HHS Public Access

Author manuscript

J Clin Child Adolesc Psychol. Author manuscript; available in PMC 2024 November 01.

Published in final edited form as:

J Clin Child Adolesc Psychol. 2023 ; 52(6): 850–865. doi:10.1080/15374416.2022.2051529.

Pilot Trial of Online Measurement Training and Feedback in Family Therapy for Adolescent Behavior Problems

Aaron Hogue[1], **Alexandra MacLean**[1], **Molly Bobek**[1], **Nicole Porter**[1], **Lila Bruynesteyn**[1], **Amanda Jensen-Doss**[2], **Craig E. Henderson**[3]

[1]Family and Adolescent Clinical Technology & Science, Partnership to End Addiction, New York, NY, USA.

[2]Department of Psychology, University of Miami, Miami, FL, USA.

[3]Department of Psychology, Sam Houston State University, Huntsville, TX, USA.

Abstract

Objective: Pragmatic procedures for sustaining high-fidelity delivery of evidence-based interventions are needed to support implementation in usual care. This study tested an online therapist training system, featuring observational coder training and self-report fidelity feedback, to promote self-report acumen and routine use of family therapy (FT) techniques for adolescent behavior problems.

Method: Therapists (N = 84) from nine substance use and mental health treatment sites reported on 185 adolescent clients. Therapists submitted baseline data on FT technique use with clients, completed a workshop introducing the 32-week training system, and were randomly assigned by site to Core Training versus Core Training + Consultation. Core Training included a therapist coder training course (didactic instruction and mock session coding exercises in 13 FT techniques) and fidelity feedback procedures depicting therapist-report data on FT use. Consultation convened therapists and supervisors for one-hour monthly sessions with an external FT expert. During the 32 weeks of training, therapists submitted self-report data on FT use along with companion session audiotapes subsequently coded by observational raters.

Results: Therapist self-report reliability and accuracy both increased substantially during training. Observers reported no increase over time in FT use; therapists self-reported a decrease in FT use, likely an artefact of their improved self-report accuracy. Consultation did not enhance therapist self-report acumen or increase FT use.

Conclusions: Online training methods that improve therapist-report reliability and accuracy for FT use may confer important advantages for treatment planning and fidelity monitoring. More intensive and/or different training interventions appear needed to increase routine FT delivery.

Keywords

family therapy; online training; usual care; adolescent behavior problems

Correspondence should be addressed to: Aaron Hogue, Partnership to End Addiction, 711 Third Avenue, Suite 500, New York, NY, 10017. ahogue@toendaddiction.org.

A highly promising strategy for increasing evidence-based interventions (EBIs) in routine behavioral care is instituting research-tested quality assurance procedures to support the broad clinical workforce in EBI delivery. Quality assurance procedures are designed to ensure that efficacious interventions are delivered with fidelity, that is, to the intended population, by appropriately trained providers, and in accord with specified principles and procedures (Hogue et al., 2013). This includes the need for quality metrics that can reliably measure EBI fidelity in everyday settings (McLeod et al., 2013) combined with data-driven learning systems (Chambers et al., 2016) in which EBI implementation activities are carried out systematically, implementation and sustainability data are regularly collected and reviewed, and continuous EBI modifications are made to increase fit and/or feasibility.

Pragmatic EBI Quality Assurance: Measurement Training and Feedback System for Implementation (MTFS-I)

To promote effective implementation of EBIs in usual care, pragmatic procedures for sustaining high-fidelity EBI delivery need to be developed. This study reports a pilot trial of an online quality assurance system intended to promote EBI implementation in community-based care: Measurement Training and Feedback System for Implementation (MTFS-I; Hogue, Dauber, et al., 2019; Hogue, Bobek, MacLean, et al., 2019). The two main MTFS-I components—therapist coder training, EBI self-report feedback—are discussed below. Broadly, the system has three user-centered design features intended to maximize its practicality and effectiveness (Lyon & Koerner, 2016). First, MTFS-I is housed in an online learning management system. Online learning is a cost-effective method that in various formats has proven comparable or superior to in-person workshops for improving clinical knowledge, self-reported use of treatment skills, and clinical proficiency (Beidas & Kendall, 2010). Community clinicians report comfort with online learning, believe it to be efficacious (Becker & Jensen-Doss, 2014), and believe it increases learning accessibility and engagement (Ehrenreich-May et al., 2016). Second, MTFS-I focuses on “practice elements” rather than manualized intervention. EBI practice elements are discrete treatment techniques that are core ingredients of multiple EBI protocols for a given disorder (Chorpita & Daleiden, 2009). Practice elements are considered easier to learn than full manuals and equip clinicians with a diverse portfolio of techniques that can be flexibly used for clients with heterogeneous clinical profiles, making them well-suited for the eclectic treatment strategies found in usual care (Hogue, Bobek, Porter, Dauber, et al., 2021; Weisz et al., 2017). Third, only 15–20 minutes is required to complete each MTFS-I learning exercise via smartphone or computer.

In the current pilot trial, MTFS-I was designed to support delivery of family therapy techniques for adolescent behavior problems. Family therapy (FT) is an evidence-based approach for many behavioral disorders presented by adolescents in routine care: conduct problems and delinquency (Dopp et al., 2017; McCart & Sheidow, 2016), depression (Weersing et al., 2017), substance misuse (Hogue et al., 2018), and eating disorders (Lock, 2015). Further, systematic reviews (Hogue et al., 2018; McCart & Sheidow, 2016) and meta-analyses (Baldwin et al., 2012; Dopp et al., 2017; Tanner-Smith et al., 2013) suggest that, compared to other evidence-based approaches, FT has perhaps the strongest empirical

support for treating adolescent conduct and substance use disorders. These are compelling reasons for intensifying efforts to promote delivery of high-fidelity FT interventions for adolescent clients in community settings.

Measurement Training: Therapist Training in EBI Observational Coding

MTFS-I makes therapist coder training the centerpiece of the learning experience. All learning exercises contain a brief video of a mock therapy session segment depicting a handful of core FT techniques for adolescent behavior problems (Hogue, Bobek, et al., 2019); these scripted vignettes contain actors in client roles and expert therapists in the clinician role. After viewing each vignette, therapists rate the extent to which specific EBI techniques were present. Immediately after submitting ratings, they receive automated feedback that compares their scores to expert-derived scores for each technique, along with evidence justifying the expert scores.

There are several potential benefits to therapist coder training. A logical first step toward boosting the capacity of community therapists to implement EBIs with fidelity is improving their ability to recognize and assess interventions they are expected to deliver (McLeod et al., 2018). Because each video vignette depicts a cohesive set of techniques delivered within a realistic sequence of client and therapist interactions, coder training on these vignettes can improve EBI assessment acumen by leveraging immediate corrective feedback on objectively rated samples of gold-task performance (Gonder et al., 2018). Coder training can also increase therapist declarative knowledge (i.e., factual knowledge and information about an EBI) as well as make initial gains in procedural knowledge (i.e., knowledge about how to deliver an EBI) via cognitive mechanisms of observational learning and evaluative processing (McLeod et al., 2018).

Observational coder training might also boost therapist reliability in reporting on their own use of EBI techniques with their routine caseloads. Studies attempting to show concordance between therapist self-ratings and observer ratings of EBI fidelity have produced mostly disappointing results (e.g., Wain et al., 2015), casting doubt on whether therapists can reliably rate their own performance. This includes research-trained clinicians delivering manualized treatment (e.g., Martino et al., 2009) as well front-line clinicians in routine care (e.g., Hurlburt et al., 2010). In theory, therapist coder training could help turn the tide on self-report acumen. To date only a handful of studies have trained practitioners to function as participant judges, whereby they (a) learn to observationally code EBIs they are expected to use with their own clients and (b) participate in coder training activities during their everyday work routine (e.g., Caron & Dozier, 2019, 2022; Isenhart et al. 2014).

A potential downstream benefit of therapist coder training is prompting increased EBI delivery in real-world practice (Caron & Dozier, 2022; Stirman, 2020). Although live coaching and guided skills practice are the most effective means to teach new clinical skills, video-based modeling has shown promise for augmenting EBI use (e.g., Beidas et al., 2014). Learning to recognize and assess EBI techniques via longitudinal coder training could solidify the clinical acceptability of the techniques and gradually strengthen therapist confidence and motivation to implement them in routine services (McLeod et al., 2018).

