Administration and Policy in Mental Health and Mental Health Services Research (2022) 49:139–151
https://doi.org/10.1007/s10488-021-01152-4

ORIGINAL ARTICLE

Online Training of Community Therapists in Observational Coding of Family Therapy Techniques: Reliability and Accuracy

Aaron Hogue[1,4] · Nicole Porter[1] · Molly Bobek[1] · Alexandra MacLean[1] · Lila Bruynesteyn[1] · Amanda Jensen-Doss[3] · Sarah Dauber[1] · Craig E. Henderson[2]

Accepted: 14 July 2021 / Published online: 23 July 2021

Abstract

A foundational strategy to promote implementation of evidence-based interventions (EBIs) is providing EBI training to therapists. This study tested an online training system in which therapists practiced observational coding of mock video vignettes demonstrating family therapy techniques for adolescent behavior problems. The study compared therapists ratings to gold-standard scores to measure therapist *reliability* (consistency across vignettes) and *accuracy* (approximation to gold scores); tested whether reliability and accuracy improved during training; and tested therapist-level predictors of overall accuracy and change in accuracy over time. Participants were 48 therapists working in nine community behavioral health clinics. The 32-exercise training course provided online instruction (about 15 min/week) in 13 core family therapy techniques representing three modules: Family Engagement, Relational Orientation, Interactional Change. Therapist reliability in rating technique presence (i.e., technique recognition) remained moderate across training; reliability in rating extensiveness of technique delivery (i.e., technique judgment) improved sharply over time, from poor to good. Whereas therapists on average overestimated extensiveness for almost every technique, their tendency to give low-accuracy scores decreased. Therapist accuracy improved significantly over time only for Interactional Change techniques. Baseline digital literacy and submission of self-report checklists on use of the techniques in their own sessions predicted coding accuracy. Training therapists to be more reliable and accurate coders of EBI techniques can potentially yield benefits in increased EBI self-report acumen and EBI use in daily practice. However, training effects may need to improve from those reported here to avail meaningful impact on EBI implementation.

Trial Registration: The parent clinical trial is registered at www.ClinicalTrials.gov, ID: NCT03342872 (registration date: 11.10.17).

Keywords Family therapy · Online training · Observational coding · Reliability · Evidence-based intervention · Usual care

Introduction

A foundational strategy to promote implementation of evidence-based interventions (EBIs) in behavioral health settings is providing EBI training to therapists. Manualized EBIs use standardized clinician training protocols, often in conjunction with ongoing expert consultation and provider certification procedures, to ensure high-fidelity model delivery and continuous quality improvement. Yet, over three decades of research show that whereas EBI training and consultation can increase EBI satisfaction and immediate skill acquisition among community practitioners, they do not reliably or appreciably increase practitioner use of EBIs with their clients, especially after interactions with experts cease (Beidas et al., 2019; Valenstein-Mah et al., 2020). Moreover,

✉ Aaron Hogue
ahogue@toendaddiction.org

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

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

[3] Department of Psychology, University of Miami, Miami, FL, USA

[4] Partnership to End Addiction, 485 Lexington Avenue, 3rd floor, New York, NY 10017, USA

training and quality procedures for manualized EBIs can be prohibitively costly and resource-intensive to sustain in everyday care. There remains great need to develop EBI training procedures that are both effective and scalable.

This study tests an online EBI training system in which therapists practice observational coding of video re-enactments from real therapy sessions. The training is a 32-exercise course on family therapy techniques for addressing conduct and substance use problems among adolescents; see "Method" section for a full description. It has three user-centered design features intended to maximize its practicality and effectiveness (Lyon & Koerner, 2016). First, the training course is housed on an online learning management system. Online training is a cost-effective way to reach large numbers of therapists and manage continuous trainings (Weingardt, 2004). In various formats it 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 training, believe it to be efficacious (Becker & Jensen-Doss, 2014), and believe it increases training accessibility and engagement (Ehrenreich-May et al., 2016). Second, only 15–20 min is required to complete each exercise, accessible via smartphone or computer. Third, the system 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 they equip clinicians with a diverse portfolio of techniques that can be flexibly used with clients showing comorbid, heterogeneous, and/or emerging clinical problems, making them well-suited for the eclectic treatment strategies that constitute usual care (Weisz et al., 2017).

