Published in final edited form as:

Adm Policy Ment Health. 2015 March ; 42(2): 229–243. doi:10.1007/s10488-014-0548-2.

Validity of Therapist Self-Report Ratings of Fidelity to Evidence-Based Practices for Adolescent Behavior Problems: Correspondence between Therapists and Observers

Aaron Hogue,

Treatment Research Division, The National Center on Addiction and Substance Abuse at Columbia University, 633 Third Avenue, 19th floor, New York, NY 10017, USA

Sarah Dauber,

Treatment Research Division, The National Center on Addiction and Substance Abuse at Columbia University, 633 Third Avenue, 19th floor, New York, NY 10017, USA

Emily Lichvar,

Treatment Research Division, The National Center on Addiction and Substance Abuse at Columbia University, 633 Third Avenue, 19th floor, New York, NY 10017, USA

Molly Bobek, and

Treatment Research Division, The National Center on Addiction and Substance Abuse at Columbia University, 633 Third Avenue, 19th floor, New York, NY 10017, USA

Craig E. Henderson

Department of Psychology, Sam Houston State University, Huntsville, TX 77340, USA

Abstract

Developing therapist-report fidelity tools to support quality delivery of evidence-based practices in usual care is a top priority for implementation science. This study tested the reliability and accuracy of two groups of community therapists who reported on their use of family therapy (FT) and motivational interviewing/cognitive-behavioral therapy (MI/CBT) interventions during routine treatment of inner-city adolescents with conduct and substance use problems. Study cases ($n = 45$) were randomized into two conditions: (a) Routine Family Therapy (RFT), consisting of a single site that featured family therapy as its standard of care for behavioral treatment; or (b) Treatment As Usual (TAU), consisting of five sites that featured non-family approaches. Therapists and trained observational raters provided FT and MI/CBT adherence ratings on 157 sessions (104 RFT, 53 TAU). Overall therapist reliability was adequate for averaged FT ratings ($ICC = .66$) but almost non-existent for MI/CBT ($ICC = .06$); moreover, both RFT and TAU therapists were more reliable in reporting on FT than on MI/CBT. Both groups of therapists overestimated the extent to which they implemented FT and MI/CBT interventions. Results offer support for the feasibility of using existing therapist-report methods to anchor quality assurance procedures for FT interventions in real-world settings, though not for MI/CBT.

Keywords

Therapist-report fidelity; Observational fidelity ratings; Adolescent behavior problems; Usual care; Family therapy; Cognitive-behavioral therapy; Motivational interviewing

The emerging discipline of implementation science is focused on elucidating the conditions under which evidence-based interventions (EBIs) can be delivered with fidelity by front-line therapists and sustained over time in community settings (McHugh and Barlow 2010). EBIs exist along a continuum of standardization defined by three intersecting dimensions: specification of prescribed intervention techniques; elaboration of rule-driven procedures for when and how interventions should be delivered; and customization for given clinical populations. EBIs include both empirically supported treatments—manualized models and brand-name programs—and evidence-based practices (EBPs)—generic, modular, or kernel/core interventions with strong empirical support (Chorpita et al. 2005; Embry and Biglan 2008; Garland et al. 2010). Empirically supported treatments contain a high degree of technique specification, formulaic implementation guidelines, and precisely defined target groups; EBPs vary more widely in standardization along each dimension (Westen et al. 2005).

Need for Efficient Fidelity Measures to Support EBI Quality Assurance (QA) in Usual Care

There is strong consensus that successful implementation of EBIs in mainstream care for behavioral health disorders requires procedures designed to ensure that interventions are delivered according to given standards, that is, to the intended population, by appropriately trained and supervised providers, and in accordance with specified principles, procedures, and techniques (Aarons et al. 2011). In laboratory settings such methods are called integrity or fidelity procedures (Perepletchikova 2011), owing to their tight focus on the delivery of the intervention package itself in the context of controlled research in which therapist, client, and organizational variables are favorable by design. In field settings such methods are usually called QA procedures (Schoenwald 2011).

Well-specified QA procedures currently exist for a wide variety of EBIs and usually contain several components (Fixsen et al. 2005): guidelines for selecting adoption-ready sites and qualified trainees; standardized provider training toolkits; fidelity checklists and related QA measurement tools designed to yield immediate performance evaluation data; procedures for ongoing consultation from model experts; continuous quality improvement procedures to evaluate implementation data collected on site, feed selected data back to therapists, and buttress agency support; and certifications granted to providers who complete training and maintain quality standards. Continuous quality improvement procedures are a distinctive component of QA systems intended to maximize the utility of implementation evaluation in routine practice via site-specific performance feedback (O'Neill et al. 2011).

A few treatment developers have tested the impact of QA systems on provider performance and client outcomes (e.g., Schoenwald et al. 2009), with promising results to date. However, despite their many strengths, validated QA systems for manualized treatments cannot fully

meet the fidelity assessment needs of the provider workforce. Model-specific QA procedures are fundamentally nontransferrable: They support only the specific manualized treatment to which they are attached. This presents two limitations to widespread applicability. First, they have limited utility for guiding “off-map” interventions such local adaptation of the given treatment for clients with uncommon or boundary-condition diagnoses, or flexible combinations of treatment components with non-manualized interventions for clients with co-occurring disorders (Gallo and Barlow 2012). Second, they may not be suitable for evaluating the broader spectrum of generic interventions that constitute EBPs commonly implemented in usual care. At minimum, QA tools developed to assess EBPs in front-line clinics should be (1) designed to capture a wide variety of practices rather than a narrow band of prespecified techniques and (2) applicable to heterogeneous client pools with diverse profiles of comorbid problems (Bearsley-Smith et al. 2008; Schoenwald 2011). To be clear, such tools would not be used to supplement or replace existing QA procedures for manualized treatments, but instead, to promote high-fidelity use of EBPs among providers who do not opt for wholesale adoption of manuals and their companion QA systems (Garland et al. 2010a).

Meeting the Need: Therapist-Report Measures of Fidelity to Core Elements of EBIs

To meet the need for efficient and flexible assessment of EBI implementation in usual care, it is essential to develop fidelity evaluation tools that share two features: (1) therapist-report format and (2) focus on core treatment techniques that are common across multiple empirically supported treatment manuals for similar populations (Garland et al. 2010a; McLeod et al. 2013). Therapist-report measures of EBI fidelity offer several methodological strengths that may afford the desired balance between rigor and relevance in QA procedures: they are quick, inexpensive, and non-intrusive; they capture the unique viewpoint of the provider delivering the interventions; and they can be completed throughout treatment, which facilitates measurement of infrequent but clinically meaningful interventions (Carroll et al. 1998; Weersing et al. 2002). Therapist-report measures can also inform quality improvement procedures via feedback loops of several kinds: as a self-check by therapists to mark their own progress in treating cases; as a supervision aid for trainers to monitor fidelity; and as administrative data for stake-holders and reviewers to evaluate therapist- and agency-level performance (Bearsley-Smith et al. 2008; Schoenwald et al. 2005).

EBI fidelity tools will also benefit from focusing on discrete intervention techniques shared across multiple treatment manuals for a given disorder. These *core element* interventions¹ (Chorpita et al. 2007) are approach-specific (i.e., identified with a particular treatment orientation and/or modality) but model-free (i.e., not inextricably bound to a single manual or intervention sequence). Within the youth treatment field, prime examples of core element EBIs include cognitive coping interventions for anxiety (Chorpita 2007), behavioral parenting training and child-only interventions for childhood conduct problems (Garland et al. 2008), and family therapy techniques for adolescent behavior problems (Baldwin et al. 2012; Hogue and Liddle 2009). A primary virtue of adopting core EBIs is that they equip

¹Core element interventions, which refer to specific treatment techniques, should not be confused with common elements/factors in therapy, which refer to non-specific relational processes germane to virtually every psychotherapy approach and thought to promote client gains (e.g., therapeutic alliance, positive regard).

clinicians with fundamental techniques that can be judiciously applied to clients presenting with diverse, comorbid, and/or emerging clinical problems (Barth et al. 2014; Chorpita et al. 2005, 2007). Core elements also represent a comfortable middle ground between molar versus molecular specification of treatment processes that is well-suited for describing the eclectic clinical practices favored in usual care (Barth et al. 2014; Garland et al. 2010b).