Feedback System: Feedback on Therapist Self-Reports of EBI Delivery

The second main component of MTFS-I is a fidelity-focused measurement feedback system. Conventional measurement feedback procedures institute a performance feedback loop in which a given quality metric is continuously monitored by a clinician to gauge case progress and support clinical decision-making. Measurement feedback has led to impressive gains in treatment outcomes across diverse clinical samples (see Shimokawa et al., 2010), success that has generated enthusiasm about the value of developing complementary procedures for EBI implementation. When attuned to implementation characteristics such as EBI fidelity, feedback systems could serve as a functional quality procedure with broad dissemination potential in behavioral health (McLeod et al., 2013). In MTFS-I, therapists are asked to submit online self-reports of EBI technique use with their caseloads; these data are converted into customizable EBI fidelity feedback reports (described in Method). To the degree that training in observational coding boosts the accuracy of therapist self-reports of EBIs, the potential accuracy and hence utility of fidelity feedback reports are concomitantly increased.

Current Study Aims

The current study was a two-phase, cluster randomized effectiveness-implementation pilot trial (fully described in Hogue, Dauber, et al., 2019) conducted at nine clinical sites. During the Baseline Phase (lasting between 8 and 12 weeks at each site), prior to beginning MTFS-I, therapists submitted self-report data on use of core FT techniques with their adolescent cases, along with audio recordings of corresponding sessions. The Implementation Phase launched 32 weeks of MTFS-I, along with continued therapist submission of self-report data and session audios. At the start of the Implementation Phase, sites were randomized to Core Training versus Core Training plus Consultation in order to test the effects of adding monthly consultation with an extramural FT expert to the MTFS-I package. A primary goal of the study was to determine whether ongoing expert consultation on MTFS-I participation and delivery of FT techniques significantly improved MTFS-I effects. Whereas booster training and consultation is invaluable for ensuring the vitality and sustainability of EBI training effects (McLeod et al., 2018), and systems-level consultation is used to reinforce service quality outcomes in behavioral treatment systems (e.g., Knight et al., 2016), to our knowledge this is the first test of the benefits of consultation for promoting the effects of online measurement training and fidelity feedback.

The first study aim examined phase effects by comparing Baseline versus Implementation Phase on therapist self-report acumen (i.e., reliability and accuracy) in documenting FT technique use, along with amount of FT technique delivery (i.e., dose), with their adolescent cases. A precursor study from this pilot trial found that study therapists participating in MTFS-I showed significant gains in aspects of observational-report reliability and accuracy while coding brief video vignettes for FT technique use (Hogue, Porter, et al., 2022). The current study of main trial outcomes tested whether study therapists made comparable gains in reliability and accuracy when reporting on their own use of the same techniques with their caseloads. It was predicted that both therapist-report acumen and FT fidelity (i.e., dose of FT delivery) would be stronger during the Implementation Phase. The second aim examined

experimental effects during the Implementation Phase by comparing Core Training versus Core Training + Consultation on therapist-report acumen and FT fidelity. It was predicted that consultation effects would promote greater gains for the Core Training + Consultation condition.

Method

The study was conducted under approval by the governing Institutional Review Board. Data were collected from March 2019 through March 2020.

Study Participants: Baseline Characteristics and Disposition toward Family Therapy

Study participants were 84 behavior therapists working in nine community-based substance use and mental health treatment clinics in various regions of a large northeastern state; see Table 1. Demographic information is missing for 11% of the sample who did not complete survey questions. Therapists (75% self-identified female, 14% male) averaged 31.7 (SD = 9.3) years of age. Self-identified race/ethnicity was 71% White Non-Hispanic, 8% Hispanic, 4% Black/African-American, 2% Asian, 4% other. A total of 83% had a master's level degree, 6% associate's or bachelor's degree; 87% were full- or part-time staff and 2% trainees. They averaged 3.6 (SD = 5.1) years of post-degree therapy experience, 1.9 (SD = 3.3) years of employment at the study clinic, and the average caseload size was 32.5 (SD = 25.3) clients across individual, group, and family session formats. These demographic and practice data are representative of the larger clinical workforce in the agencies from which the study sample was drawn, based on author consult with participating agency directors.

At Baseline each participant provided data on their disposition toward family therapy using the following scale: 0 = None, 1 = A little, 2 = Moderate, 3 = Considerable, 4 = High. Four indices were rated (see Table 1): *Degree of personal allegiance to the family systems treatment approach*: 4% none, 25% a little; 36% moderate, 20% considerable, 6% high (M[SD]: 3.0[.94]); *Self-rating of family therapy skills*: 0% none, 33% a little; 37% moderate, 16% considerable, 5% high (M[SD]: 2.9[.87]); *Degree of importance for their personal caseload to include family members in sessions and other treatment activities*: 0% none, 6% a little; 25% moderate, 39% considerable, 21% high (M[SD]: 3.8[.84]); *Degree of importance for my personal caseload to learn or review family therapy techniques*: 0% none, 11% a little; 23% moderate, 31% considerable, 26% high (M[SD]: 3.8[.98]).

Study Sites and Clients

Study therapists and their clients were affiliated with nine outpatient behavioral treatment sites: 4 were licensed as substance use treatment clinics, 2 licensed as mental health treatment clinics, and 3 co-licensed to deliver both substance use and mental health services. Clinics were located in suburban (n = 7) and urban (n = 2) locations and had been in operation for an average of 35.5 (SD = 12.0) years. Collectively, clinics reported that their clients generally spend an average of 0.85 (SD = 0.15) hours per week in individual sessions and 1.5 (SD = 0.7) hours per week in group sessions. All clinics expressed desire to increase routine use of FT techniques.

Clients (n = 185) were adolescents referred for outpatient care and their families. Clinical record data collected on study clients indicated that adolescents self-identified as 41% female, 31% male, 28% unreported; they averaged 17.0 (SD = 2.3; range 13–21) years of age. Self-identified race/ethnicity was 45% White Non-Hispanic, 9% Hispanic, 7% Black/African-American, 9% a different category or multiple categories, and 30% unreported. In addition, eight clinics provided site-level trend data regarding the percentage of adolescent-aged clients who presented with the following primary referral problems (averaged across clinics): substance use (42%), mental health/other (58%).

Study Procedures: Eligibility, Baseline Training Workshops, and Site Randomization

All volunteering behavior therapists employed by study sites were eligible to participate in the study if they met the following criteria: routinely treated clients age 13–21 years; agreed to submit Baseline Phase data (at least two audiorecorded sessions and corresponding self-report checklists on use of FT techniques in the given session) prior to participating in Baseline training workshops; and agreed to submit Implementation Phase data (session recordings and self-report checklists) for adolescent-aged clients after workshops concluded for the duration of the study (32 weeks). Virtually every therapist at each study site who met these criteria consented to participate. After submitting Baseline Phase data, therapists attended two on-site 90-minute Baseline workshops. The first introduced the online therapist coder training procedures, and the second introduced 13 core FT techniques that were the foci of ongoing training. After Baseline workshops concluded, sites were randomized to study condition (described in Study Conditions section): Core Training (5 sites) versus Core Training + Consultation (4 sites).

CONSORT Data: Participant and Data Flow

Study site and therapist flow are documented in the CONSORT diagram (Figure 1). Of 21 potential sites approached to collaborate, 12 (57%) hosted an on-site research orientation meeting for clinical staff and 9 (43%) were ultimately randomized as study sites. All randomized sites remained active for the study duration. Of the 84 therapists from randomized sites who consented to participate in the study, 29 (35%) attrited before completing Baseline training workshops because they left clinic employ (n = 17), failed to submit Baseline data (n = 11), or asked to leave the study (n = 1). Of the 55 therapists who completed Baseline workshops and initiated training interventions, 16 (29%) attrited because they did not submit any data during the Baseline and Implementation Phase data from 39 therapists reporting on 164 clients.

Examining the pool of 84 therapists who consented to participate in the study, we compared the study sample (N = 39) to the attrited sample (n = 45) on baseline therapist characteristics; see Table 1. Therapists who attrited did not differ from the study sample on any baseline characteristic. We also compared the study versus attrited samples on disposition toward family therapy and there were no significant differences. Also, there was no difference between study conditions in the percentage of therapists who attrited after consenting to participate: 49% in Core Training (CT), 55% in Core Training + Consultation (CTC).

Examining the study sample of N = 39 therapists, we compared CT (21 therapists at five sites) to CTC (18 therapists at four sites) on baseline therapist characteristics and disposition toward family therapy; see Table 1. There were no significant between-condition differences. We also compared study conditions on data submitted. During the Baseline Phase, CT submitted 112 checklists (per therapist: M[SD] = 5.1[3.6]) and 60 audios (per therapist: 2.9[1.5]); CTC submitted 150 checklists (9.4[9.9]) and 83 audios (5.2[3.7]); there was no significant difference in number of checklists submitted by study condition, however, CTC therapists submitted significantly more audios [$t(37) = -2.65$, $p = .01$]. During the Implementation Phase, CT therapists submitted 202 checklists (M[SD] = 10.1[8.6]) and 151 audios (7.9[8.1]); CTC submitted 230 checklists (13.5[9.9]) and 154 audios (10.3[9.5]); there were no significant between-condition differences.