Potential Benefits of EBI Training via Observational Coding Procedures

The EBI training system tested in this study makes observational coding the centerpiece of the learning experience. Each training exercise contains a brief video of a mock therapy session segment depicting a handful of core family therapy techniques for adolescent behavior problems (Author, 2019a, b); these scripted vignettes contain actors in client roles and expert family therapists in the clinician role. After viewing each vignette, therapists rate the extent to which specific techniques were present. Immediately after submitting their ratings, they receive automated feedback that reveals and justifies gold-standard scores (i.e., scores previously determined via expert consensus) for each technique.

These procedures mimic well-established observational coder training methods used in treatment research to verify EBI integrity (Hogue et al., 1996). A logical first step toward boosting the capacity of community therapists to implement EBIs with fidelity is improving their ability to accurately recognize and assess interventions they are expected to deliver (McLeod et al., 2018). Because each vignette depicts a cohesive set of techniques delivered within a realistic sequence of family 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). Overall, potential immediate benefits of coder training are increasing therapist declarative knowledge (i.e., factual knowledge and information about an EBI) as well as making 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).

Because practicing therapists are the learners in this system, a potential downstream benefit of therapist coder training is prompting increased EBI delivery in real-world practice (Stirman et al., 2010). Although live coaching and guided skills practice are the most effective means to acquire new clinical skills, video-based modeling has shown promise for augmenting use of EBIs of several kinds (e.g., Beidas et al., 2014; Weingardt et al., 2009). Learning to recognize and assess EBI techniques via longitudinal coder training could solidify the clinical acceptability of the techniques, further advance procedural knowledge of the techniques by modeling their delivery across diverse clinical contexts, and gradually strengthen therapist confidence and motivation to implement them in everyday care (McLeod et al., 2018).

Extant Research on Training Community Therapists in Observational Coding of EBIs

Despite the promise of a coder training approach, there is little extant research on whether community therapists can be accurate observational raters of EBIs. On the one hand, practitioners are not naturally proficient *self*-raters of their own treatment sessions, tending to significantly overestimate (or occasionally, underestimate) EBI use on post-session self-report measures (e.g., Caron et al., 2020). On the other hand, a few studies have shown that practitioners can be proficient *observational* raters when acting as research-trained fidelity judges in controlled trials of manualized EBIs (Carroll et al., 2000; Hogue et al., 2008). Such studies train clinicians to function as non-participant judges who receive expert instruction, monitoring, and financial compensation for supplying external ratings of trial session recordings.

Rarer still are studies that train 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. We found only two studies of this kind in the psychosocial treatment research literature; both examined whether therapist coder training augmented ongoing EBI consultation by helping recently trained therapists sustain EBI fidelity. Isenhart et al. (2014) described a group consultation process wherein two clinicians rated each others' sessions to offer feedback regarding EBI adherence. Caron and Dozier (2019) asked seven therapists trained in a parenting intervention for at-risk families to rate over 300 total segments of their own sessions; agreement between therapist and supervisor ratings was regularly reviewed during consultation meetings. Therapists demonstrated good-to-excellent agreement with supervisors for two targeted EBI techniques and fair-to-poor agreement for two others. There was also evidence that therapist use of the targeted EBI increased during the period of consultation that featured therapist coding.

Study Context and Specific Hypotheses

The current study tested whether therapist coder training could help community practitioners learn to assess family therapy techniques reliably and accurately. Family therapy is an evidence-based approach to treating adolescent conduct and substance use problems (Hogue et al., 2018; McCart & Sheidow, 2016), and it has an exemplary record of success in these clinical domains even when compared to other evidence-based approaches (Baldwin et al., 2012; Riedinger et al., 2017; Tanner-Smith et al., 2013). Yet, family therapy is not as widely practiced as other approaches in youth behavioral care (Riedinger et al., 2017). Thus there are compelling reasons to advance family therapy training opportunities for members of the clinical workforce who regularly treat adolescents. Note that we are also developing similar training resources focused on cognitive-behavioral therapy (Author, 2020).