Importantly, the field is beginning to investigate whether core EBIs can be clinically effective in usual care settings. For example, a recent randomized trial (Bearman et al. 2013; Weisz et al. 2012) tested a transdiagnostic treatment protocol consisting of modules for core treatment elements that systematic reviews had found to be most common in cognitive-behavioral and behavioral parent training manuals for childhood anxiety, depression, and disruptive disorders. Similar to empirically supported treatments, this modular protocol, called MATCH-ADC, was highly standardized with regard to technique specification, formulaic intervention selection, and client customization, including specific decision rules for when to implement which core interventions based on evolving case circumstances. Results showed that the standardized modular protocol outperformed both disorder-specific treatment manuals and TAU in promoting improvements in child outcomes across the board (Weisz et al. 2012) and maintained this advantage over TAU at two-year follow-up (Chorpita et al. 2013); moreover, the modular protocol was associated with greater perceived benefits and satisfaction by participating therapists (Korathu-Larson et al. 2011).

One Major Hurdle to Clear: Validity of Therapist Self-Reports of EBI Implementation

Currently there is only a handful of therapist-report fidelity tools designed to assess core EBIs commonly used in everyday practice, including measures for adult depression (Hepner et al. 2010), adult substance use (Gifford et al. 2012), childhood disruptive behavior (Hurlburt et al. 2010), and broad child psychotherapy approaches (Bearsley-Smith et al. 2008; Weersing et al. 2002). Although developing new therapist-report tools for other clinical populations could advance EBI dissemination in several ways, there remains a major hurdle to clear: Studies attempting to confirm therapist self-reports of EBI implementation via observational ratings have mostly produced disappointing results, casting doubt on the accuracy with which therapists can judge their own performances. Studies with adult samples have logged modest to weak correspondence between therapist and observer reports of fidelity to motivational interviewing (Martino et al. 2009; Miller et al. 2004) and cognitive-behavioral interventions (Brosan et al. 2008; Carroll et al. 1998). In the youth arena, Hurlburt et al. (2010) found that observers reported substantially less occurrence and lower intensity of EBIs for childhood conduct problems compared to therapists in front-line mental health care. In an exception to this trend, Chapman et al. (2013) reported that therapists, trained observational raters, and treatment experts were generally consistent in reporting on contingency management interventions for adolescent substance use, though therapists did report significantly greater use of these interventions than did observational raters. Overall, it appears that weak correlations between clinicians and observers are due in large part to therapist bias in overestimating the extent to which they deliver EBIs; this bias holds true for research-trained therapists delivering manualized treatments (Carroll et al. 1998; Martino et al. 2009) as well clinicians in routine practice (Brosan et al. 2008; Hurlburt et al. 2010). Simply stated, therapist-report fidelity measures cannot be (fully) trusted until

they establish reasonable correspondence with non-participant ratings of EBI implementation.

Development of a Therapist-Report Fidelity Tool for Adolescent Behavior Problems: ITT-ABP

In a previous study (Hogue et al. 2012) we introduced the Inventory of Therapy Techniques for Adolescent Behavior Problems (ITT-ABP), a post-session therapist-report measure of 25 techniques representing four theoretical approaches to treating ABPs: cognitive-behavioral therapy (CBT), family therapy (FT), motivational interviewing (MI), and drug counseling (DC). The derivation of ITTABP items is detailed in the “Measures” section; the items are listed in Table 2. These four approaches were included in the ITT-ABP because each has a substantial base of empirical support for addressing adolescent conduct, delinquency, and substance use problems (for reviews see Becker and Curry 2008; Chorpita et al. 2011; see also Winters et al. 2007) and are widely endorsed in front-line settings (Cook et al. 2010; Gifford et al. 2012). The CBT, FT, and MI approaches each boast several manualized versions that have proven efficacious for treating a range of ABPs, which makes them ideal candidates to supply core EBIs capable of addressing a variety of treatment needs. This is especially important in adolescent behavioral care, wherein comorbidity and multisystem involvement are more rule than exception (Hawkins 2009; Kazak et al. 2010). As described in Methods, ITT-ABP items derive from two observational fidelity scales that were validated during controlled trials of empirically supported treatments. There is considerable enthusiasm for using observational measures forged in controlled trials as the foundation for developing QA tools for usual care (Gifford et al. 2012; Schoenwald et al. 2011), on the premise that this will yield instruments with both strong psychometric properties and clinical validity in measuring EBPs.

Also as described in “Method” section, the ITT-ABP has demonstrated factor validity, discriminant validity, and convergent validity in a sample from the parent trial of the current study (Hogue et al. 2012). Factor analysis identified three clinically coherent EBI scales: FT, MI/CBT, and DC interventions. These factor scales correspond with the four observational fidelity scales used as source measures for the ITT-ABP—albeit with a pooled MI/CBT factor, a combination regularly found in clinical practice (e.g., Dennis et al. 2004).

Current Study Aims

The purpose of the current study was to compare therapist ratings on the ITT-ABP to ratings made by non-participant observers, thereby conducting a rigorous test of the criterion validity of the ITT-ABP: To what degree did scores on this therapist-report measure of fidelity to core EBIs for ABPs correspond with objective ratings made by highly trained coders? Study therapists and clients were participating in a randomized naturalistic trial testing the effectiveness of usual care treatment for adolescent disorders. Clients were randomized into two study conditions: (a) Routine Family Therapy (RFT), consisting of a single site that featured family therapy as its routine standard of care for behavioral treatment; or (b) Treatment As Usual (TAU), consisting of five sites in the same catchment area that did not espouse family therapy as a primary approach. All study therapists operated

under everyday practice conditions without benefit of external training, supervision, or clinical resources from the research team.

The study had two specific aims. The first aim was to test the reliability and accuracy of therapist-report ITTABP scores. We tested reliability by calculating intraclass correlation coefficients (ICCs) for therapist-observer correspondence on individual items and factor scales for 157 sessions (104 RFT, 53 TAU) from the parent trial. We tested accuracy by conducting mean comparisons between therapists and observers for the factor scales. Based on previous studies of clinician-observer correspondence on EBI fidelity ratings, we expected that therapist reliability would be moderate-to-poor, and that therapists would overestimate the extent to which they delivered EBIs in session. The second aim was to examine whether therapists demonstrated differential reliability and accuracy when rating approach-concordant versus approach-discordant EBIs. We predicted that family therapists would be most reliable and accurate in reporting on their fidelity to FT interventions; we had no predictions about whether TAU therapists would show differences in reliability and accuracy in reporting on FT versus MI/CBT fidelity. As an exploratory third aim, we examined whether therapist reliability differed according to years of professional experience. Note that the current study did not include analysis of the DC scale of the ITTABP for any aim, due to lack of variance in these scores stemming from failure to collect session recordings from the only TAU clinic that specialized in addiction treatment (as explained in the “Method” section).

Method

The parent study from which these data were collected was conducted between 2006 and 2012 under approval by the governing Institutional Review Board.

Study Clients

Demographic and psychiatric characteristics for $n = 45$ clients (28 RFT, 17 TAU) in the current study are listed in Table 1, for the entire sample and separately by study condition. The study sample contains all clients from the parent study who provided at least one recorded session, totaling 157 sessions (all recorded sessions had corresponding therapist-report data). Clients were adolescents [53 % male; mean age 15.2 years ($SD = 1.5$)] and their primary caregivers. Self-reported race/ethnicity was Hispanic (69 %), African American (11 %), multiracial (13 %), and other (7 %). Households were headed by single parents (65 %), two parents (23 %), grandparents (9 %), or other (2 %); 50 % earned less than \$15,000 per year and 13 % received public assistance. Adolescents were referred from schools (79 %), or other sources (22 %); 16 % were involved in the juvenile justice system at referral.

Rates of psychiatric diagnosis were assessed by research staff using the Mini International Neuropsychiatric Interview (Version 5.0; Sheehan et al. 1998), based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV; American Psychiatric Association 2000). Psychiatric diagnoses were given for meeting threshold based on either adolescent or caregiver report, with the following sample prevalence rates: oppositional defiant disorder (ODD) = 91 %, attention-deficit/hyperactivity disorder = 76 %, conduct disorder (CD) = 46 %, mood disorder or dysthymia = 41 %, substance use disorder (SUD) = 23 % (16 %

cannabis use, 14 % alcohol), generalized anxiety disorder = 16 %, posttraumatic stress disorder = 14 %. A total of 74 % of the sample was diagnosed with more than one disorder; of these comorbid cases, 6 % met criteria for ODD ? CD only, with the remaining 94 % showing some other comorbidity profile. Between-condition differences in demographic and diagnostic variables (see Table 1) were observed for one variable only: More caregivers in TAU had graduated high school compared to RFT ($\chi^2(1) = 5.7, p < .05$).