Study Conditions: Training Interventions and Intervention Uptake

Core Training (CT).—The 32-week MTFS-I training system provided instruction in three clinical modules containing 13 core FT techniques for treating adolescent substance use and conduct problems. FT techniques were drawn from an empirical distillation process driven by analysis of observational fidelity coding data from manualized FT models (Hogue, Bobek, et al., 2019); items are listed in Table 2. There were two CT components. The first was a *therapist coder training course* hosted on a web-based learning management system. A total of 32 training exercises were released weekly via protected weblinks distributed by email. Exercises remained accessible until completed by a given trainee. Each exercise contained two synergistic parts. (1) *Didactic Instruction*: slides containing brief descriptions and exemplar therapist statements for selected techniques. Each exercise presented three total techniques (one per slide) during didactic instruction. In order to promote skill in differentiated coding, only two of these techniques appeared in the vignette that followed. (2) *Mock Session Coding*: one 5–8 minute scripted video vignette modeling multiple techniques, followed by a standardized coding activity. Vignettes depicted expert family therapists working with actors in family roles, all re-enacting therapy scenarios drawn from real cases. Each vignette illustrated several FT techniques ranging from low to high extensiveness, and collectively the vignettes depicted a diverse group of therapists and families and showcased a range of therapist styles and presenting problems. After viewing the vignette, trainees were guided through a coding activity designed to grow their ability to recognize and evaluate FT technique delivery. Trainees were instructed to rate selected techniques on a 5-point Likert-type scale according to the thoroughness and frequency with which each appeared in the vignette: 0 (*Not at all*), 1 (*A little bit*), 2 (*Moderately*), 3 (*Quite a bit*), 4 (*Extensively*). Trainees rated five selected techniques per coding exercise; to sharpen technique recognition and discrimination, only three of the selected techniques actually appeared in the vignette. Upon completing the coding activity, trainees received immediate scoring feedback in the form of gold scores determined via consensus scoring by FT coding experts (including authors AM, MB, NP, LB) who themselves observationally rated each vignette during MTFS-I construction. Gold scores for the three depicted techniques were accompanied by verbatim statements (uttered by the therapist in the vignette) exemplifying each technique; these statements served as both scoring evidence and teaching exemplars of the given technique.

The second core training component was a *fidelity measurement feedback system* focused on therapist self-report data on delivery of FT techniques. Therapists were asked to complete self-report checklists on use of the 13 core FT techniques after every session (i.e., inclusive of youth only, family member only, and conjoint sessions) with their adolescent caseload; they rated each technique on a 5-point scale identical to that described above for the therapist coder training course (see Measures for a full description of the checklist: ITT-CEFT). Research staff compiled checklist data into monthly fidelity feedback reports summarizing cumulative data for each therapist, including: (a) family member participation in sessions; (b) mean values of each FT technique item and the FT total scale (average of all 13 items), aggregated at the client level; (c) mean values of the FT total scale. Fidelity feedback reports were emailed directly to therapists. In addition, clinical supervisors at each site were emailed monthly feedback reports containing non-identified checklist data on each FT technique item and the total scale, aggregated at the site level. Two different sample versions of MTFS-I feedback reports can be found in Hogue, Dauber and colleagues (2019) and Hogue, Bobek, MacLean, and colleagues (2019), respectively.

Core Training + Consultation (CTC).—In addition to both components of Core Training, therapists in CTC received clinical consultation from an external FT expert (M. Bobek) via teleconference during the 32 weeks of core training. Separately at each site, consultation convened therapists and their supervisors for one-hour monthly sessions. At each group's discretion, consultation sessions could focus on one or more of five topics: technical assistance and support of therapist participation in CT components; review of FT techniques and their applicability in routine care; review and processing of the previous month's online coding vignettes; review and processing of site-level or therapist-level fidelity feedback reports; review and feedback on session audio recordings; discussion of a therapist-prepared case summary.

Measures: Therapist- and Observer-Report of FT Fidelity

Inventory of Therapy Techniques—Core Elements of Family Therapy (ITT-CEFT).—The ITT-CEFT is a therapist-report fidelity checklist designed to collect post-session data on delivery of core FT treatment techniques for adolescent substance use and conduct problems. It consists of three clinical modules containing 13 techniques; all are listed in Table 2. The ITT-CEFT operationalizes FT fidelity in the form of extensiveness (i.e., quantity, or dose) scores (Hogue et al., 1996). Therapists report the extent—defined as thoroughness and/or frequency—to which each technique was utilized in a just-completed session, based on a 5-point Likert-type scale: 0 = *Not at all*, 1 = *A little bit*, 2 = *Moderately*, 3 = *Quite a bit*, 4 = *Extensively*.

The original pool of ITT-CEFT items was derived from prior work (Hogue, Bobek, et al., 2019) that analyzed observational ratings of 302 therapy sessions to identify model-shared treatment techniques from the respective fidelity scales of manualized, empirically supported FT models for adolescent behavior problems. The final pool of 13 ITT-CEFT items was examined in a previous study associated with the current pilot trial (Hogue, Bobek, Porter, MacLean, et al., 2022). Confirmatory factor analysis from 189 sessions justified the three-factor structure; fit indices were: $\chi^2(51) = 99.17$, $p < 0.001$; RMSEA = .07 (90% CI: .05 -

.09); CFI = .90. Internal consistency for each derived module was adequate as indicated by inter-item correlations within module: Cronbach's α range .66 - .74. There was meaningful differentiation among modules as indicated by the pattern of bivariate correlations between modules, wherein each correlation was $r < .70$ (Kline, 1979). Importantly for the current study, these previous analyses also indicated that study therapists were adequately reliable in reporting on their own use of FT techniques averaged at the module level. Therapist ratings of their own sessions on the ITT-CEFT were compared to observer ratings of audio recordings of the same 189 sessions from the Implementation Phase of the trial. One-way random intraclass correlation coefficients (ICCs; Shrout & Fleiss, 1979) were calculated for the mean scores on each FT module; therapist-observer ICCs ranged from .64 to .75, indicating adequate reliability (Koo & Li, 2016). Overall, these psychometric properties indicate that ITT-CEFT data can analyzed as sufficiently reliable training outcome data for the current study.

Inventory of Therapy Techniques—Core Elements of Family Therapy: Observational Version (ITT-CEFT-O).—The ITT-CEFT-O contains 13 core FT technique items identical to those on the ITT-CEFT. It also contains observational scoring guidelines.

Study Procedures: Data Collection and Observational Coding

Throughout the study, therapists were encouraged to submit self-report checklist data and companion audio recordings after sessions for as many adolescent clients and sessions as possible, regardless of session composition (i.e., which persons participated in the given session). Therapists submitted self-report data by completing an online survey powered by Qualtrics with fields for recording session composition information and item scores; session audio recordings were submitted via a secure online upload to protected research archives.

All submitted audio recordings were not included in the current study; coding resources permitted us to code a maximum of one Baseline Phase audio and four Implementation Phase audios per client. The Baseline Phase audio was randomly selected. For Implementation Phase data, for any client with five or more audios submitted, four were randomly selected for study inclusion; when this occurred, we prioritized selecting from among those sessions for which companion checklist data indicated that a family member attended. Of the 68 clients for whom therapists submitted audios in the Implementation Phase, 24 (35%) provided five or more sessions, 4 (6%) provided four, 7 (10%) three, 10 (15%) two, and 23 (34%) one.

Observational coders (n = 14) were research personnel consisting of undergraduates and graduates with a bachelor's degree (n = 9) and graduates with master's level training in social work, psychology, or a related field (n = 5). Coders were trained during weekly virtual meetings over the course of two months using review of the ITT-CEFT-O coding manual, in-group coding and review of practice recordings, and exercises to increase understanding of scale items. Study coding commenced once all coders reached a collective threshold reliability of ICC = .65 for the preponderance of items; thereafter, the group met biweekly for supportive training. Sessions were scored in their entirety (average about 55 minutes).

Two coders were assigned to score each session; coders were randomly paired with each other across the session sample using a randomized block design (Fleiss, 1981).

Plan of Analysis

The study design incorporated therapy sessions nested within clients who were nested within therapists, who in turn were nested within sites. We accounted for non-independence of data by using multilevel, mixed effect modeling to test study hypotheses, with random effects for Client, Therapist, and Site. Analyses used full information maximum likelihood estimation (FIML) to accommodate and reduce potential bias due to missing data under the assumption that data were missing at random (Schafer & Graham, 2002). All analyses were conducted using the statistical platform R (R Studio Team, 2020); specific R packages used to conduct analyses and generate coefficients are listed below.