This study was part of a randomized trial evaluating family therapy training and consultation procedures across nine behavioral health treatment sites. In the parent trial (Author, 2019a, b), all enrolled practitioners were invited to complete therapist coder training in family therapy techniques. In addition, half of the sites were selected to receive monthly consultation meetings with a family therapy expert. As described below and listed in Table 1, all practitioners were trained in 13 core family therapy techniques. These techniques are further grouped into three modules that were empirically derived from factor analyses on observational ratings of family therapy sessions (Author, 2019a, b) and constitute foundational areas of clinical and research focus in evidence-based family therapy models for adolescents (e.g., Robbins et al., 2011): family engagement, relational orientation, interactional change.

The current study had three aims focused on the effectiveness of the coder training system: (1) Compare therapist ratings to gold-standard scores assigned to the assorted family therapy techniques demonstrated in each vignette, calculating overall therapist *reliability* (consistency across vignettes) and *accuracy* (approximation to gold scores). Essentially, indices of reliability address the question, "How much does the length of the ruler change at each measurement?"; in contrast, indices of accuracy address a validity-related question, "How close is the actual ruler length to the true length?" (see Hallgren, 2012). (2) Test whether therapist reliability and accuracy improved over the duration of training, both overall and separately for each of the three foundational modules. (3) Determine whether overall accuracy, as well as change in accuracy over time, were associated

Table 1 Roster and brief coding prompts of the 13 family therapy techniques presented in the training system

Family therapy module	Family therapy technique	Brief coding prompt
Family engagement	Parent collaboration	Attempts to collaborate with parent(s) by instilling hope and/or involving them in treatment goals
Family engagement	Love and commitment	Enhances parental feelings of love and commitment
Family engagement	Parent ecosystem	Focuses on parent's non-parenting life as an adult person
Family engagement	Adolescent goal collaboration	Formulates family-oriented treatment goals with the adolescent
Relational orientation	Relational focus	Adopts a relational/systemic focus
Relational orientation	Focus on process	Asks clarifying questions and focuses on relational process, not content
Relational orientation	Reframe	Utilizes meaning-change interventions toward a new and/or more positive view
Relational orientation	Relational reframe	Reframes adolescent symptoms as relational problems that need relationship solutions
Relational orientation	Family-focused rationale	Offers a family-focused rationale for introducing a new skill, activity, or focus in therapy
Interactional change	Prepare for interactions	Prepares various participants separately for future interactions in or out of session
Interactional change	Stimulate dialogue	Prompts interactions among family members when they do not occur spontaneously
Interactional change	Coach and process	Coaches and processes family interactions in session
Interactional change	Teach family skills	Conducts in-session exercises, rehearsal, discussion, and/or feedback related to developing or practicing new behaviors

with therapist-level predictors of potential training effects: assigned study condition, baseline self-rated proficiency in family therapy, baseline digital literacy, proportion of training exercises completed, perceived utility of the training vignettes, and number of submitted therapist-report activity checklists (measuring therapist use of the 13 core family therapy techniques with their routine cases).

Method

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

Study Participants

Study participants included 48 therapists working in nine community-based mental health and substance use treatment clinics in various regions of a large northeastern state. Therapists (88% self-identified female, 12% male) averaged 31.4 (SD = 9.7) years of age. Self-identified race/ethnicity was 79% White Non-Latinx, 8% Latinx, 6% Black/African-American, 2% Asian, and 5% Other. A total of 96% had a master's level degree and 4% an associate's or bachelor's degree (data on master's degree specialization were not obtained); all participants indicated they were full-time or part-time staff—that is, none was an intern or resident. They averaged 3.6 (SD = 4.4) years of post-degree therapy experience and 2.3 (SD = 3.4) years of employment at the study clinic. In response to the query "Average caseload (total # individual, family, or group)", average reported caseload size was 35.1 (SD = 27.7) clients across individual, group, and family session formats.

