session with a client with only one to two iterations of the logical FAP interaction, but with a subjective level of potency or meaningfulness that is yet to be quantified by a fidelity measure. Potency may well be the most important variable and while it may relate to frequency the two are conceptually distinguishable. Maitland et al., (2016) did well to first establish FAP as a promising treatment for interpersonal relating and suggested that the FAP Scale may serve as a proxy for the mechanism of action; however, in our study, which differed from Maitland's (2016) in the number of sessions observed for adherence, perhaps the mechanism was difficult to capture related to varying degrees of both dose and meaningfulness. As such, we suggest research explore some of these empirical

concerns through the use of large sample RCTs in order to explore both the question of dose (frequency) and meaningfulness. The design of a study may be structured to address these concerns, particularly for the question of FAP dose. A large sample, 3-arm RCT, where one arm consists of 6 sessions of FAP (representing "high FAP"), the second of 6 sessions of SL (representing "low FAP"), and the third of 6 sessions of 3 each FAP or SL in random order (representing "medium FAP"), might provide information regarding the impact of dose on treatment response. Furthermore, a structured study design with adjustment to the implementation of FAP may address the concern of meaningfulness of interactions. Haworth and colleagues utilized an analogue design to address reinforcement and Rule 3 (reinforce CRB2), a hypothetically important component in "meaningfulness". These researchers asked a series of closeness generating questions to participants and demonstrated the increase in participant self-disclosure (target "CRB") being directly related to whether or not participants received meaningful responses from researchers ("rule 3"; Haworth et al., 2015). With this same idea of meaningfulness in mind, another 3-arm RCT is suggested examining: 1) comprehensive FAP ("meaningful group"); 2) the use of the FAP 5-rules where the therapists are not trained in authentic rule 3 meaningful responses as they were in Haworth et al., (2015) or forgo meaningful responses ("superficial FAP"), and 3) supportive listening but no FAP.

Limitations

A prominent limitation of this study is in the actual empirical question underlying the purpose of the study. How to quantify "FAP interactions" remains to be an important question and may need to be clarified in order to elucidate the mechanism of action. For instance, when we consider the amount of FAP consistent interactions, is it simply a matter of doing more of the FAP 5-rules or having more "meaningful" moments of engaging the rules. If it is the latter, the

question of quantifying "meaningfulness" becomes imperative. Within this study, the coders were instructed to takes notes, tally FAP item instances beyond the required scale, and count IV turns. As such, deconstructing these additional notes and data may lend itself valuable to answering this question in secondary analyses or future research.

Further, several methodological limitations exist within the current study, some as a result of the study of origin (i.e, Maitland et al., 2016). One such limitation is the FAP sessions were provided in the RCT by a single therapist, which may limit the generalizability of our reliability data. Likewise, the RCT data was smaller sample size and therefore may not provide the power necessary to identify mediator relationships or for more general inferences to be drawn. Further, several session tapes were lost to technological error (e.g., camera recording error, computer transfer error, etc), which prevents us from analyzing complete data. While having expertise and training in FAP, the three coders were all graduate students, which may preclude generalization to the larger population of potential FAP coders. The author of the study was also the main coder, as well as the main coding trainer for the two other coders of the study.

Conclusions

The present study examined the utility of a fidelity measure for distinguishing Functional Analytic Psychotherapy as an independent variable for treatment outcome research. Other metrics were compared in this study (i.e., FAP instances, IV turns count) and all were shown to be reliable and predictable methods of demonstrating fidelity. These data suggest that the FAP scale of the FAP-AF, a simple 5-item measure, can reliably substantiate the implementation of FAP making it potentially useful tool for future efficacy research as it is more time efficient than other more tedious and time-consuming fidelity techniques (i.e., Kanter et al., 2005). The FAP scale of the FAP-AF did at least as good as IV turns count, a metric previously utilized in the

literature; however, due to the design of the measure, the FAP-AF may provide more detailed information to researchers who intend to also explore the 5-rules of FAP. Furthermore, as the FAP-AF was found to be capable of distinguishing FAP from a supportive listening condition even for the first 15 minutes of sessions (Maitland et al., 2016), it is efficient at capturing FAP as an independent variable and distinguishing it from other treatments. None of the fidelity measures served as statistical mediators of outcome. The extent to which the FAP-AF is useful as a process measure remains to be determined in future research. Specifically, despite this study's lack of support for the FAP scale (and Item 8 of the FAP scale) to serve as a proxy for FAP's mechanism of action, it is possible with an increased sample size, or enhancements to capture the meaningfulness of engaging the FAP 5-rules, that future studies may be able to shed light on whether the FAP-AF functions as a statistical mediator of outcome, consistent with predictions from FAP.