The first set of analyses compared Baseline Phase versus Implementation Phase. Two classes of Phase effects were examined: therapist self-report acumen (i.e., concordance with observers), and dose of FT interventions delivered. Therapist self-report acumen was operationalized in terms of both *reliability* (consistency with and approximation to gold scores across rating occasions) and *accuracy* (agreement with gold scores; see LeBreton & Senter, 2008). Therapist reliability was calculated using the one-way random intraclass correlation coefficient with average rater scores ($ICC_{(1,2)}$; Shrout & Fleiss, 1979) via the irr package in R (Gamer et al., 2019). Therapist accuracy was assessed using an estimate of inter-informant agreement, r_{wg} (James et al., 1984), via the multilevel package in R (Bliese, 2016). FT intervention dose effects were operationalized using composite scores (averaging across items) for FT Total Score and also for its three subscales: Family Engagement, Relational Orientation, Interactional Change. For all FT dose analyses, therapist and observer ratings were analyzed separately. We tested between-Phase effects for therapist accuracy and FT dose using mixed-effect ANOVA models via the lmer package (Bates et al., 2015) along with the lmerTest package to estimate p values (Kuznetsova et al., 2017). Effect sizes were indexed by Eta squared (η^2) and generated via the effectsize package (Ben-Shachar, 2021). Note that there were not sufficient numbers of data points per therapist to generate stable ICC profiles at the Phase level to allow between-Phase comparisons of therapist reliability.

The second set of analyses examined experimental effects during the Implementation Phase by comparing CT versus CTC. First, we used the procedures described above to evaluate therapist reliability in each experimental condition using $ICC_{(1,2)}$. We then examined comparative growth in therapist accuracy and FT dose over the course of the Implementation Phase using a multilevel modeling framework. Again, for reliability data there were not sufficient numbers of data points to model growth on a weekly basis. Multilevel models were tested via the lme4 and lmerTest packages, in two stages. First, we examined unconditional multilevel models to test for overall sample linear growth in therapist accuracy using r_{wg} coefficients and FT dose variables. Then, random effects were added to account for data nesting, and experimental condition (CT or CTC) was added as a fixed effect to examine comparative growth in accuracy and FT dose. Cohen's $\hat{f}^2$ was used to index effect sizes generated using the effectsize package.

Results

Preliminary Analyses of Intervention Uptake

Study activities at all sites were halted in March 2020 due to the COVID-19 pandemic. Because sites initiated study activities on a staggered schedule, two of the nine sites (5 of 39 therapists) had access to only 31 of the 32 training exercises. In CT, therapists completed an average of 21.8 (SD = 10.0) exercises; 85% of therapists completed at least 8 exercises, 75% 16 exercises, and 60% 24 exercises. In CTC, therapists completed an average of 19.3 (SD = 8.8) exercises; 89% of therapists completed at least 8 exercises, 66% 16 exercises, and 43% 24 exercises. There was no between-condition difference in exercises completed.

Therapists provided data on perceived clinical utility of the CT components at conclusion of training exercises, every four weeks starting week 8. They responded to two queries: *How relevant and/or useful were the illustrated techniques in video vignettes to your clinical practice?*; and *How relevant and/or useful was the fidelity feedback report to your clinical practice*? For each query they completed a 5-point rating scale: 0 (*Not at all*), 1 (*A little bit*), 2 (*Moderately*), 3 (*Quite a bit*), 4 (*Extensively*). In CT, the average utility score for illustrated techniques was 3.3 (SD = 1.0); average utility score for feedback reports was 2.3 (.98). In CTC, the average utility score for illustrated techniques was 3.2 (.76); average utility score for feedback reports was 2.1 (.93). There was no significant between-condition difference for either utility score.

Regarding uptake of the monthly Consultation sessions in CTC, sites convened an average of 5.3 (SD = 0.81) sessions; each CTC therapist attended an average of 3.9 (1.5) sessions. The expert consultant completed a checklist after every session documenting the number of session minutes spent on each of the five consultation topics. Across all sites, average time in each Consultation session was allotted as follows: 18 (SD = 9.5) minutes on CT technical assistance; 11 (16.1) minutes on review of coding vignettes; 6 (9.5) minutes on review of feedback reports; 8 (14.8) minutes on review of session audios; and 11 (10.0) minutes on clinical case discussion.

Phase Effects: Therapist Self-Report Acumen (Reliability, Accuracy) and FT Dose

Therapist self-report reliability data comparing Baseline versus Implementation are shown in Table 3. ICC magnitudes can be interpreted based on Cicchetti's (1994) criteria, which are ubiquitous in observational coding research on behavioral interventions: below .40 is poor, .40–.59 is fair, .60–.74 is good, and .75–1.0 is excellent; and/or Koo and Li's (2016) criteria recommended for behavioral measurement theory more broadly: below .50 is poor, .50–.74 is fair, .75–.90 is good, and .91–1.0 is excellent. ICC for FT Total Score was .35 (poor) during Baseline, with subscales ranging from .25 (poor) to .43 (fair/poor). During Implementation, ICC for EF Total Score was .72 (good/fair), with subscales ranging from .60 (good/fair) to .72 (good/fair). Therapist self-report accuracy results tested with mixed effects ANOVA models are shown in Table 4. We adopted Cohen's (1988) guidelines for classifying effect size magnitude (η^2): 0.01 = small, 0.06 = medium, and 0.14 = large. For therapist accuracy, results indicated stronger therapist-observer agreement in Implementation than Baseline for all variables: FT Total Score [$F(1, 101.78) = 18.84$, $p < .001$, $\eta^2 =$

0.16], Family Engagement [F(1, 100.22) = 5.76, p = .02, η^2 = 0.05], Relational Orientation [F(1, 111.93) = 5.88, p = .02, η^2 = 0.05], and Interactional Change [F(1, 103.26) = 11.74, p <.001, η^2 = 0.10]. These data demonstrate that therapist training activities during the Implementation Phase precipitated a two-fold increase in therapist reliability, along with a significant increase in therapist accuracy, for every FT intervention module.

As a companion to the mean scores listed in Table 2, in Table 4 we list statistical results comparing groups for FT intervention dose. Results differed by reporter. According to therapist-report data, there was a higher average FT dose during Baseline than Implementation for all four variables: FT Total Score [F(1, 243.50) = 37.35, p < .001, η^2 = 0.13], Family Engagement [F(1, 244.10) = 9.27, p < .01, η^2 = 0.04], Relational Orientation [F(1, 251.05) = 32.13, p < .001, η^2 = 0.11], and Interactional Change [F(1, 244.18) = 26.09, p <.001, η^2 = 0.10]. In contrast, according to observer reports, there was no significant between-Phase effect in FT dose for any variable.

Experimental Effects: Core Training versus Core Training + Consultation

Therapist reliability data during the Implementation Phase comparing CT to CTC are also shown in Table 3. For the CT condition, ICC for FT Total Score was .64 (good/fair); subscales ranged from .43 (fair/poor) to .60 (fair/poor). For the CTC condition, ICC for FT Total Score was .67 (good/fair); subscales ranged from .56 (fair) to .65 (good/fair).

Results of the multilevel models testing overall and comparative growth (CT versus CTC) in therapist accuracy and FT dose are shown in Table 5. We used the following guidelines for indexing effect size magnitude (f^2): 0.10 is small, 0.25 is medium, 0.40 is large (Cohen, 1988). Results of unconditional models testing for change in the average slope of the accuracy coefficient (r_{wg}) showed significant linear growth for Interactional Change (B = 0.01, SE = 0.01, p = 0.03, β = <0.001, f^2 = 0.03) indicating therapists became more accurate over time rating items related to coaching family interactions in session. Unconditional models testing therapist accuracy did not show change for FT Total Score, Family Engagement, or Relational Orientation. Unconditional models testing for increase in FT dose did not show change for either therapist- or observer-report data. Conditional models testing for comparative increases (CT versus CTC) did not show between-condition differences in either therapist accuracy or FT dose.

Discussion

Study results suggest that online measurement training and feedback in family therapy for adolescent behavior problems produced substantial improvements in community therapist acumen in reporting on their use of core FT interventions with their routine cases. Therapist self-report reliability transitioned from uniformly poor during pre-training to uniformly fair-to-good across all FT intervention modules, showing a two-fold increase overall. Similarly, therapist self-report accuracy improved significantly across all FT interventions, with small-to-medium effect sizes for each FT module and a large effect size for the FT total scale. However, contrary to hypotheses, observational raters reported no change in the actual dose of FT interventions delivered in sessions after training started, and even more surprisingly, therapists reported a decrease in their own use of FT. Also contrary to hypotheses, there

were no differences between experimental conditions on any self-report acumen or FT dose variable, indicating that this study's expert consultation procedures did not add value to MTFS-I components.