Study Procedures: Parent Randomized Trial

All therapists working at partner treatment sites were eligible to participate in the parent randomized trial (Author, 2019a, b) on a voluntary basis if they met the following criteria: routinely treated clients age 13–21 years; agreed to submit at least two audiorecorded sessions and corresponding self-reported use of family therapy techniques (i.e., session activity checklists) prior to initiating coder training, as part of research efforts to collect baseline data on family therapy technique use; agreed to submit additional session recordings and activity checklists after coder training commenced. Virtually every therapist at each study site who met these criteria consented to participate. Prior to initiating the training course, trial-consenting therapists attended two on-site 90-min baseline workshops: The first introduced the online learning management system and its coder training procedures; the second introduced 13 family therapy techniques (see Table 1) that were the foci of training. After baseline workshops concluded, sites were randomized to study condition: Five sites employing 43 therapists were assigned to the coder training condition, and four sites employing 41 therapists were assigned to the coder training plus expert consultation condition. The coder training system is described below (see Training Intervention). The four sites assigned to coder training plus consultation received monthly one-hour group consultation from a family therapy expert for the purpose of reviewing therapist experiences of the coder training system and encouraging use of the family therapy techniques in routine practice.

Study Procedures: Participant Enrollment and Attrition

Because the current study focused on coding reliability and accuracy, an additional inclusion criterion was that therapists in the parent trial needed to have completed at least three coder training exercises, thereby providing a minimally sufficient sample of therapist-level coding acumen to inform reliability statistics. For this reason, the current study should be couched as examining "training participation" effects rather than "intent-to-train" effects. As described above, 84 therapists consented to participate in the parent trial; of these, 48 (57%) ultimately met all inclusion criteria for the current study. Of the 36 therapists in the parent trial who did not meet inclusion criteria, 17 left clinic employment prior to course activation, 11 did not submit sufficient baseline session data, 1 withdrew from the parent trial, and 7 completed fewer than three coder training exercises. The 36 therapists who did not meet inclusion criteria (attrited sample) did not differ from the 48 retained therapists (study sample) on any baseline characteristics. Therapists completed at least 3, and on average 19.5, training exercises (SD = 9.5) during the 32-exercise training course; one new exercise per week was released to therapists (N.B.: A handful of therapists hired by their respective clinics after study launch initiated the training course mid-stream [e.g., their first training exercise was Exercise 9 of the course] so as to be aligned with site peers in weekly session exposure). Out of the 48 therapists enrolled in the study, 44 (92%) completed at least 1 exercise during weeks 1–8 of the training course, 43 (90%) completed at least one exercise during weeks 9–16, 35 (73%) completed at least one exercise during weeks 17–24, and 24 (50%) completed at least one exercise during weeks 25–32. Therapists who discontinued the training course after week 24 (the point of largest dropoff) did not differ from training completers on any baseline characteristic. 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 (6 of the 48 therapists) had access to only 31 of the 32 exercises.

Training Intervention: Online Therapist Coder Training in Core Family Therapy Techniques for Adolescent Behavior Problems

The online therapist coder training system provides instruction in 13 core family therapy techniques for treating adolescent conduct and substance use problems; the techniques are listed in Table 1, along with brief descriptions used as immediate coding prompts. The 13 techniques were drawn from an empirical distillation process driven by factor analysis of observational adherence coding data from manualized family therapy models (Author, 2019a, b) and comprise three clinically coherent Modules: (1) *Family Engagement*: Parent Collaboration, Love and Commitment, Parent Ecosystem, Adolescent Goal Collaboration; (2) *Relational Orientation*: Relational Focus, Focus on Process, Reframe, Relational Reframe, Family-Focused Rationale; (3) *Interactional Change*: Prepare for Interactions, Stimulate Dialogue, Coach and Process, Teach Family Skills. Previous research conducted in usual care settings has shown that community therapists can deliver these techniques with strong fidelity (Hogue et al., 2017) and show positive impacts on long-term adolescent outcomes (Hogue et al., 2015, 2017); moreover, greater use of these core techniques is associated with better outcomes when delivered by family therapists and non-family therapists alike (Henderson et al., 2019).