The parent randomized trial contained 205 participants, of which 157 (77 %) attended at least one treatment session at the assigned site, including 74 % of RFT cases and 79 % of TAU cases. Of these 157 treatment attenders, 45 cases (29 %) were included in the current study by virtue of having a recorded treatment session. Significant differences in the current study sample versus the originally randomized cases were observed for two demographic and diagnostic variables: Adolescents in the current sample were younger ($t(203) = 2.48, p < .05$) and less likely to have more than one diagnosis ($\chi^2(1) = 4.46, p < .05$). These same difference were observed when comparing the study sample to the subgroup of 157 cases that attended treatment: younger age ($t(155) = 2.46, p < .05$) and less likely to have multiple diagnoses ($\chi^2(1) = 3.70, p = .05$).

Client Recruitment, Randomization, and Enrollment in the Parent Trial

Clients were part of a parent study designed to identify adolescents with untreated behavioral health problems, enroll them in available outpatient treatment services, and assess treatment effects up to one year later. Research staff developed a referral network of high schools, family service agencies, and youth programs serving a large inner-city area. Staff contacted referred families by phone and offered them an opportunity to participate in a two-part baseline interview to assess the reason for study referral and discuss study enrollment. After completion of the baseline interview, adolescents who met diagnostic criteria for an APB (ODD, CD, and/or SUD) and whose families were interested in receiving outpatient treatment were randomly assigned to RFT or TAU. One TAU site that specialized in addiction treatment was withheld from randomization of ODD and CD cases, and one TAU site that did not accept substance users was withheld from randomization of SUD cases.

Parent study enrollment rates were as follows: 806 families were referred for eligibility screening; of these, 434 (54 %) completed a screen, with 264 not able to be contacted and 108 refusing to participate. Of the screen completers, 297 (68 %) were study eligible; of these, 205 (69 %) completed the baseline interview (45 could not be contacted after screening and 47 refused the baseline) and were randomized into the parent study. For the 157 cases that attended treatment, the average number of sessions for each case across the sample was 8.5 (SD = 9.8); conditions did not differ on average number of sessions attended. These rates are comparable to treatment engagement and attendance rates broadly reported for child mental health services (Garland et al. 2013).

Study Sites and Therapists

All six treatment sites were outpatient clinical settings that accepted study cases as standard referrals. Sites were in close proximity and easily accessible via public transportation. No

external training or financial support of any kind was provided to treat study cases, and therapists were not required to alter their clinical practices in any way. All therapists at each site who treated adolescent clients and who volunteered to participate were accepted into the study; approximately 40 % of therapists across sites volunteered. Each site prescribed weekly treatment sessions and offered in-house psychiatric support. Therapists at each site routinely received a comparable amount of weekly individual and/or group supervision. Details on site treatment practices are contained below in the “Treatment Fidelity” section.

Routine Family Therapy (RFT)—The RFT condition consisted of one community mental health clinic (CMHC) that featured family therapy as the standard-of-care approach for behavioral interventions with youth. At no time had the site imported a manualized FT model or contracted for extramural implementation support. RFT therapists ($n = 14$, who treated 28 cases total) were licensed Marriage and Family Therapists, social workers with training in family therapy, or advanced clinical trainees with family therapy experience. All therapists received regular in-house training and supervision to promote family-based case conceptualization and use of FT treatment techniques. Participating therapists ranged from 28 to 59 years; 8 were female, 7 were Hispanic American, and as a group they averaged 3.1 years ($SD = 4.3$) postgraduate therapy experience (note: demographic information was not collected 5 RFT therapists).

Treatment As Usual (TAU)—This condition included a set of five clinics in order to sample the full spectrum of outpatient treatment options widely available for ABPs. Among the five TAU sites were two CMHCs, two outpatient clinics in child and adolescent psychiatry departments of teaching hospitals, and one addictions treatment clinic with an adolescent program that featured group-based treatment with supportive individual sessions. Although the addiction treatment site provided self-report data, it did not permit recording of sessions due to sensitivity about treatment group participants who were not enrolled in the research study. For this reason, in the current study adolescents with substance use disorders are underrepresented in TAU (though not in RFT), so that utilization of drug counseling techniques is likely underrepresented as well compared to ITT-ABP therapist-report data previously described in Hogue et al. (2012). Among the TAU sites, only the addictions clinic espoused a strong organizational allegiance to a specific EBI—in this case, drug counseling—as revealed during focus group activities for developing the ITT-ABP (see “Measures” section). No TAU site contained a supervisor or staff therapist with extensive training in family therapy or an MFT degree, and no site appeared to promote or feature implementation of FT techniques. Across the four sites that provided data for the current study, participating therapists ($n = 11$, who treated 17 cases total) ranged in age from 25 to 39 years; 8 were female, 6 were European American, and as a group they averaged 2.9 years ($SD = 2.3$) postgraduate therapy experience.

Treatment Fidelity

RFT Adherence to Signature FT Techniques—In a separate study (Hogue and Dauber 2013) we utilized benchmarking analyses to compare FT fidelity scores for sessions held at the RFT site to fidelity scores from an efficacy trial of an empirically supported treatment for ABPs, multidimensional family therapy (MDFT). We randomly selected 15

archived RFT sessions that were videotaped on site prior to the start of the current study. We coded these sessions for adherence to FT treatment techniques using the same validated FT observational fidelity measure previously used to code MDFT sessions from the efficacy trial (the fidelity measure and MDFT scores are detailed in Hogue et al. 2006). Among the signature FT techniques coded for both samples were: convening multiple family members in sessions; specifying treatment goals that are family-based; working to bring about in-session change in family interaction patterns to decrease emotional negativity, increase positive attachments and communication, and improve family problem-solving; and intervening with caregivers to improve parenting skills. Then, using a probability sampling method known as statistical process control analysis (Deming 1986), we plotted the within-sample variance in mean FT adherence scores for the RFT site against the “benchmark” adherence data produced by MDFT. Scores for the RFT site clustered closely around the average FT adherence score for MDFT, and no score for any RFT session fell beyond two standard deviations of the benchmark MDFT mean. These analyses indicated that treatment delivered at the RFT site prior to enrollment of study cases adhered closely to gold-standard fidelity levels for FT techniques achieved by a manualized treatment.

RFT Differentiation from TAU—In previous research on the parent study (Hogue et al. 2012), we examined (a) therapist baseline reports of allegiance and skill (i.e., proficiency) in the four EBIs and (b) therapist post-session reports of EBI adherence from 822 study sessions using the ITT-ABP and found: (1) Prior to seeing study cases, RFT therapists reported strongest proficiency in FT techniques, whereas TAU therapists reported strongest proficiency in CBT and MI; (2) While treating study cases, RFT therapists reported greater utilization of treatment techniques associated with the FT approach than techniques associated with CBT, MI, or DC; (3) RFT therapists reported significantly higher average use of FT techniques than the average level reported by TAU therapists, and this difference remained significant when the TAU sample was restricted to therapists from the two CMHC sites; (4) TAU therapists reported greater use of CBT, MI, and DC interventions than did RFT therapists.

RFT and TAU Session Participants—At the end of each session study therapists were asked to document who attended the session: adolescent, caregiver, and/or other person. For the 157 sessions in the current study, in RFT 67 % of sessions included the adolescent, 80 % a caregiver, and 13 % another person. In TAU, 100 % of sessions included the adolescent, 6 % a care-giver, and 0 % another person. These percentages, which are equivalent to the participation data recorded for the parent study, indicate that RFT strongly emphasized family work, whereas TAU produced a modicum of family sessions.

Measures

Inventory of Therapy Techniques-Adolescent Behavior Problems (ITT-ABP)—

The 25-item ITT-ABP is a QA tool designed to collect post-session therapist-report data on implementation of discrete treatment techniques associated with the CBT, FT, MI, and DC approaches. ITT-ABP items (see Table 2) were derived from four observational fidelity scales for empirically supported treatments, using an instrument development process described in Hogue et al. (2012). This included two focus conducted at each study site with

all available therapists to review prospective ITT-ABP items, gather information about fit between prospective items and the treatment practices favored at each site, and trim items to create the final set of four ITT-ABP scales. The ITTABP measures EBI adherence, that is, the thoroughness/frequency with which each technique was utilized in a just-completed session, based on a 5-point Likert-type scale: 1 = *Not at all*, 2 = *A little bit*, 3 = *Moderately*, 4 = *Considerably*, 5 = *Extensively*.