Demonstrating that therapist coder training and fidelity feedback can systematically increase therapist self-report reliability and accuracy is a breakthrough finding. Past research has shown that practitioners are not naturally proficient self-raters of their EBI use in session. The reliability of community clinicians reporting on their own delivery of EBIs is generally poor (e.g., Brookman-Frazee et al., 2021), with a few exceptions (e.g., Hogue et al., 2015). And self-report accuracy is weak virtually without exception, as clinicians tend to over-report both the number (breadth) and extensiveness (depth) of EBIs they have delivered (e.g., Brookman-Frazee et al., 2021; Hurlburt et al., 2010). Current study results, among a handful of others (e.g., Caron & Dozier, 2019, 2022), indicate that online training in observational coding can help turn the tide, boosting therapists to achieve adequate-or-better levels of reliability and reducing the proclivity to inflate self-report of EBI use.

The second main finding was disappointing to be sure: There was no evidence that MTFS-I made any impact on the level of FT interventions delivered in routine care by study therapists. Based on observer ratings of recorded sessions, therapists delivered a meager dose of FT both before and after training commenced, with extensiveness ratings for the FT total score falling on average between scale anchor values of 0 (*Not at all*) and 1 (*A little bit*). With ample room to move the needle for every FT module, no progress occurred for any. Our interpretation of the seemingly counterintuitive finding for therapist ratings—that FT interventions actually waned after training commenced—is that the apparent decrease in dose is a mirage generated by the established increase in self-report accuracy. That is, as therapists strengthened their acumen in judging their own delivery of FT, their tendency to inflate their scores was correspondingly mitigated. This interpretation is supported by the fact that whereas therapist scores for FT dose during the Baseline Phase were higher than observer scores across the board, during the Implementation Phase all therapist scores decreased to equivalency with observer scores. In one sense, this correction of self-report dose inflation could be couched as a coder training victory. Still, if future studies continue to find that "lighter touch" training systems such as MTFS-I cannot move the needle on FT delivery, behavioral health agencies determined to increase FT utilization may need to default to purveyor-driven manualized FT models that are expensive and resource-demanding (Hogue et al., 2013) yet produce demonstrable training and outcome effects (e.g., Baldwin et al., 2012).

The third main finding, also contrary to hypotheses, was that expert FT consultation had no impact on either therapist self-report acumen or delivery of FT interventions. The Core Training and Core Training + Consultation conditions logged equivalent gains in reliability and accuracy, and each had no discernible gain in FT dose. Expert supervision and booster training have well-known benefits for promoting EBI use (McLeod et al., 2018; Valenstein-Mah et al., 2020), benefits that did not emerge in the current study. We suspect that study consultation methods were not sufficiently extensive. To minimize agency burden, the CTC condition prescribed monthly one-hour consultation meetings, well below the intensity featured in most EBI training programs (see Beidas & Kendall, 2010;

Valenstein-Mah et al., 2020). Training uptake data corroborate this interpretation: Therapists attended an average of 3.9 (SD = 1.5) consultation sessions over the course of training, which is quite modest exposure to expert clinical guidance. Also, consultant log data show that about one-third of consultation time was devoted to technical assistance in using the online training system itself. Future efforts in online EBI training should look to maximize (within agency tolerance levels) the amount of expert consultation provided as well as divorce technical support from clinical consultation activities. We also suspect that training efforts will be more successful to the degree that agencies set firm professional development expectations that clinicians demonstrate increased EBI delivery.

Current results extend findings from precursor analyses of the same sample. Hogue, Porter, and colleagues (2022) examined therapist performance in scoring the mock video vignettes presented during coder training sessions. They found that therapist reliability in observationally scoring FT dose within the MTFS-I system improved over time, and that therapist tendency to give low-accurate scores (i.e., highly discrepant from gold standard scores) declined. Thus, progress in coding acumen while rating mock video interventions (precursor study) ran parallel to progress in self-reporting acumen while rating their own interventions (current study). Also, the precursor analyses found that accuracy in rating mock videos improved significantly only for the Interactional Change module; in the current study, accuracy in rating their own completed sessions likewise improved only for Interactional Change. Many contend that FT techniques focused on arranging, coaching, and processing interactions among family members in session are the *sine qua non* foundations of the FT approach (Minuchin & Fishman, 1981). These parallel findings across both precursor and current analyses of our pilot trial data constitute a measure of preliminary conceptual validity for the coder training component of MTFS-I, though experimental evidence is needed to verify and elucidate the actual mechanisms of observed training effects (McLeod et al., 2018).

To this very point, the study design did not accommodate dismantling of MTFS-I effects attributable to the coder training versus fidelity feedback components. Still, gains in therapist reliability and accuracy appear to have been fueled by participation in the vignette coding exercises. In both conditions, therapists completed an average of about 20 training sessions, with at least two-thirds completing 16 sessions, indicating substantial sample uptake of the coder training component. Also, in both conditions, average therapist-reported clinical utility for the coder training exercises was a full scale point higher than that for the fidelity feedback reports. Future research should focus on discovering whether and how fidelity measurement feedback procedures can be properly leveraged to optimize EBI training and implementation (Hogue et al., 2013).

Study Strengths and Limitations

There were several study strengths. The study sampled community therapists operating in everyday clinical settings and reporting on routine caseloads. These are conditions of high ecological validity that support generalizability of findings to real-world practice. Therapist and client participants were relatively diverse demographically and representative of the regions from which the sample was drawn. Results appear generalizable to the

originally recruited sample: There were no differences in baseline therapist characteristics between the retained sample included in the study versus the attrited sample. Still, the 57% attrition rate (from 84 recruited to 39 retained) attests to the need for strategies to increase workforce participation in EBI training (Jensen-Doss et al., 2020); one intriguing option is mounting EBI training studies in organizations wherein the training itself is mandatory for clinical staff. There was virtually no evidence of randomization failure, as the CT and CTC conditions were equivalent on all indices of therapist characteristics, disposition toward the FT approach, and data submission save one (CTC therapists submitted significantly more audio recordings). Data on sample disposition toward the FT approach signaled mostly moderate-to-considerable personal allegiance to, self-rated skill level in, and caseload-based importance of FT; thus, the sample appeared reasonably primed to learn and use core FT techniques.

There were also numerous study limitations. There was a relatively small number of participating sites and therapists, which did not supply a nationally representative sample of the usual care workforce. Collection of recorded sessions was decidedly non-random: Study therapists were asked to record and upload as many sessions as possible, but only a small fraction of convened sessions was ultimately submitted, driven by whatever selection biases held sway for a given therapist. These sampling gaps open the door to sampling biases of several kinds (e.g., overrepresentation of therapist-preferred clients/sessions and/or clients with less-flexible treatment plans, underrepresentation of clients with erratic attendance or who refused to be recorded) that encroach on study generalizability. Analyses controlled for therapist nesting effects but did not investigate individual therapist differences in self-report acumen or FT delivery. With a focus solely on core techniques, MTFS-I does not capture the “contours” of EBI delivery (Schoenwald et al., 2011) defined by the parameters of a given treatment (i.e., service delivery aspects of implementation: to whom, where, how often) and by its prescribed treatment themes and session content (Garland et al., 2010). Asking therapists to self-rate the more easily defined targets and foci of their interventions, rather than discrete treatment techniques that are often multifaceted and interwoven, sets the measurement bar a notch lower, which might engender even better reliability and accuracy (see Hogue et al., 2014). Another limitation was MTFS-I focus on the extensiveness (i.e., dose) rather than the expertise (i.e., competence) with which therapists delivered FT techniques; therapist expertise in implementing EBI techniques is highly germane to quality practice but notoriously difficult to judge reliably even by observers, let alone therapists themselves (Webb et al., 2010). Finally, the study did not examine whether therapist experience level influenced study findings. Some may deem the sample’s average of 3.6 (SD = 5.1) years of post-degree experience to be relatively junior; notably, our previous work (Hogue et al., 2015) found that therapist experience did not consistently predict either EBI self-report accuracy or EBI utilization in usual care.

Clinical Implications

The launch point for disseminating online clinician training is creating systems that are engaging, user-friendly, and pragmatic. MTFS-I is designed to minimize staff time commitment and maximize component flexibility. Its feedback report templates can be tailored to suit the needs of clinicians, supervisors, administrators, and/or regulatory

agencies; supervisors have appreciable latitude for how to incorporate report data into supervision meetings and electronic health records; and the learning management platform itself is built for dynamic adaptation over time as procedures become routinized within a given agency. MTFS-I can be readily adapted for approaches other than FT; the authors are currently testing a version focused on core CBT techniques for adolescent conduct and substance use problems (Hogue, Bobek, MacLean, et al., 2019).

Regarding therapist perceptions of MTFS-I acceptability, our earlier study of MTFS-I training effects (Hogue, Porter, et al., 2022) found that trainee ratings of the utility of the observational coder training exercises were not related to improvements over time in their reliability and accuracy in coding video vignettes. Of note, during informal group exit interviews conducted with each site at the conclusion of the current study, therapists' consensus suggestions about potential MTFS-I improvements mirrored those that typically emerge from agency-hosted "innovation tournaments" aimed at identifying effective EBI training practices for routine settings (e.g., Sibley et al., 2022; Stewart et al., 2019): emphasize client engagement, train agency supervisors in best-practice supervision, increase use of group feedback and roleplay in supervision, and use fidelity feedback that centralizes session recordings.