The therapist coder training course was hosted on a web-based learning management system. In the current study, training exercises were released weekly via protected weblinks distributed by email to enrolled therapists. Each exercise remained accessible until completed by a given therapist, allowing them to stockpile exercises for later completion. Each exercise contained two synergistic parts. (1) *Didactic Instruction*: slides containing brief descriptions and exemplar therapist statements for selected techniques. Each exercise presents three total techniques (one per slide) during didactic instruction. In order to promote skill in differentiated coding, only two of these techniques appear in the vignette that follows. (2) *Mock Session Coding*: 5–8 min scripted video vignette modeling multiple techniques, followed by a standardized coding activity. Vignettes depict expert family therapists working with actors in family roles, all re-enacting therapy scenarios drawn from real cases. Each vignette illustrates several techniques that range from low to high extensiveness, and collectively they present a diverse group of therapists and families and showcase a range of therapist styles and presenting problems. After viewing the vignette, therapists are guided through a coding activity designed to grow their ability to recognize and evaluate technique delivery. Therapists are instructed to rate selected techniques on a 5-point Likert-type scale according to the thoroughness and frequency with which each was exhibited in the companion vignette: 0 (*Not at all*), 1 (*A little bit*), 2 (*Moderately*), 3 (*Quite a bit*), 4 (*Extensively*). Therapists then rate five selected techniques; to sharpen technique recognition and discrimination, only three of the selected techniques actually appear in the vignette. Upon completing the coding activity, therapists receive immediate scoring feedback in the form of gold-standard scores determined via consensus scoring by family therapy coding experts. Gold scores for the three depicted techniques are supported by evidence in the form of verbatim statements (uttered by the therapist in the vignette) that exemplify each technique.

Measures

Family therapy proficiency (Hogue et al., 2014) was captured using a two-item self-report variable that averaged therapists' own judgments about their (1) degree of allegiance to and (2) perceived technical skill in family therapy techniques, rated on a 5-point scale from 0 (*None*) to 4 (*High*). *Digital literacy* was captured using a two-item self-report variable that averaged therapists' judgement about their aptitude in (1) using web-based tools and programs and (2) reading and understanding data graphs and figures (see Davis, 1989). *Training exercises completed* was automatically logged for each therapist by the learning management system; an exercise was deemed "completed" only if the link was active for longer than one minute (to disqualify obvious instances of "click through"). This study used a variable representing the proportion of assigned exercises completed by each therapist. *Perceived vignette utility* was calculated by averaging scores from two survey items completed online by each therapist at the end of training exercises every 4th week, starting with exercise eight: How effective were the video vignettes in demonstrating family therapy techniques? How relevant and/or useful were the demonstrated techniques in video vignettes to your clinical practice? Utility scores were based on a 5-point scale from 0 (*Not at all*) to 4 (*Extensively*). *Therapist-report family therapy activity* was measured using a validated post-session therapist-report checklist (Hogue et al., 2014) indicating the extent to which each of 13 family therapy techniques (identical in content and scoring to those in the online training course described above) were delivered in the given session. Therapists were asked to complete the checklist after every therapy session involving an adolescent client, to complement and enhance the effects of online training.

Dependent Variables and Plan of Analysis

Dependent variables (operationalized below) were drawn from coding data entered by therapists into the online training system. As described above, therapists rated the thoroughness and/or frequency with which they observed family therapy techniques in video vignettes using a 5-point scale:

0 (*Not at all*), 1 (*A little bit*), 2 (*Moderately*), 3 (*Quite a bit*), 4 (*Extensively*).

Study analyses occurred in three stages. In Stage 1, descriptive statistics were calculated for the 13 family therapy techniques presented across video vignettes in all course exercises as well as therapist reliability and accuracy. *Part I. Reliability:* Two indices of therapist interrater reliability with gold-standard scores were calculated: (1) intraclass correlation coefficient (ICC$_{(1,2)}$; Shrout & Fleiss, 1979) and (2) Cohen's kappa (κ; Cohen, 1960). We used one-way random ICCs with averaged rater scores and absolute agreement estimation to assess the level of agreement between therapist and gold-standard scores for each family therapy technique. We adopted reliability standards for ICCs based on Cicchetti's (1994) criteria for classifying coefficient magnitudes: Below .40 is poor, .40–.59 fair, .60–.74 good, and .75–1.00 excellent. Next, we calculated Cohen's kappa to supplement ICC data by capturing therapists' ability simply to recognize when given interventions appeared. To facilitate kappa calculations the therapist-report and gold-standard scores were re-coded to become dichotomous: either presence (scores 1 through 4 were recoded as "1") or absence ("0") of each item in each vignette. We interpreted kappa magnitudes based on the Landis and Koch (1977) criteria: <0 is no agreement, 0–.20 slight, .21–.40 fair, .41–.60 moderate, .61–.80 substantial, and .80–1.0 perfect. *Part II. Accuracy:* Therapist accuracy was assessed in two ways: (a) sensitivity and specificity; and (b) mean differences. We used sensitivity and specificity to understand therapist ability to discriminate accurately between target items (i.e., the three techniques, from among the five that therapists were asked to code for a given vignette, that actually appeared in the vignette) and contrast items (i.e., the two techniques that did not appear). Sensitivity was calculated as the proportion of correctly identified target items (true positives) to the total number of target items presented across the training course. Specificity was calculated as the proportion of correctly identified contrast items (true negatives) to the total number of contrast items. Finally, we compared mean scores of therapist ratings averaged across all therapists, for individual items and the averaged Module scores and Scale Total score, to corresponding mean gold-standard scores using one-sample *t*-tests.