ITT-ABP items representing the FT and CBT approaches were drawn from the Therapist Behavior Rating Scale (TBRs; Hogue et al. 1998), a macroanalytic observational tool designed to identify therapeutic techniques prescribed by FT and CBT for treating ABPs. The TBRs has shown strong psychometric properties in studies of treatment adherence (Diamond et al. 2007; Hogue et al. 1998, 2008a), therapist competence (Hogue et al. 2008a), and fidelity-outcome links (Hogue et al. 2006, 2008b) with samples including drug-using, conduct-disordered, and depressed teens. ITT-ABP items representing the MI and DC approaches were drawn from the Motivational Enhancement Therapy and Twelve Step Facilitation sub-scales of the Yale Adherence and Competence Scale (YACS; Carroll et al. 2000). The YACS is a general system for rating fidelity in delivering behavioral treatments for adult substance use. It has demonstrated strong reliability and validity for MI and DC fidelity scores in an efficacy trial comparing CBT, 12-step facilitation, and clinical management for adult drug users (Carroll et al. 2000) as well as multisite effectiveness trials comparing manualized MI to usual care (Ball et al. 2007; Carroll et al. 2006).

Previous construct validity analyses for this instrument (Hogue et al. 2012) used principal components analysis (PCA) on half of the 822 ITT-ABP protocols collected to that point (888 were collected by the end of the parent study), followed by confirmatory factor analysis on the remaining half that yielded adequate fit indices: $\chi^2(272) = 388.01$, $p < .001$; RMSEA = .03 (90 % CI: .025–.039); CFI = .96; TLI = .96. This process identified three clinically coherent scales with strong internal consistency: FT scale (8 items: PCA item-factor loading range .73–.46, Cronbach's $\alpha = .79$), MI/CBT scale (8 items: PCA range .81–.52, $\alpha = .87$), and DC scale (9 items: PCA range .97–.44, $\alpha = .90$).

Inventory of Therapy Techniques-Adolescent Behavior Problems:

Observational Version (ITT-ABP-O)—The ITT-ABP-O contains 25 items for rating adherence to CBT, FT, MI, and DC techniques that are identical to those on the ITT-ABP. The observational coding manual contains lengthy descriptions and exemplars of each item to aid coders in making reliable ratings; the content for this manual was taken directly from the source observational fidelity scales, with minimal editing of YACS items for age appropriateness.

Fidelity Data Collection Procedures and Sampling Rates

ITT-ABP Data—Prior to seeing study cases, study therapists participated in a one-hour group session introducing the purpose of the ITT-ABP and briefly reviewing all items. At this meeting, therapists were given an ITT-ABP scoring manual containing one-page descriptions of every item, including exemplar statements of what a therapist might say in session when implementing the given technique; item descriptions and exemplars were

distilled from the corresponding item materials contained in the TBRS and YACS coding manuals. Therapists were then asked to complete a checklist after every session with study clients. During the study, therapists were prompted to submit completed checklists for active cases on a weekly basis but received no further ITT-ABP training and no feedback on submitted data. Over the course of the parent study 888 ITT-ABP checklists were submitted (57 % for RFT sessions, 43 % for TAU).

ITT-ABP-O Data—At the beginning of the study and periodically throughout, therapists were asked to videorecord or (at client request) audiorecord half of their treatment sessions. Thus the study attempted to record 444 sessions of the 888 sessions for which there was an ITT-ABP checklist. A total of 157 recordings were ultimately collected (104 RFT sessions, 53 TAU sessions), which represents 35 % of the 444 sessions targeted for recording. Of the 287 checklist sessions targeted for recording but not ultimately collected, 118 (41 %) belonged to families who did not consent to allow recordings; 75 (26 %) belonged to cases treated in the addiction clinic of the TAU condition, which did not permit recordings; and 95 (33 %) were not recorded for miscellaneous reasons that included therapist negligence, change in therapist and/or family agreeability to record, and equipment problems. Recordings included 66 % RFT and 34 % TAU sessions; 24 % were audiorecordings; 44 % were from the Early phase of treatment (sessions 1–5) and 56 % from the Later phase (session 6 or higher). In RFT, the 14 therapists provided a range of 2–19 recordings, and the 28 cases yielded a range of 1–19 recordings; in TAU, the 11 therapists provided a range of 1–21 recordings, and the 17 cases yielded a range of 1–21 recordings. No between-condition differences were found for total number of sessions per client or for number of early sessions or later sessions.

Observational Coding Procedures

Observational Raters—ITT-ABP-O raters ($n = 8$) were research staff with master's level training in social work, psychology, or a related field. The group contained 5 European American women, 2 European American men, 1 African American woman, 1 Hispanic American woman, and 1 Asian American woman. None had prior fidelity coding experience.

Training and Coding Procedures—Raters were trained during weekly small-group meetings over the course of two months using review of the ITTABP-O coding manual, in-group coding and review of practice recordings, and exercises to increase understanding of scale items. Study coding commenced once all raters reached a collective threshold reliability of $ICC = .65$ for the preponderance of items; thereafter, the group met biweekly for supportive training and monitoring of rater drift until coding was completed. Sessions were coded in their entirety (average about 55 min). Two raters were assigned to code all recordings and were randomly paired with one another across the sample using a randomized block design (Fleiss 1981).

Plan of Analysis

Inter-observer reliability was calculated for each item using the intraclass correlation coefficient (ICC; Shrout and Fleiss 1979). After acceptable ICCs were established, item

scores were averaged across observational raters to yield a single observer score for each item. Item-level reliability between observers and therapists was then calculated using the ICC. Following item-level analysis, item scores were averaged to create the FT factor scale and the MI/CBT factor scale. As described in the “Measures” section, the drug counseling items were not included in analyses due to negligible rates of endorsement in the study sample. All remaining analyses were conducted on the averaged factor scores. To examine reliability of therapist judgments, therapist-observer correspondence on both factor scores was calculated using the ICC for the full sample, separately by Early versus Later phase of treatment. To test for differences in the relative magnitude of ICCs (i.e., strength of correlation) between two sets of raters, ICCs were compared using an *r*-to-*z* transformation and a subsequent *z*-test (Thorndike 2007). To examine therapist accuracy for the full sample and within treatment phase, therapist and observer mean factor scores were compared using repeated measures ANOVA that included Therapist as a covariate to control for between-therapist effects. Then, two sets of within-condition analyses were conducted to determine whether therapists showed differential facility in rating approach-concordant (i.e., RFT therapists rating FT, TAU therapists rating MI/CBT) versus approach-discordant (i.e., RFT therapists rating MI/CBT, TAU therapists rating FT) techniques: therapist-observer ICCs to examine differential reliability, followed by comparisons of therapist versus observer mean scores to examine differential accuracy. Finally, differential reliability was tested in Novice versus Veteran therapists within both study conditions.

Results

Preliminary Analyses: Inter-Observer Reliability

One-way random ICCs were calculated for each item on the FT and MI/CBT scales, and also for the mean scale scores, to examine inter-observer reliability (see Table 2). According to Cicchetti's (1994) criteria for classifying the utility of ICC magnitudes, below .40 is poor, .40–.59 is fair, .60–.74 is good, and .75–1.00 is excellent. ICCs ranged from .63 to .88 for the FT items, demonstrating strong agreement between observational raters; the ICC for the mean scale score of all 8 FT items was .87. For the MI/CBT items, ICCs were good to excellent for most items, ranging from .60 to .75. Two items on the MI/CBT scale had only fair reliability but were retained in order to preserve the integrity of the original ITT-ABP scale: Reflective Listening had an ICC of .54 and Cognitive Monitoring had an ICC of .48. The inter-observer ICC for the mean scale score of all 8 MI/CBT items was .76. Given the overall high agreement, scores from the two raters were averaged to obtain a single observer rating for each item; these averaged scores were used in all remaining analyses.

Reliability of Therapist Judgment: Item-Level Analyses

Reliability between observers and therapists was examined with one-way random ICCs comparing the observational and therapist-reported scores on each item. Analyses are presented in Table 2. Among the FT items, therapists and observers achieved good reliability for Parental Monitoring (ICC = .72) and Coached Interactions (ICC = .70). The ICC for Family Attachment and Communication was near the acceptable range (ICC = .52). All other items on the FT scale showed weak reliability. On the MI/CBT scale, only one

item had acceptable reliability between therapists and observers: Behavioral Interventions had an ICC of .53. All other items on this scale showed weak reliability.