Beyond system accessibility and flexibility, there may be sizable clinical advantages conferred by developing training methods that improve self-report reliability and accuracy for EBIs in usual care. For example, Caron and Dozier (2022) found that clinician self-coding accuracy predicted initial fidelity and growth in fidelity to a parent training intervention. There may also potential impacts on data-informed quality procedures such as adaptive treatment planning, person-centered supervision processes, and fidelity measurement and monitoring (Barth et al., 2014; Chambers et al., 2016; McLeod et al., 2013).

Yet, much remains to be accomplished to realize the ultimate goal of designing pragmatic training systems that effectively increase EBI delivery. This goal was not achieved in the current pilot test of MTFS-I. Two training system enhancements are logical next steps toward making tangible inroads. First, as suggested above, more extensive consultation with EBI experts might better scaffold clinicians to transition from recognizing and evaluating EBI techniques (i.e., coder training effects) to deploying those techniques in line services (i.e., EBI dose effects). Second, there are cutting-edge online procedures designed to mimic guided-practice-and-feedback processes that are the bulwark of behavioral intervention trainings. These methods include scripted interactive role-plays, rehearsal and feedback sessions with simulated clients and/or therapists (real or animated), and remote competency feedback on live-capture practice assignments with active cases (e.g., Kobak et al., 2017; Mastroleo et al., 2020). Such methods are worthy, and perhaps necessary, rehearsal-based complements to cognitive-based exercises such as coder training (Beidas et al.; 2014; McLeod et al., 2018).

Acknowledgments

Preparation of this article was supported by the National Institute on Drug Abuse (R34DA044740; PI: Hogue). The trial registration number (ClinicalTrials.gov) is NCT03342872.

The authors gratefully acknowledge the contributions of our colleagues Sarah Dauber, Guy Diamond, Cori Hammond, Suzanne Levy, Bryce McLeod, and Michael Southam-Gerow.

References

Baldwin SA, Christian S, Berkeljon A, & Shadish WR (2012). The effects of family therapies for adolescent delinquency and substance abuse: A meta-analysis. Journal of Marital and Family Therapy, 38, 281–304. [PubMed: 22283391]

Barth RP, Kolivoski KM, Lindsey MA, Lee BR, & Collins KS (2014). Translating the common elements approach: Social work's experiences in education, practice, and research. Journal of Clinical Child & Adolescent Psychology, 43, 301–311. [PubMed: 24245958]

Bates D, Mächler M, Bolker B, & Walker S (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67, 1–48.

Ben-Shachar MS (2021). Indices of effect size and standardized parameters. R package version 0.4.5. Retrieved: https://cran.rproject.org/web/packages/effectsize/effectsize.pdf

Becker EM, & Jensen-Doss A (2014). Therapist attitudes towards computer-based trainings. Administration and Policy in Mental Health and Mental Health Services Research, 41, 845–854. [PubMed: 24150441]

Beidas RS, Cross W, & Dorsey S (2014). Show me, don't tell me: Behavioral rehearsal as a training and analogue fidelity tool. Cognitive and Behavioral Practice, 21, 1–11. [PubMed: 25382963]

Beidas RS, & Kendall PC (2010). Training therapists in evidence-based practice: A critical review of studies from a systems-contextual perspective. Clinical Psychology: Science and Practice, 17, 1–30. [PubMed: 20877441]

Bleise P (2016). Multilevel functions. R package version 2.6. Retrieved: https://cran.r-project.org/web/packages/multilevel/multilevel.pdf

Brookman-Frazee L, Stadnick NA, Lind T, Roesch S, Terrones L, Barnett ML, … & Lau AS (2021). Therapist-observer concordance in ratings of EBP strategy delivery: Challenges and targeted directions in pursuing pragmatic measurement in children's mental health services. Administration and Policy in Mental Health and Mental Health Services Research, 48(1), 155–170. [PubMed: 32507982]

Caron EB, & Dozier M (2019). Effects of fidelity-focused consultation on clinicians' implementation: An exploratory multiple baseline design. Administration and Policy in Mental Health and Mental Health Services Research, 46(4), 445–457. [PubMed: 30783903]

Caron EB, & Dozier M (2022). Self-coding of fidelity as a potential active ingredient of consultation to improve clinicians' fidelity. Administration and Policy in Mental Health and Mental Health Services Research, 49, 237–254. DOI: 10.1007/s10488-021-01160-4 [PubMed: 34499299]

Chambers DA, Feero WG, & Khoury MJ (2016). Convergence of implementation science, precision medicine, and the learning health care system: a new model for biomedical research. JAMA, 315, 1941–1942. [PubMed: 27163980]

Chorpita BF, & Daleiden EL (2009). Mapping evidence-based treatments for children and adolescents: Application of the distillation and matching model to 615 treatments from 322 randomized trials. Journal of Consulting and Clinical Psychology, 77, 566–579. [PubMed: 19485596]

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

Dopp AR, Borduin CM, White I, Mark H, & Kuppens S (2017). Family-based treatments for serious juvenile offenders: A multilevel meta-analysis. Journal of Consulting and Clinical Psychology, 85, 335. [PubMed: 28333535]

Ehrenreich-May J, Dimeff LA, Woodcock EA, Queen AH, Kelly T, Contreras IS, … & Kennedy SM (2016). Enhancing online training in an evidence-based treatment for adolescent panic disorder: a randomized controlled trial. Evidence-Based Practice in Child and Adolescent Mental Health, 1, 241–258.

Fleiss JL (1981). Balanced incomplete block designs for inter-rater reliability studies. Applied Psychological Measurement, 5, 105–112.

Gamer M, Lemon J, Fellows I, & Singh Puspendra (2012). Various coefficients of interrater reliability and agreement. R package version 0.84.1. Retrieved from https://cran.r-project.org/web/packages/irr/irr.pdf

Garland AF, Hurlburt MS, Brookman-Frazee L, Taylor RM, & Accurso EC (2010). Methodological challenges of characterizing usual care psychotherapeutic practice. Administration and Policy in Mental Health and Mental Health Services Research, 37, 208–220. [PubMed: 19757021]

Gonder J, Metlay W, & Shapiro T (2018). Testing assumptions: Can performance rating feedback result in objective performance improvements? Journal of Management and Innovation, 4(2).

Hogue A, Bobek M, Dauber S, Henderson CE, McLeod BD, & Southam-Gerow MA (2019). Core elements of family therapy for adolescent behavior problems: Empirical distillation of three manualized treatments. Journal of Clinical Child and Adolescent Psychology, 48, 29–41. [PubMed: 30657722]

Hogue A, Bobek M, MacLean A, Porter N, Jensen-Doss A, & Henderson CE (2019). Measurement training and feedback system for implementation of evidence-based treatment for adolescent externalizing problems: Protocol for a randomized trial of pragmatic clinician training. Trials, 20, 1–12. DOI: 10.1186/s13063-019-3783-8. [PubMed: 30606236]

Hogue A, Bobek M, Porter N, Dauber S, Southam-Gerow MA, McLeod BD, & Henderson CE (2021). Core elements of family therapy for adolescent behavioral health problems: Validity generalization in community settings. Journal of Clinical Child and Adolescent Psychology, 1–13. DOI: 10.1080/15374416.2021.1969939.

Hogue A, Bobek M, Porter N, MacLean A, Bruynesteyn L, Jensen-Doss A, & Henderson CE (2022). Therapist self-report of fidelity to core elements of family therapy for adolescent behavior problems: Psychometrics of a pragmatic quality indicator tool. Administration and Policy in Mental Health and Mental Health Services Research, 49, 298–311. DOI: 10.1007/s10488-021-01164-0 [PubMed: 34476623]

Hogue A, Dauber S, Bobek M, Jensen-Doss A, & Henderson CE (2019). Measurement training and feedback system for implementation of family-based services for adolescent substance use: Protocol for a cluster randomized trial of two implementation strategies. Implementation Science, 14(1), 1–12. [PubMed: 30611302]

Hogue A, Dauber S, Henderson CE, & Liddle HA (2014). Reliability of therapist self-report on treatment targets and focus in family-based intervention. Administration and Policy in Mental Health and Mental Health Services Research, 41, 697–705. [PubMed: 24068479]

Hogue A, Dauber S, Lichvar E, Bobek M, & Henderson CE (2015). Validity of therapist self-report ratings of fidelity to evidence-based practices for adolescent behavior problems: Correspondence between therapists and observers. Administration and Policy in Mental Health and Mental Health Services Research, 42, 229–243. [PubMed: 24711046]

Hogue A, Henderson CE, Becker SJ, & Knight DK (2018). Evidence Base on Outpatient behavioral treatments for adolescent substance use, 2014–2017: Outcomes, treatment delivery, and promising horizons. Journal of Clinical Child and Adolescent Psychology, 47, 499–526. [PubMed: 29893607]

Hogue A, Liddle HA, & Rowe C (1996). Treatment adherence process research in family therapy: A rationale and some practical guidelines. Psychotherapy: Theory, Research, Practice, & Training, 33, 332–345.