In Stage 2, we examined change in therapist reliability and accuracy over time. For reliability data (ICC and kappa), there were not sufficient numbers of data points to model change on a weekly basis (i.e., attempts to estimate distinct reliability coefficients for each of 32 exercise weeks did not converge). Instead, we examined change over time in two ways: composite sores, and discrepancy scores. First, we grouped reliability data into four 8-week intervals and re-calculated ICC and kappa statistics; this allowed us to compare therapist and gold-standard scores on composite indices averaged across the 8 exercise weeks represented in each time interval. Second, we calculated a discrepancy score for every item for every vignette by taking the absolute value of therapist score minus gold-standard score. We then analyzed these discrepancy scores in two ways: (a) To generate a basic index of sample-wide accuracy that could be compared directly with reliability data, for each 8-week interval we counted the total number of low-accuracy scores across all therapists and items. A discrepancy score was categorized as "low-accuracy" if its value was >1 (i.e., discrepancy scores of 2, 3, or 4). (b) To examine change in accuracy over time, we used multilevel modeling to test the slope of discrepancy scores (averaged across all five items for each vignette) over the 32-exercise course. We conducted multilevel modeling with weeks (i.e., coding exercises) nested within therapists using the software package Mplus (version 8.2; Muthén & Muthén, 1998–2021) and activating the option 'twolevel' (weeks within therapists). Although it was also the case that therapists were nested within agencies, a preliminary examination of the intraclass correlation coefficient revealed that the ratio of between-cluster variance to total variance for Agency was nearly zero (ICC = .004), indicating that ignoring nesting at this third level would have negligible impact on study results (Kreft & DeLeeuw, 1998). In comparison, the ICC for Therapist was .085. Therefore, for modeling parsimony we did not include Agency as an additional cluster variable. We used Full Information Maximum Likelihood (FIML) estimation to accommodate and reduce potential bias due to missing data. Effect sizes were indexed by the standardized regression coefficient (β), which is available in Mplus for multilevel models only with Bayes estimation. Because our data satisfied FIML assumptions (e.g., multivariate normality), separate models were conducted to estimate effect sizes. We tested change in accuracy using the Discrepancy Total score (averaging all 13 items) and then separately using each of the Module average scores (Family Engagement, Relational Orientation, Interactional Change); for the Module analyses we used Bonferroni correction to adjust for family-wise error.

In Stage 3, multilevel modeling was used to test predictors of change in therapist accuracy using the model specifications described above. Due to data limitations at the Module level that prohibited predictive analyses, we tested predictors of change for the Discrepancy Total score only. We screened predictor variables for association with Discrepancy Total using a threshold of $p < .10$. For parsimony and to preserve power, only predictors exceeding this threshold were included in the final conditional model. A total of six variables were screened for association with change in Discrepancy Total: three time-varying predictors (proportion of exercises completed, vignette utility score, number of submitted activity checklists) and three time-invariant baseline predictors (study condition, family therapy proficiency, digital literacy). We also screened interactions between study condition and the three time-varying predictors. Time-varying predictors and interaction terms exceeding the threshold were entered in level 1 to model within-person change over time; time-invariant predictors exceeding the threshold were entered in level 2 to model individual differences between therapists. As described above, Discrepancy Total score was entered as the dependent variable and therapist nesting effects were adjusted for using multilevel modeling.