Appendix A

FAP Adherence Form (FAP-AF)

FAP Adherence Form

Participant # _____ **Session # _____**

Instructions: Read over every question before watching the session video, making notes as needed, rewinding as needed, and rate and score the adherence form at the end.

1. To what extent was the therapist's behavior mainly directed toward attempts to understand the daily life social relationships from the client's vantage point

0	1	2	3
never	once	twice	3+ times

2. Did the therapist engage in reflective and empathic listening in reaction to the client?

0	1	2	3
never	once	twice	3+ times

3. Did the therapist prompt/encourage the client to discuss daily life social relations?

0	1	2	3
never	once	twice	3+ times

4. Did the therapist turn the focus of the session on the client's feelings/emotional reactions to events in his/her daily life social relations?

0	1	2	3
never	once	twice	3+ times

5. Did the therapist turn the focus of the session on the clients in-session behavior?

0	1	2	3
never	once	twice	3+ times

6. Did the therapist compare in-session events to the participant's daily life?

0	1	2	3
never	once	twice	3+ times

7. Did the therapist prompt/encourage the client to engage in particular responses in the session?

0	1	2	3
never	once	twice	3+ times

8. Did the therapist share his/her reaction to the client's in session behavior?

0	1	2	3
never	once	twice	3+ times

9. Did the therapist check with the participant to see his/her response to the therapist sharing his/her reaction?

0	1	2	3
never	once	twice	3+ times

10. Did the therapist assign the client for homework to engage in specific out of session behaviors that followed from in-session interactions?

Yes Partially No

Appendix B

HSIRB Approval Letter

WESTERN MICHIGAN UNIVERSITY
Human Subjects Institutional Review Board

Date: March 25, 2016

To: Scott Gaynor, Principal Investigator
Lindsey Knott, Student Investigator
Rachel Petts, Student Investigator
Rebecca Rausch, Student Investigator

From: Amy Naugle, Ph.D., Chair

Re: HSIRB Project Number 16-03-33

This letter will serve as confirmation that your research project titled "The Relationship between In-Vivo Focus and Client Outcomes in Functional Analytic Psychotherapy" has been **approved** under the **exempt** category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note: This research may **only** be conducted exactly in the form it was approved. You must seek specific board approval for any changes in this project (e.g., ***you must request a post approval change to enroll subjects beyond the number stated in your application under "Number of subjects you want to complete the study***)." Failure to obtain approval for changes will result in a protocol deviation. In addition, if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

Reapproval of the project is required if it extends beyond the termination date stated below.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: **March 24, 2017**

1903 W. Michigan Ave., Kalamazoo, MI 49008-5456
PHONE: (269) 387-8293 FAX: (269) 387-8276
CAMPUS SITE: 251 W. Walwood Hall

REFERENCES

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.

Callaghan, G. M., & Follette, W. C. (2007). Manual for the functional analytic psychotherapy rating scale. *The Behavior Analyst Today, 9*(1), 57-97.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*(1), 7-18.

Darrow, S. B., & Follete, W., (2014). Where is the beef?: Reply to Kanter, Holman, and Wilson. *Journal of Contextual Behavioral Science*. *3*(2014). 69-73.

Descutner, C. J., & Thelen, M. H. (1991). Development and validation of a Fear-of-Intimacy Scale. *Psychological Assessment: A Journal of Consulting and Clinical Psychology.* 3(2). 218-225

Donohoe, J.W., & Palmer, D.C. (2004). *Learning and Complex Behavior.* Richmond, MA: Ledgetop Publishing.

Follette, W. C., & Bonow, J. T. (2009). The challenge of understanding process in clinical behavior analysis: The case of functional analytic psychotherapy. *The Behavior Analyst, 32*(1), 135-148.

Gaynor, S. T., & Lawrence, P. S. (2002). Complementing CBT for depressed adolescents with Learning through In Vivo Experience (LIVE): Conceptual analysis, treatment description, and feasibility study. *Behavioural and Cognitive Psychotherapy,* 30(1),

79–101.

Gifford, E. V., Kohlenberg, B. S., Hayes, S. C., Pierson, H. M., Piasecki, M. P., Antonuccio, D. O., & Palm, K. M. (2011). Does acceptance and relationship focused behavior therapy contribute to bupropion outcomes? A randomized controlled trial of Functional Analytic Psychotherapy and Acceptance and Commitment Therapy for smoking cessation. *Behavior Therapy*, 42(4), 700–715.