Hogue A, Ozechowski TJ, Robbins MS, & Waldron HB (2013). Making fidelity an intramural game: Localizing quality assurance procedures to promote sustainability of evidence-based practices in usual care. Clinical Psychology: Science and Practice, 20, 60–77.

Hogue A, Porter N, Bobek M, MacLean A, Bruynesteyn L, Jensen-Doss A, Dauber S, & Henderson CE (2022). Online training of community therapists in observational coding of family therapy techniques: Reliability and accuracy. Administration and Policy in Mental Health and Mental Health Services Research, 49(1), 139–151. 10.1007/s10488-021-01152-4 [PubMed: 34297259]

Hurlburt MS, Garland AF, Nguyen K, & Brookman-Frazee L (2010). Child and family therapy process: Concordance of therapist and observational perspectives. Administration and Policy in Mental Health and Mental Health Services Research, 37(3), 230–244. [PubMed: 19902347]

Isenhart C, Dieperink E, Thuras P, Fuller B, Stull L, Koets N, & Lenox R (2014). Training and maintaining motivational interviewing skills in a clinical trial. Journal of Substance Use, 19, 164–170.

James LR, Demaree RG, & Wolf G (1984). Estimating within-group interrater reliability with and without response bias. Journal of Applied Psychology, 69, 85–98.

Jensen-Doss A, Smith AM, Walsh LM, Ringle VM, Casline E, Patel Z, … & Webster R (2020). Preaching to the choir? Predictors of engagement in a community-based learning collaborative. Administration and Policy in Mental Health and Mental Health Services Research, 47, 279–290.

Kline P (1979). Psychometrics and Psychology. London: Academic Press

Kobak KA, Wolitzky-Taylor K, Craske MG, & Rose RD (2017). Therapist training on cognitive behavior therapy for anxiety disorders using internet-based technologies. Cognitive Therapy and Research, 41, 252–265. [PubMed: 28435174]

Koo TK, & Li MY (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15, 155–163. [PubMed: 27330520]

Kreft I, & De Leeuw J (1998). Varying and random coefficient models. Introducing Multilevel Modeling, 35–56.

Kuznetsova A, Brockhoff PB, & Christensen RHB (2017). lmerTest Package: Tests in linear mixed effects models. Journal of Statistical Software, 82, 1–26.

LeBreton JM, & Senter JL (2008). Answers to 20 questions about interrater reliability and interrater agreement. Organizational research methods, 11(4), 815–852.

Lock J (2015). An update on evidence-based psychosocial treatments for eating disorders in children and adolescents. Journal of Clinical Child and Adolescent Psychology, 44, 707–721. [PubMed: 25580937]

Lyon A, & Koerner K (2016). User-centered design for psychosocial intervention development and implementation. Clinical Psychology: Science and Practice, 23(2), 180–200. [PubMed: 29456295]

Martino S, Ball S, Nich C, Frankforter TL, & Carroll KM (2009). Correspondence of motivational enhancement treatment integrity ratings among therapists, supervisors, and observers. Psychotherapy research, 19(2), 181–193. [PubMed: 19396649]

Mastroleo NR, Humm L, Williams CM, Kiluk BD, Hoadley A, & Magill M (2020). Initial testing of a computer-based simulation training module to support clinicians' acquisition of CBT skills for substance use disorder treatment. Journal of Substance Abuse Treatment, 114, 108014. [PubMed: 32527511]

McCart MR, & Sheidow AJ (2016). Evidence-based psychosocial treatments for adolescents with disruptive behavior. Journal of Clinical Child and Adolescent Psychology, 45, 529–563. [PubMed: 27152911]

McLeod BD, Cox JR, Jensen-Doss A, Herschell A, Ehrenreich-May J, & Wood JJ (2018). Proposing a mechanistic model of clinician training and consultation. Clinical Psychology: Science and Practice, 25(3), e12260. [PubMed: 30713369]

McLeod BD, Southam-Gerow MA, Tully CB, Rodriguez A, & Smith MM (2013). Making a case for treatment integrity as a psychosocial treatment quality indicator for youth mental health care. Clinical Psychology: Science and Practice, 20(1), 14–32. [PubMed: 23935254]

Minuchin S, & Fishman HC (1981). Family therapy techniques. Cambridge, MA: Harvard University Press.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/.

RStudio Team (2020). RStudio: Integrated Development for R. Boston, MA. Retrieved from http://www.rstudio.com/

Schafer JL, & Graham J (2002). Missing data: Our view of the state of the art. Psychological Methods, 7, 147–177. [PubMed: 12090408]

Schoenwald SK, Garland AF, Chapman JE, Frazier SL, Sheidow AJ, & Southam-Gerow MA (2011). Toward the effective and efficient measurement of implementation fidelity. Administration and Policy in Mental Health and Mental Health Services Research, 38(1), 32–43. [PubMed: 20957425]

Shimokawa K, Lambert MJ, & Smart DW (2010). Enhancing treatment outcome of patients at risk of treatment failure: meta-analytic and mega-analytic review of a psychotherapy quality assurance system. Journal of consulting and clinical psychology, 78(3), 298. [PubMed: 20515206]

Shrout P, & Fleiss J (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420–428. 10.1037/0033-2909.86.2.420 [PubMed: 18839484]

Sibley MH, Ortiz M, Rios-Davis A, Zulauf-McCurdy CA, Graziano PA, & Bickman L (2022). Stakeholder-generated implementation strategies to promote evidence-based ADHD treatment in community mental health. Administration and Policy in Mental Health and Mental Health Services Research, 49, 44–58. [PubMed: 33988847]

Stewart RE, Williams N, Byeon YV, Buttenheim A, Sridharan S, Zentgraf K, … & Beidas RS (2019). The clinician crowdsourcing challenge: using participatory design to seed implementation strategies. Implementation Science, 14, 1–8. [PubMed: 30611302]

Stirman SW (2020). Commentary: Challenges and opportunites in the assessment of fidelity and related constructs. Administration and Policy in Mental Health and Mental Health Services Research, 47, 932–934. [PubMed: 32715432]

Tanner-Smith EE, Wilson SJ, & Lipsey MW (2013). The comparative effectiveness of outpatient treatment for adolescent substance abuse: A meta-analysis. Journal of Substance Abuse Treatment, 44, 145–158. [PubMed: 22763198]

Valenstein-Mah H, Greer N, McKenzie L, Hansen L, Strom TQ, Wiltsey Stirman S, … & Kehle-Forbes SM (2020). Effectiveness of training methods for delivery of evidence-based psychotherapies: a systematic review. Implementation Science, 15, 1–17. [PubMed: 31900167]

Wain RM, Kutner BA, Smith JL, Carpenter KM, Hu MC, Amrhein PC, & Nunes EV (2015). Self-report after randomly assigned supervision does not predict ability to practice motivational interviewing. Journal of substance abuse treatment, 57, 96–101. [PubMed: 25963775]

Webb CA, DeRubeis RJ, & Barber JP (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. Journal of Consulting and Clinical Psychology, 78, 200–211. [PubMed: 20350031]

Weersing RV, Jeffreys M, Do M, Schwartz K, & Bolano C (2017). Evidence base update of psychosocial treatments for child and adolescent depression. Journal of Clinical Child and Adolescent Psychology, 46, 11–43. [PubMed: 27870579]

Weisz JR, Bearman S, Santucci L, & Jensen-Doss A (2017). Initial test of a principle-guided approach to transdiagnostic psychotherapy with children and adolescents. Journal of Clinical Child and Adolescent Psychology, 46, 44–58. [PubMed: 27442352]

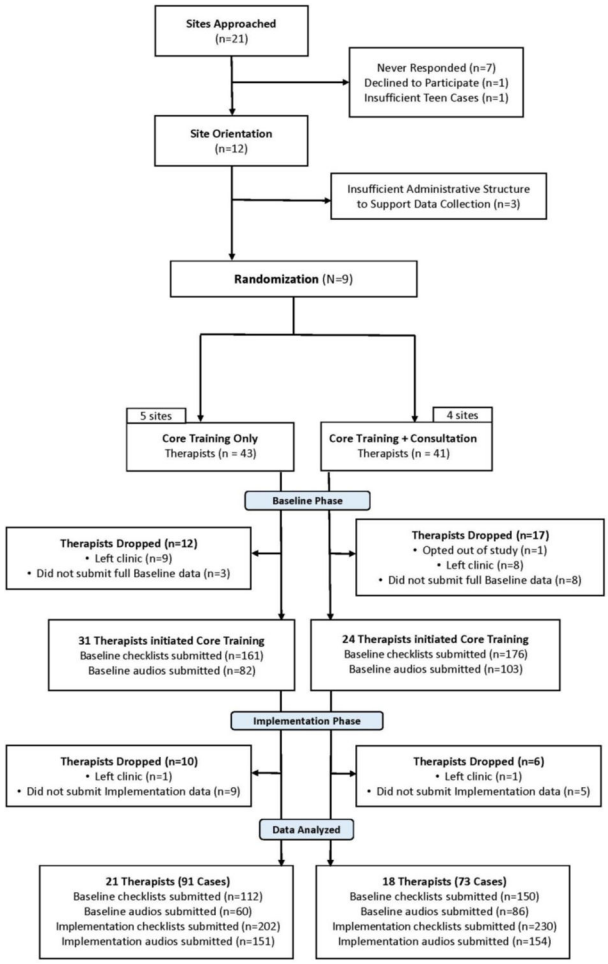

Figure 1.
CONSORT Diagram

Table 1.