Reliability and Accuracy of Therapist Judgment: Scale-Level Analyses

Therapist and observer mean scale scores and ICCs are presented in Table 3 for the full sample and separately by treatment phase: Early phase = sessions 1–5, Later phase = sessions 6 and above. For the FT scale, therapist reliability was adequate in the full sample (ICC = .66) as well as for both Early phase (ICC = .69) and Later phase (ICC = .63) sessions. FT scale reliability was even stronger (ICC = .75) when deleting the outlier item Established Agenda, which had a rare negative reliability (ICC = −.62) indicative of discrepancies in basic item definition between therapists and observers. We then tested the strength of correlation for the inter-observer ICC (.87) versus the therapist-observer ICC (.66) for the full sample; results ($z = 4.74, p < .001$) demonstrated that observers agreed to a greater extent with one another than with therapists. For the MI/CBT scale, reliability was almost non-existent for the full sample (ICC = .06) as well as for Early phase (ICC = .10) and Later phase (ICC = .01) sessions. Again, the magnitude of correlation for the inter-observer ICC (.76) was significantly higher than that of the therapist-observer ICC (.06) for the full sample ($z = 8.21, p < .001$).

To assess therapist accuracy, repeated measures ANOVA were conducted to compare therapist and observer mean scores on the FT and MI/CBT scales for the full sample and separately by phase. For the FT scale, therapists rated themselves significantly higher on FT interventions compared to observers in the full sample ($F(1, 155) = 20.06, p < .001, \eta_p^2 = .12$), as well as for Early phase ($F(1, 67) = 13.42, p < .001, \eta_p^2 = .17$) and Later phase ($F(1, 86) = 7.48, p < .01, \eta_p^2 = .08$). Similarly, therapists rated themselves significantly higher than observers on MI/CBT interventions in the full sample ($F(1, 155) = 13.54, p < .001, \eta_p^2 = .08$) and for Early phase sessions ($F(1, 67) = 14.71, p < .001, \eta_p^2 = .18$). For later phase sessions, the difference was not significant ($F(1, 86) = 2.18, p = .14$).

Differential Reliability and Accuracy for Approach-Concordant Versus Approach-Discordant EBIs

Therapist-observer ICCs within the RFT and TAU conditions, as well as therapist and observer mean scores on the FT and MI/CBT scales, were examined to determine whether therapists demonstrated differential reliability and accuracy for approach-concordant versus approach-discordant EBIs (see Table 3). RFT therapists, all of whom were family therapists by training, demonstrated higher reliability for FT interventions (ICC = .52 for all items, and .64 after deleting Established Agenda) compared to MI/CBT interventions (ICC = .13), suggesting that family therapists did demonstrate more reliability for approach-concordant interventions. In contrast, TAU therapists, none of whom were family therapists, did not achieve good reliability for either FT (ICC = .39 for all items, .57 after deleting Established Agenda) or MI/CBT (ICC = −.06) interventions, although reliability was noticeably higher for FT interventions, indicating that non-family therapists were in relatively better agreement with observers about implementing FT techniques than techniques associated with other theoretical orientations.

Differential accuracy was also examined by comparing within-condition therapist and observer mean scores representing the extensiveness of EBI implementation. Within RFT, therapists rated themselves significantly higher than observers on both FT interventions ($F(1, 102) = 19.28, p < .001, \eta_p^2 = .16$) and MI/CBT interventions ($F(1, 102) = 16.11, p < .001, \eta_p^2 = .14$). Unexpectedly, within TAU there was no significant difference between therapists and observers on FT ($F(1, 51) = .46, p = .50$) or MI/CBT ($F(1, 51) = .00, p < .99$).

Differential EBI Implementation and Self-Report Reliability for Novice versus Veteran Therapists

To examine whether therapist reliability differed according to years of professional experience, we created two subgroups: Novice versus Veteran clinicians. Of the 25 therapists in the study, 15 (9 RFT, 6 TAU) were placed in the Novice subgroup ($M_{\text{experience}} = 1.0$ years, $SD = .65$, range 0–2) and 10 (5 RFT, 5 TAU) in the Veteran subgroup ($M_{\text{experience}} = 5.2$ years, $SD = 3.2$, range 3–13); as intended the Veteran subgroup was significantly more experienced ($t(9.5) = -4.2, p < .05$). Table 4 contains the mean scale scores for FT and MI/CBT interventions reported by each subgroup and by observational coders separately by condition, along with corresponding therapist-observer ICCs. According to observer report there were no subgroup differences whatsoever in utilization of EBIs: Within RFT, novices and veterans were judged to be comparable in the extent to which they implemented both FT ($F(2, 103) = 1.39, p = .24$) and MI/CBT ($F(2, 103) = .54, p = .46$); within TAU, the subgroups were also rated as comparable in delivery of FT ($F(2, 52) = .26, p = .62$) and MI/CBT ($F(2, 52) = .61, p = .44$). The self-report ratings provided by TAU therapists showed the same pattern: Novice and Veteran scores indicated a comparable utilization of both FT ($F(2, 52) = .82, p = .37$) and MI/CBT ($F(2, 52) = 2.29, p = .14$). Departing from this trend, Veteran therapists in RFT reported greater utilization of FT ($F(2, 103) = 17.76, p < .001, \eta_p^2 = .15$) and also MI/CBT ($F(2, 103) = 45.54, p < .001, \eta_p^2 = .30$) than their Novice counterparts.

Of greatest interest was whether Novices and Veterans were differentially reliable in reporting on EBI delivery. Results for the FT scale were intriguing: Within RFT, Novices ($ICC = .55$) were moderately reliable overall and marginally more reliable than Veterans ($ICC = .33; z = 1.78, p = .07$); in contrast, TAU Veterans ($ICC = .62$) were much more reliable than their basically unreliable Novice counterparts ($ICC = .17; z = 3.23, p < .001$). For the MI/CBT scale, similar to the full-sample results reported above, therapist self-reports showed uniformly poor reliability across subgroups and conditions (see Table 4).

Discussion

Study results indicated that family therapists working under routine practice conditions were fairly reliable overall with trained observers in reporting on the extent to which they implemented family therapy interventions, though unreliable in reporting on MI/CBT interventions. Non-family therapists also showed relatively stronger correspondence with observers for FT interventions than for MI/CBT, although in absolute terms, their reliability was low for FT and non-existent for MI/CBT. Family therapists also significantly

overestimated their use of FT and MI/CBT relative to observer ratings, though non-family therapists did not. For the most part, therapist reliability and accuracy did not differ across phase of treatment.

These results provide promising support for the feasibility of using therapist-report adherence monitoring methods to anchor QA procedures for family therapy interventions in real-world settings. In this study, agency-hired family therapists were reliable with observational raters in judging the extent to which they delivered core FT techniques when treating community-based referrals for adolescent problem behaviors. Interestingly, novice family therapists showed greater agreement with raters than did veterans, perhaps because they were more conservative overall in rating themselves, although the small subgroup sizes preclude confident generalizations. Given that study therapists received only a single training session in using the self-report coding system, there is reason to believe that their reliability could be substantially improved if ongoing training and feedback in reliable documentation of FT adherence were to be incorporated into routine QA procedures (Hogue et al. 2013; Ward et al. 2013).

Beyond therapist *reliability*, it may be more challenging to boost therapist *accuracy* in self-reports. As true for every therapist-report study of EBI adherence to date involving either research-hired (e.g., Carroll et al. 1998; Martino et al. 2009) or agency-hired (e.g., Brosan et al. 2008; Hurlburt et al. 2010) clinicians, we found that family therapists overestimated the extent to which they implemented both FT and MI/CBT techniques. That is, family therapists could reliably document their relative use of FT techniques in any given session—more versus less FT—but they consistently overstated the actual quantity of FT as documented by observers. This kind of EBI adherence score inflation appears deeply rooted in benign reporter biases of several kinds, for example, perceived effort in delivering an intervention, more inclusive personal framework for evaluating an intervention—biases that also might, from the reverse perspective, contribute to some degree of observer under-reporting. Such biases may be most sensibly approached as a fundamentally intractable factor. If so, the best option might be to develop inflation-adjustment formulae that can be applied using national or local norms. This would allow for a favorable (or at least workable) compromise when using therapist reports of EBI adherence for QA purposes. Notably, TAU therapists did not significantly overestimate their use of EBI techniques, though the modest number of sessions sampled in this condition may have dampened ability to detect reporter differences.