Hayes, S. C., Masuda, A., Bissett, R., Luoma, J., & Guerrero, L. F. (2004). DBT, FAP and ACT: How empirically oriented are the new behavior therapy technologies? *Behavior Therapy*, 35(1), 35–54.

Holman, G., Kohlenberg, R. J., Tsai, M., Haworth, K., Jacobson, E., & Liu, S. (2012). Functional Analytic Psychotherapy is a framework for implementing evidence-based practices: The example of integrated smoking cessation and depression treatment. *International Journal of Behavioral Consultation and Therapy*, 7(2–3), 58–62.

Kanter, J. W., Holman, G. & Wilson, K., (2014). Where is the love? Contextual Behavioral Science and Behavior Analysis. *Journal of Contextual Behavioral Science*. *3*(2014). 69-73.

Kanter, J. W., Manbeck, K. E., Kuczynski, A. M., Maitland, D. W. M., Villas-Boas, A., & Ortega, M. A. R. (2017). A comprehensive review of research on Functional Analytic Psychotherapy. *Clinical Psychology Review*. 58(2017). 141-156.

Kanter, J.W., Schildcrout, J. S., & Kohlenberg, R. J., (2005). The in vivo process in cognitive therapy for depression: Frequency and benefits. *Psychotherapy Research*. *15*(4). 366-373.

Kohlenberg, R.J., Kanter, J.W., Bolling, M. Y., Parker, C., & Tsai, M. (2002). Enhancing cognitive therapy for depression with functional analytic psychotherapy: Treatment guidelines and empirical findings. *Cognitive and Behavioral Practice, 9*(3), 213-229.

Kohlenberg, R. J., & Tsai, M. (1991). *Functional Analytic Psychotherapy: A guide for creating intense and curative therapeutic relationships*. New York, NY: Plenum.

Landes, S. J., Kanter, J. W., Weeks, C. E., & Busch, A. M. (2013). The impact of the active components of Functional Analytic Psychotherapy on idiographic target behaviors. *Journal of Contextual Behavioral Science*, 2(1), 49–57.

Maitland, D. W. M., & Gaynor, S. T. (2012). Promoting efficacy research on functional analytic psychotherapy. *International Journal of Behavioral Consultation and Therapy. 7*(2), 63-71.

Maitland, D. W. M., & Gaynor, S. T. (2016). Functional analytic psychotherapy compared with supportive listening: An alternating treatments design examining distinctiveness, session evaluations, and interpersonal functioning. *Behavior Analysis: Research and Practice. 16*(2). 52-64.

Maitland, D. W. M., Petts, R. A., Knott, L. E., Briggs, C. A., Moore, J., & Gaynor, S. T. (2016). A randomized controlled trial of Functional Analytic Psychotherapy versus watchful waiting: enhancing social connectedness and reducing anxiety and avoidance. *Behavior Analysis: Research and Practice. 16*(3). 103-122

Mangabeira, V., Kanter, J., & Del Prette, G. (2012). Functional analytic psychotherapy (FAP): A review of publications from 1990 to 2010. *International Journal of Behavioral Consultation and Therapy, 7*(2-3), 78-89.

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology*. 75(6), 829-841.

Rounsaville, B. J., Carroll, K. M., and Onken, L. S. (2001). A stage model of behavioral therapies research: getting started and moving on from stage I. *Clinical Psychology Science Practice.* 8:133-142.

Singh, R. J. & O'Brien, W. H. (2018). A quantitative synthesis of functional analytic psychotherapy single-subject research. *Journal of Contextual Behavior Science*. 7(2018). 35-46.

Tsai, M., Kohlenberg, R. J., Kanter, J. W., Kohlenberg, B., Follette, W. C., & Callaghan, G. M. (Eds.). (2009). *A guide to functional analytic psychotherapy: Awareness, courage, love, and behaviorism. New York, NY: Springer Science + Business Media.*

Tolin, D. F., McKay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically supported treatment: recommendations for a new model. *Clinical Psychology: Science and Practice.* Doi: 10.111/cpsp.12122

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology.* 61(4), 620-630.

Weeks, C. E., Kanter, J. W., Bonow, J. T., Landes, S. J., & Busch, A. (2012). Translating the theoretical into practical: A logical framework of functional analytic psychotherapy interactions for research, training, and clinical purposes. *Behavior Modification, 36*(1), 87-119.

Zimmerman, M., & Mattia, J. L. (2001). A self-report scale to help make psychiatric diagnoses: The Psychiatric Diagnostic Screening Questionnaire. *Archives of General Psychiatry*. 58(8), 787-794.