Baseline Therapist Characteristics and Disposition toward Family Therapy

	Full Sample[a] *N* = 84	Study Sample *N* = 39	CT *n* = 21	CTC *n* = 18
		M (SD) or *n (%)*		
Age	31.7 (9.3)	30.8 (9.1)	30.9 (7.2)	30.4 (9.9)
Race/Ethnicity				
White Non-Hispanic	60 (71)	31 (79)	16 (73)	15 (77)
Hispanic	7 (8)	4 (10)	2 (9)	2 (12)
Black/African-American	3 (4)	2 (5)	2 (9)	--
Asian	2 (2)	1 (3)	1 (5)	--
Other	3 (4)	1 (3)	--	--
Sex				
Female	62 (74)	34 (87)	18 (86)	16 (88)
Male	12 (14)	5 (13)	3 (14)	2 (12)
Education				
Master's	70 (83)	34 (97)	19 (96)	17 (100)
Associates/Bachelor's	5 (6)	1 (3)	1 (4)	--
Employment				
Staff (full-time or part-time)	73 (87)	39 (100)	21 (100)	17 (100)
Trainees	2 (2)	--	--	--
Post-degree experience (years)	3.6 (5.1)	3.7 (4.8)	3.2 (4.2)	4.4 (5.6)
Employment at clinic (years)	1.9 (3.3)	2.1 (3.9)	1.4 (1.6)	3.1 (5.5)
Average caseload	32.5 (25.3)	37.6 (29.5)	36.9 (20.4)	38.5 (38.9)
Disposition toward FT				
Allegiance to FT	3.0 (.94)	3.1 (1.1)	2.9 (1.2)	3.4 (.93)
Self-rating FT skills	2.9 (.87)	3.0 (.92)	2.8 (.87)	3.2 (.95)
Importance to include family	3.8 (.84)	3.9 (.84)	4.0 (.81)	3.8 (.81)
Importance to learn FT	3.8 (.98)	3.7 (.99)	3.8 (.99)	3.7 (.99)

[a]10 therapists did not complete demographic questionnaire

Table 2.

Descriptive statistics for outcome variables.

	Baseline (N=74)		Implementation (N=192)	
	Observer[a]	Therapist	Observer	Therapist
	Mean (SD)	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*
FT Total Score	0.87 (0.43)	1.22 (0.81)	0.82 (0.47)	0.78 (0.58)
Family Engagement[b]	0.41 (0.44)	0.84 (1.00)	0.43 (0.52)	0.57 (0.74)
Relational Orientation[c]	1.24 (0.53)	1.75 (0.87)	1.23 (0.56)	1.19 (0.75)
Interactional Change[d]	0.26 (0.40)	0.93 (0.91)	0.27 (0.43)	0.46 (0.63)

Note. N=266 sessions.

[a]Ratings were averaged across observers.

[b]Average of the following items: Parent Collaboration, Love and Commitment, Parent Ecosystem, Adolescent Goal Collaboration.

[c]Average of the following items: Relational Focus, Focus on Process, Reframe, Relational Reframe, Family-Focused Rationale.

[d]Average of the following items: Prepare for Future Interactions, Stimulate Dialogue, Coach and Process, Teach Family Skill.

Table 3.

Intra-class correlation coefficients* for reliability between inter-observer ratings and therapist-therapist ratings on the ITT-CEFT.

	Baseline (N=74)	Implementation (N=192)	Core Training (N=98)	Core Training + Consultation (N=94)
	ICC $_{(1, 2)}$	*ICC* $_{(1, 2)}$	*ICC* $_{(1, 2)}$	*ICC* $_{(1, 2)}$
FT Total Score	.35	.72	.64	.67
Family Engagement	.43	.72	.51	.65
Relational Orientation	.42	.60	.60	.56
Interactional Change	.25	.63	.43	.60

Note. N=266 sessions.

* Intraclass correlation coefficients were calculated using the one-way random formula ($ICC_{1,2}$; Shrout & Fleiss, 1979).

Table 4.

Mixed effects ANOVA models testing Phase effects: Baseline versus Implementation.

	df	F	*p*	η^2
Therapist Accuracy [a]				
FT Total Score	101.78	18.84	<.001	0.16
Family Engagement	100.22	5.76	0.02	0.05
Relational Orientation	111.93	5.88	0.02	0.05
Interactional Change	103.26	11.74	<.001	0.10
FT Dose: Therapist-Report [b]				
FT Total Score	243.50	37.35	<.001	0.13
Family Engagement	244.10	9.27	<.01	0.04
Relational Orientation	251.05	32.13	<.001	0.11
Interactional Change	244.18	26.09	<.001	0.10
FT Dose: Observer Ratings				
FT Total Score	256.48	1.30	.56	--
Family Engagement	255.27	1.01	.30	--
Relational Orientation	261.12	0.47	.43	--
Interactional Change	259.39	0.57	.45	--

Note. N=266 sessions. Effect sizes (η^2) were calculated only for significant effects.

[a] r_{wg} coefficients assessing agreement between therapist-report and observer ratings were used in the model.

[b] Results were consistent using the full sample of therapist-report checklists, inclusive of checklists for which corresponding audio-recordings were not coded (N=432 sessions).

Table 5.

Results of multilevel models testing overall (Unconditional) and comparative (Conditional) growth in Therapist Accuracy and FT Dose during Implementation.

	Change Over Time[a]				Group Differences in Change over Time[b]			
	B	B (SE)	*p*	f^2	β	B (SE)	*P*	f^2
Therapist Accuracy [c]								
FT Total Score	<0.01	<0.01 (<0.00)	0.43	--	<0.01	<0.01 (<0.01)	0.21	--
Family Engagement	<0.01	<0.01 (0.01)	0.81	--	<−0.00	−0.01 (<0.01)	0.87	--
Relational Orientation	<0.01	<0.01 (<0.00)	0.70	--	<−0.00	−0.01 (<0.01)	0.86	--
Interactional Change	−0.01	0.01 (0.01)	0.03	--	0.01	0.01 (<0.01)	0.03	--
Therapist-Report FT Dose [d]								
FT Total Score	<0.01	<0.01 (<0.01)	0.89	--	<0.01	0.01 (<0.01)	0.36	--
Family Engagement	<0.01	<0.01 (<0.01)	0.63	--	<0.01	<0.01 (<0.01)	0.88	--
Relational Orientation	<0.01	<0.01 (<0.01)	0.75	--	0.01	0.01 (0.01)	0.23	--
Interactional Change	<−0.01	−0.01 (<0.01)	0.46	--	<−0.01	<0.01 (<0.01)	0.94	--
Observer-Report FT Dose								
FT Total Score	<0.01	<0.01 (<0.01)	0.12	--	<0.01	0.01 (<0.01)	0.36	--
Family Engagement	<0.01	<0.01 (<0.01)	0.24	--	<0.01	<0.01 (<0.01)	0.88	--
Relational Orientation	<0.01	0.01 (0.01)	0.37	--	0.01	0.01 (0.01)	0.23	--
Interactional Change	<0.01	<0.01 (<0.01)	0.11	--	<−0.01	<0.01 (<0.01)	0.94	--

Note. N= 192 Implementation phase sessions. SE = Standard error. Effect sizes (f^2) were calculated only for significant findings.

[a] Unconditional models tested the average slope of each outcome over the Implementation phase.

[b] Conditional models tested study Condition as the predictor variable (coded 0 = Core Training, 1 = Core Training + Consultation) and included therapist random effects.

[c] r_{wg} coefficients assessing agreement between therapist-report and observer ratings were used in the model.

[d] Results were consistent using the full sample of therapist-report checklists, inclusive of checklists for which corresponding audio-recordings were not coded (N=432 sessions).