Results for therapist reports of MI/CBT interventions were more discouraging. Although this study assessed the degree to which community therapists used MI/CBT interventions in treating APB youth, which has intrinsic value for documenting the extent to which EBIs for a given disorder appear in everyday care, neither study condition contained therapists who made pre-study claims about practicing MI/CBT exclusively or routinely. Even so, the current sample of TAU therapists was unreliable in documenting their de facto implementation of MI/CBT, and this was equally true for both family and non-family therapists, and for both veteran and novice therapists across study condition. Perhaps FT is better suited for therapist-report formats because most FT techniques are linked to the physical presence of family members; perhaps the techniques themselves are more easily

operationalized in a manner that avails reliable self-reporting; perhaps this particular sample of community therapists lacked an adequate level of expertise in the MI and CBT approaches to engender reliable reporting. Whatever the reasons, given past failures to establish therapist reliability in fidelity reporting on MI (e.g., Martino et al. 2009; Miller et al. 2004) and CBT (e.g., Brosan et al. 2008; Carroll et al. 1998) across various client populations, the field remains stuck at the drawing board in creating viable therapist-report tools for these two approaches, both of which are widely endorsed in the clinical workforce (Cook et al. 2010; Gifford et al. 2012).

Study Strengths and Limitations

The main study strength was high ecological validity: It sampled community therapists who worked without benefit of extramural resources and who treated clients diagnosed with an array of clinical disorders for whom comorbidity was the norm. Also, therapists completed the brief ITTABP measure after participating in a one-hour introductory training, as part of their routine session documentation responsibilities. These conditions support the generalizability of the findings to real-world practice.

There were two major study limitations. First, there was a small number of participating sites and therapists, which prevented testing or controlling for site differences, did not supply a representative sample of the usual care provider workforce, and generated minimal power to detect differences between subgroups of therapists (novices versus veterans), which disallows firm conclusions regarding the influence of clinician experience on self-report reliability. Second, the collection of recorded sessions was decidedly non-random: Study therapists were asked to record only half of their sessions and free to decide which sessions to record; moreover, investigators obtained recordings for only 35 % of the requested total. These sampling gaps open the door to sampling biases of several kinds (e.g., over-representation of therapist-preferred clients and sessions, underrepresentation of clients with erratic attendance or who refused to be recorded) and thus significantly limit the generalizability of the sampled data to the intended population. Likewise, only about 40 % of all site therapists volunteered to participate. Also, due to treatment attrition the study sample was younger and contained fewer adolescents with multiple diagnoses than the originally randomized sample.

Analyses did control for between-therapist effects when comparing mean scores but did not directly investigate individual therapist differences in self-report scores due to lack of power. Approach concordance and discordance comparisons were acutely drawn within RFT—all therapists in this condition had strong allegiance to the FT approach—but less so within TAU, as those therapists did not uniformly espouse a strong allegiance to MI/CBT (as reported in Hogue et al. 2012). The RFT condition submitted twice as many recordings as TAU, an unintended discrepancy that may have resulted from the larger critical mass of having 14 RFT therapists located in one site versus 11 TAU therapists spread across five sites. This afforded a comparably greater opportunity to evaluate FT techniques than MI/CBT techniques.

Study Implications

Study results provide partial evidence for the criterion validity of the ITT-ABP: Community therapists were reliable and consistent with observational raters in reporting on EBI adherence for the FT scale, but not the MI/CBT scale. If the FT scale remains valid when tested in larger and more diverse therapist samples, it would gain standing as a valuable new QA tool in efforts to ensure quality implementation of FT in routine care for ABPs. Specifically, the FT scale could serve both prescriptive and evaluative purposes in the difficult task of achieving fidelity in front-line settings (see Regan et al. 2013). Therapists could guide core-element case planning (see Barth et al. 2014) for ABP cases based on the FT techniques contained in the scale. Supervisors, administrators, and therapists themselves could utilize routine therapist-report documentation of FT fidelity to establish benchmarks for quality performance and provide corrective feedback at the case or clinician level (McLeod et al. 2013). Measuring EBI fidelity will become steadily more important in the wake of system-level changes in health-care that include tracking and incentivizing the use of multilevel quality indicators to assess competence in service delivery (Hoagwood et al. 2013).

As a brief therapist-report fidelity measure, the ITT-ABP is intended to be cost-efficient and readily manageable using intramural agency resources (Hogue et al. 2013). Moreover, as demonstrated in this study, the FT scale can be used reliably with minimal training. These features align with the instrument's modest degree of standardization: The eight FT items refer to generic treatment techniques that are not embedded within procedural guidelines for when, how, or with whom to implement them. Thus there is concordance between the design of the instrument and the (lack of) of standardization among the “loose confederation” of interventions that it measures—it is intended to identify a generic approach, not a specific model. In contrast, the fidelity tools used by therapists implementing the modular MATCH-ADC protocol (Weisz et al. 2012) were technically elaborate and procedure-driven (Ward et al. 2013), matching the extensive standardization of the protocol itself. These two examples underscore a general principle: Fidelity measures and procedures can and should be calibrated according to the degree of specificity, flexibility, and customization needed to effectively capture the treatment constructs of interest and meet the desired evaluation goals (Regan et al. 2013).

Finally, the FT scale of the ITT-ABP might be profitably combined with the therapist-report scale for family-based contingency management of adolescent substance use recently validated by Chapman et al. (2013), thereby creating a broad package of therapist-report fidelity assessment technology to guide implementation of evidence-based, family-based approaches for ABPs. This fidelity package for the FT approach would offer a large boost to clinicians, agencies, and systems struggling to deliver high-quality family intervention services to multiproblem adolescents in various sectors of care (Kazak et al. 2010).

Acknowledgments

Preparation of this article was supported by grant R01DA023945 from the National Institute on Drug Abuse. The authors would like to acknowledge the dedicated work of the observational coders for this project (in addition to authors EL and MB): Daniela Caraballo, Benjamin Goldman, Jacqueline Horan, Candace Johnson, Akiyo Kodera, Emily McSpadden, Annie Niermann, and Gabi Spiewak. We are also extremely grateful to the community

therapists working at six study sites who generously agreed to participate in a research study and submit recordings of their sessions in the hopes of advancing the clinical science in their field.

References

- Aarons GA, Hurlburt M, Horowitz S. Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health and Mental Health Services Research*. 2011; 38:4–23. [PubMed: 21197565]
- American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed., text revision. APA; Washington, DC: 2000.
- Baldwin SA, Christian S, Berkeljon A, Shadish W, Bean R. The effects of family therapies for adolescent delinquency and substance abuse: A meta-analysis. *Journal of Marital and Family Therapy*. 2012; 38:281–304. [PubMed: 22283391]
- Ball SA, Martino S, Nich C, Frankforter TL, Van Horn D, Crits-Christoph P, et al. Site matters: Multisite randomized trial of motivational enhancement therapy in community drug abuse clinics. *Journal of Consulting and Clinical Psychology*. 2007; 75:556–567. [PubMed: 17663610]
- Barth RP, Kolivoski KM, Lindsey MA, Lee BR, Collins KS. Translating the common elements approach: Social work's experiences in education, practice, and research. *Journal of Clinical Child and Adolescent Psychology*. 2014; 43(2):301–311. [PubMed: 24245958]
- Bearman SK, Weisz JR, Chorpita BF, Hoagwood K, Ward A, Ugueto AM, et al. More practice, less preach? The role of supervision processes and therapist characteristics in EBP implementation. *Administration and Policy in Mental Health and Mental Health Services Research*. 2013; 40:518–529. [PubMed: 23525895]
- Bearsley-Smith C, Sellick K, Chesters J, Francis K. Treatment content in child and adolescent mental health services: Development of the Treatment Recording Sheet. *Administration and Policy in Mental Health and Mental Health Services Research*. 2008; 35:423–435. [PubMed: 18679789]
- Becker S, Curry J. Outpatient interventions for adolescent substance abuse: A quality of evidence review. *Journal of Consulting and Clinical Psychology*. 2008; 76:531–543. [PubMed: 18665683]
- Brosan L, Reynolds S, Moore RG. Self-evaluation of cognitive therapy performance: Do therapists know how competent they are? *Behavioural and Cognitive Psychotherapy*. 2008; 36:581–587.
- Carroll KM, Ball SA, Nich C, Martino S, Frankforter TL, Farentinos C, et al. Motivational interviewing to improve treatment engagement and outcome in individuals seeking treatment for substance abuse: A multisite effectiveness study. *Drug and Alcohol Dependence*. 2006; 81:301–312. [PubMed: 16169159]
- Carroll KM, Nich C, Rounsaville BJ. Utility of therapist session checklists to monitor delivery of coping skills treatment for cocaine abusers. *Psychotherapy Research*. 1998; 8:307–320.
- Carroll KM, Nich C, Sifry R, Nuro KF, Frankforter TL, Ball SA, et al. A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug & Alcohol Dependence*. 2000; 57:225–238. [PubMed: 10661673]
- Chapman JE, McCart MR, Letourneau EJ, Sheidow AJ. Comparison of youth, caregiver, therapist, trained, and treatment expert raters of therapist adherence to a substance abuse treatment protocol. *Journal of Consulting and Clinical Psychology*. 2013; 81:674–680. [PubMed: 23668668]
- Chorpita, BF. Modular cognitive-behavioral therapy for childhood anxiety disorders. Guilford; New York: 2007.
- Chorpita BF, Becker KD, Daleiden EL. Understanding the common elements of evidence-based practice: Misconceptions and clinical examples. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2007; 46:647–652. [PubMed: 17450056]
- Chorpita B, Daleiden E, Ebesutani C, Young J, Becker K, Nakamura B, et al. Evidence-based treatments for children and adolescents: An updated review of indicators of efficacy and effectiveness. *Clinical Psychology: Science and Practice*. 2011; 18:154–172.
- Chorpita BF, Daleiden EL, Weisz JR. Modularity in the design and application of therapeutic interventions. *Applied & Preventive Psychology*. 2005; 11:141–156.
- Chorpita BF, Weisz JR, Daleiden EL, Schoenwald SK, Palinkas LA, Miranda J, et al. Long-term outcomes for the Child STEPs randomized effectiveness trial: A comparison of modular and

- standard treatment designs with usual care. *Journal of Consulting and Clinical Psychology*. 2013; 81:999–1009. [PubMed: 23978169]
- Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*. 1994; 6:284–290.
- Cook JM, Biyanova T, Elhai J, Schnurr PP, Coyne JC. What do psychotherapists really do in practice? An internet study of over 2,000 practitioners. *Psychotherapy: Theory Research, Practice, Training*. 2010; 47:260–267.
- Deming, WE. *Out of the crisis*. MIT Press; Cambridge, MA: 1986.
- Dennis ML, Godley SH, Diamond G, Tims FM, Babor T, Donaldson J, et al. The Cannabis Youth Treatment (CYT) study: Main findings from two randomized trials. *Journal of Substance Abuse Treatment*. 2004; 27:197–213. [PubMed: 15501373]
- Diamond GM, Diamond GS, Hogue A. Attachment-based family therapy: Adherence and differentiation. *Journal of Marital and Family Therapy*. 2007; 33:177–191. [PubMed: 17437458]
- Embry DD, Biglan A. Evidence-based kernels: Fundamental units of behavioral influence. *Clinical Child and Family Psychology Review*. 2008; 11:75–113. [PubMed: 18712600]
- Fixsen, DL.; Naoom, SF.; Blasé, KA.; Friedman, RM.; Wallace, F. *Implementation research: A synthesis of the literature*. University of South Florida, Louis de la parte Florida Mental Health Institute, The National Implementation Research Network; Tampa, FL: 2005. (FMHI Publication #231)
- Fleiss JL. Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement*. 1981; 5:105–112.
- Gallo KP, Barlow DH. Factors involved in clinician adoption and nonadoption of evidence-based interventions in mental health. *Clinical Psychology: Science and Practice*. 2012; 19:93–106.
- Garland AF, Bickman L, Chorpita BF. Change what? Identifying quality improvement targets by investigating usual mental health care. *Administration and Policy in Mental Health and Mental Health Services Research*. 2010a; 37:15–26. [PubMed: 20177769]
- Garland AF, Haine-Schlagel R, Brookman-Frazee L, Baker-Ericzen M, Trask E, Fawley-King K. Improving community-based mental health care for children: Translating knowledge into action. *Administration and Policy in Mental Health and Mental Health Services Research*. 2013; 40:6–22. [PubMed: 23212902]
- Garland AF, Hawley KM, Brookman-Frazee LI, Hurlburt M. Identifying common, core elements of evidence-based practice for children with disruptive behavior disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2008; 47:505–514. [PubMed: 18356768]
- Garland AF, Hurlburt MS, Brookman-Frazee L, Taylor RM, Accurso EC. Methodological challenges of characterizing usual care psychotherapeutic practice. *Administration and Policy in Mental Health and Mental Health Services Research*. 2010b; 37:208–220. [PubMed: 19757021]
- Gifford EV, Tavakoli S, Weingardt KR, Finney JW, Pierson HM, et al. How do components of evidence-based psychological treatment cluster in practice? A survey and cluster analysis. *Journal of Substance Abuse Treatment*. 2012; 42:45–55. [PubMed: 21943809]
- Hawkins EH. A tale of two systems: Co-occurring mental health and substance abuse disorders treatment for adolescents. *Annual Review of Psychology*. 2009; 60:197–227.
- Hepner KA, Azocar F, Greenwood GL, Miranda J, Burnam MA. Development of a clinician report measure to assess psychotherapy for depression in usual care settings. *Administration and Policy in Mental Health and Mental Health Services Research*. 2010; 37:221–229. [PubMed: 20186569]
- Hoagwood K, Atkins M, Ialongo N. Unpacking the black box of implementation: The next generation of policy, research and practice. *Administration and Policy in Mental Health and Mental Health Services Research*. 2013; 40:451–455. [PubMed: 23942647]
- Hogue A, Dauber S. Assessing fidelity to evidence-based practices in usual care: The example of family therapy for adolescent behavior problems. *Evaluation and Program Planning*. 2013; 37:21–30. [PubMed: 23314000]
- Hogue A, Dauber S, Chinchilla P, Fried A, Henderson CE, Inclan J, et al. Assessing fidelity in individual and family therapy for adolescent substance abuse. *Journal of Substance Abuse Treatment*. 2008a; 35:137–147. [PubMed: 17997268]

- Hogue A, Dauber S, Henderson CE. Therapist self-report of evidence-based practices in usual care for adolescent behavior problems: Factor and construct validity. *Administration and Policy in Mental Health and Mental Health Services Research*. 2012; 41(1):126–139. doi:10.1007/s10488-012-0442-8. [PubMed: 23124275]
- Hogue A, Dauber S, Samuolis J, Liddle HA. Treatment techniques and outcomes in multidimensional family therapy for adolescent behavior problems. *Journal of Family Psychology*. 2006; 20:535–543. [PubMed: 17176187]
- Hogue A, Henderson CE, Dauber S, Barajas P, Fried A, Liddle HA. Treatment adherence, competence, and outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology*. 2008b; 76:544–555. [PubMed: 18665684]
- Hogue A, Liddle HA. Family-based treatment for adolescent substance abuse: Controlled trials and new horizons in services research. *Journal of Family Therapy*. 2009; 31:126–154. [PubMed: 21113237]
- Hogue A, Liddle H, Rowe C, Turner R, Dakof GA, LaPann K. Treatment adherence and differentiation in individual versus family therapy for adolescent drug abuse. *Journal of Counseling Psychology*. 1998; 45:104–114.
- Hogue A, Ozechowski TJ, Robbins MR, Waldron HB. Making fidelity an intramural game: Localizing quality assurance procedures to promote sustainability of evidence-based practices in usual care. *Clinical Psychology: Science and Practice*. 2013; 20:60–77.
- Hurlburt MS, Garland AF, Nguyen K, Brookman-Frazee L. Child and family therapy process: Concordance of therapist and observational perspectives. *Administration and Policy in Mental Health and Mental Health Services Research*. 2010; 37:230–244. [PubMed: 19902347]
- Kazak A, Hoagwood K, Weisz JR, Hood K, Kratochwill T, Vargas L, et al. A meta-systems approach for evidence-based practice for children and adolescents. *American Psychologist*. 2010; 65:85–97. [PubMed: 20141264]
- Korathu-Larson, P.; Park, A.; Fierstein, M.; Chorpita, BF. Therapist satisfaction with multiple designs of evidence-based treatments: Did therapists match with MATCH?. The Association of Behavioral and Cognitive Therapies (ABCT); Toronto, Canada: 2011.
- Martino S, Ball S, Nich C, Frankforter TL, Carroll KM. Correspondence of motivational enhancement treatment integrity ratings among therapists, supervisors, and observers. *Psychotherapy Research*. 2009; 19:181–193. [PubMed: 19396649]
- McHugh R, Barlow DH. The dissemination and implementation of evidence-based psychological treatment: A review of current efforts. *American Psychologist*. 2010; 65:73–84. [PubMed: 20141263]
- McLeod BD, Southam-Gerow MA, Tully CB, Rodriguez A, Smith MM. Making a case for treatment integrity as a psychosocial treatment quality indicator for youth mental health care. *Clinical Psychology: Science and Practice*. 2013; 20:14–32. [PubMed: 23935254]
- Miller WR, Yahne CE, Moyers TB, Martinez J, Pirritano M. A randomized trial of methods to help clinicians learn motivational interviewing. *Journal of Consulting and Clinical Psychology*. 2004; 72:1050–1062. [PubMed: 15612851]
- O'Neill SM, Hempel S, Lim YW, Danz MS, Foy R, Suttorp MJ, et al. Identifying continuous quality improvement publications: What makes an improvement intervention 'CQI'? *British Medical Journal Quality & Safety*. 2011; 20:1011–1019.
- Perepletchikova F. On the topic of treatment integrity. *Clinical Psychology: Science and Practice*. 2011; 18:148–153. [PubMed: 21769167]
- Regan J, Daleiden EL, Chorpita BF. Integrity in mental health systems: An expanded framework for managing uncertainty in clinical care. *Clinical Psychology: Science and Practice*. 2013; 20:78–98.
- Schoenwald SK. It's a bird, it's a plane, it's...fidelity measurement in the real world. *Clinical Psychology: Science and Practice*. 2011; 18:142–147. [PubMed: 21691439]
- Schoenwald SK, Garland AF, Chapman JE, Frazier SL, Sheidow AJ, Southam-Gerow MA. Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*. 2011; 38:32–43. [PubMed: 20957425]

- Schoenwald SK, Letourneau EJ, Halliday-Boykins C. Predicting therapist adherence to a transported family-based treatment for youth. *Journal of Clinical Child and Adolescent Psychology*. 2005; 34(4):658–670. [PubMed: 16232063]
- Schoenwald SK, Sheidow AJ, Chapman JE. Clinical supervision in treatment transport: Effects on adherence and outcomes. *Journal of Consulting and Clinical Psychology*. 2009; 77:410–421. [PubMed: 19485583]
- Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini International Neuropsychiatric Interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*. 1998; 59:22–33. [PubMed: 9881538]
- Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*. 1979; 86:420–428. [PubMed: 18839484]
- Thorndike, R. Fisher's Z transformation.. In: Salkind, N., editor. *Encyclopedia of measurement and statistics*. Sage; Thousand Oaks, CA: 2007. p. 361-365.
- Ward AM, Regan J, Chorpita BF, Starace N, Rodriguez A, Okamura K, et al. Tracking evidence based practice with youth: Validity of MATCH and standard manual consultation records. *Journal of Clinical Child and Adolescent Psychology*. 2013; 42:44–55. [PubMed: 22809135]
- Weersing RV, Weisz JR, Donenberg GR. Development of the Therapy Procedures Checklist: A therapist-report measure of technique use in child and adolescent treatment. *Journal of Clinical Child Psychology*. 2002; 31:168–180.
- Weisz JR, Chorpita BF, Palinkas LA, Schoenwald SK, Miranda J, Bearman SK, et al. Testing standard and modular designs for psychotherapy treating depression, anxiety, and conduct problems in youth: A randomized effectiveness trial. *Archives of General Psychiatry*. 2012; 69:274–282. [PubMed: 22065252]
- Westen D, Novotny CM, Thompson-Brenner H. EBP does not equal EST: Reply to Crits-Christoph et al. (2005) and Weisz et al. (2005). *Psychological Bulletin*. 2005; 131:427–433.
- Winters KC, Stinchfield R, Latimer WW, Lee S. Long-term outcome of substance-dependent youth following 12-step treatment. *Journal of Substance Abuse Treatment*. 2007; 33:61–69. [PubMed: 17588490]

Table 1

Demographic and clinical characteristics: full sample and study condition differences

	Full sample <i>n</i> = 45	RFT <i>n</i> = 28	TAU <i>n</i> = 17
Male	53 %	50 %	59 %
Adolescent age (M/SD)	15.2 (1.5)	15.3 (1.7)	15.0 (1.3)
Ethnicity			
Hispanic	69 %	79 %	53 %
Black	11 %	7 %	18 %
More than one race	13 %	11 %	18 %
Other race	7 %	4 %	12 %
Family composition			
Single parent	65 %	74 %	50 %
Two parents	23 %	15 %	38 %
Grandparent	9 %	11 %	6 %
Other	2 %	0 %	6 %
Caregiver graduated high school *	67 %	54 %	88 %
Caregiver employed	59 %	52 %	71 %
Caregiver income greater than \$15 K	50 %	41 %	65 %
Caregiver receiving public assistance	13 %	18 %	6 %
Oppositional defiant disorder	91 %	89 %	94 %
Conduct disorder	46 %	48 %	41 %
Attention-deficit/hyperactivity disorder	76 %	73 %	81 %
Depression diagnosis	41 %	44 %	35 %
Substance use disorder	23 %	30 %	12 %
Generalized anxiety disorder	16 %	15 %	18 %
Posttraumatic stress disorder	14 %	11 %	18 %
More than one diagnosis	74 %	77 %	71 %
Picked up by police past year	22 %	25 %	18 %
Probation/parole past year	2 %	0 %	6 %
# of delinquent acts past month (M/SD)	2.9 (3.0)	2.9 (3.4)	2.9 (2.4)
Days used substances past month (M/SD)	3.0 (7.4)	3.9 (8.7)	1.1 (2.6)

All data reported are column percents, unless otherwise indicated

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

Table 2

Item-level inter-observer reliability, and therapist-observer correspondence, on ITT-ABP-O items in the FT and MI/CBT factor scales

	Inter-observer ICC	Therapist-observer ICC
Family therapy (FT) scale		
Established agenda	.75	-.62
Parental monitoring	.77	.72
Prepared for family session	.69	.23
Coached family interactions	.88	.70
Family attachment and communication	.82	.52
Normative adolescent development	.76	.30
Targeted family members for change	.72	.46
Core relational family themes	.63	.31
MI/CBT scale		
Affirmed client's ability to change	.75	.45
Reflective listening	.54	.30
Explored readiness to change	.60	.01
Behavioral interventions	.63	.53
Taught coping skills	.60	.12
Awareness of discrepancies	.65	.23
Commitment to change plan	.66	.23
Cognitive monitoring	.48	.09

Table 3

Therapist and observer mean scores and reliability on the ITT-ABP-O factor scales: full sample, by condition, and by treatment phase

	Family therapy scale			MI/CBT scale		
	Therapist Mean (SD)	Observer Mean (SD)	ICC	Therapist Mean (SD)	Observer Mean (SD)	ICC
Full sample	2.18 (.68)	1.8 (.52)	.66	2.08 (.62)	1.6 (.37)	.06
RFT	2.42 (.63)	2.0 (.45)	.52	2.05 (.60)	1.6 (.40)	.13
TAU	1.73 (.53)	1.4 (.36)	.39	2.14 (.67)	1.6 (.32)	-.06
Early phase	2.21 (.72)	1.8 (.54)	.69	2.06 (.68)	1.6 (.30)	.10
Later phase	2.17 (.65)	1.8 (.50)	.63	2.10 (.58)	1.7 (.42)	.01

Table 4

Novice and Veteran self-report and observer mean scores on the ITT-ABP-O factor scales by study condition

	RFT			TAU		
	Novice M (SD)	Veteran M (SD)	Mean difference	Novice M (SD)	Veteran M (SD)	Mean difference
FT scale						
Self-report	2.18 (.63)	2.66 (.54)	$p < .001$	1.78 (.47)	1.64 (.63)	ns
Observer-report	1.92 (.42)	2.15 (.45)	ns	1.42 (.36)	1.37 (.38)	ns
ICC ^a	.55	.33	–	.17	.62	–
MI/CBT scale						
Self-report	1.95 (.44)	2.16 (.71)	$p < .001$	2.24 (.62)	1.95 (.73)	ns
Observer-report	1.76 (.42)	1.53 (.36)	ns	1.65 (.30)	1.57 (.36)	ns
ICC ^a	.00	.19	–	–.34	.26	–

^a ICCs were conducted to compare self-report and observer-report scores for Novice and Veteran subgroups within each condition