Results

Therapist Overall Reliability and Accuracy in Rating Family Therapy Techniques

Indices of therapist reliability and accuracy in rating 13 core family therapy techniques are presented in Table 2. *Reliability.* According to Cicchetti's (1994) classification for ICC magnitudes, items ranged from poor to excellent: Excellent, 1 item (ICC = .83); Good, 4 items (ICC range = .64 to .69); Fair, 4 items (range = .40 to .58); and Poor, 4 items (range = − .55 to .31). Based on Landis and Koch's (1977) classification for kappa magnitudes, items ranged from Slight to Substantial agreement: Substantial, 2 items (κ range = .66 to .69); Moderate, 1 item (κ = .59); Fair, 6 items (range = .30 to .39); Slight, 3 items (range = .03 to .16). *Accuracy.* Therapist sensitivity in identifying target items

Table 2 Reliability and accuracy statistics for 13 family therapy techniques presented in the training system, averaged across 32 weeks

Family therapy Modules & techniques	Average concordance: continuous	Average concordance: dichotomous		Average score: gold-standard	Average score: therapist	Equality of means
	$ICC_{(1,2)}$	Cohen's kappa	Sensitivity[a] %/ specificity[b]%	M (SD)	M (SD)	*t*-test
Family engagement						
1. Parent collaboration	0.64	0.59	89/71	1.17 (1.36)	1.28 (1.27)	9.37***
2. Love and commitment	0.65	0.66	98/62	1.69 (1.43)	1.97 (1.27)	9.93***
3. Parent ecosystem	0.58	0.35	88/59	0.47 (0.78)	1.08 (1.24)	9.24***
4. Adolescent goal collab.	0.40	0.34	81/53	1.75 (1.66)	1.38 (1.24)	− 6.32***
Relational orientation						
5. Relational focus	− 0.55	–	93/–	3.39 (0.80)	2.19 (1.09)	− 15.11***
6. Focus on process	0.11	0.16	98/15	1.72 (1.32)	2.34 (1.05)	11.73***
7. Reframe	0.30	0.13	88/27	1.90 (1.19)	1.88 (1.13)	− .33
8. Relational reframe	0.31	0.03	83/21	1.33 (1.27)	1.82 (1.25)	7.65***
9. Family-focused rationale	0.58	0.33	88/44	1.58 (1.11)	1.68 (1.22)	1.68
Interactional change						
10. Prepare for interactions	0.83	0.39	87/58	1.08 (1.67)	1.48 (1.54)	3.83***
11. Stimulate dialogue	0.67	0.32	99/30	1.63 (1.53)	2.28 (1.33)	9.51***
12. Coach and process	0.69	0.69	95/72	1.50 (1.37)	2.02 (1.54)	6.82***
13. Teach family skills	0.53	0.30	91/38	1.17 (1.17)	1.78 (1.37)	6.41***
Family engagement score	–	–	–	**1.27 (1.23)**	**1.37 (1.10)**	**2.74****
Relational orientation score	–	–	–	**1.86 (1.00)**	**1.98 (.94)**	**3.45****
Interactional change score	–	–	–	**1.10 (1.16)**	**1.72 (1.28)**	**13.05*****
Scale total score	–	–	–	**1.59 (0.41)**	**1.78 (0.73)**	**9.88*****

Kappa was not calculated for Relational Focus due to 100% appearance as a target item (it never appeared as a contrast item)

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

[a] Rate of identifying true target items. Targets items were items that were depicted in the vignette and received a gold-standard score of 1–4. Each weekly vignette coding activity included three target items. Items were presented as target items 36–100% of total appearances (mean = 64% positive appearances)

[b] Rate of identifying true contrast items. Contrast items were items that were not depicted in the vignette and received a gold-standard score of 0. Each weekly vignette coding activity contained two contrast items

ranged from 83 to 99%; specificity in identifying contrast items ranged from 15 to 72%. One sample *t*-tests comparing mean therapist scores to gold-standard scores revealed significant differences ($p < .05$) for 11 individual items, all three Modules, and Scale Total score. Notably, therapist scores were higher than gold-standard scores on 9 individual items, all Module scores, and Scale Total score; they were lower on 2 items.

Growth in Therapist Reliability and Accuracy Over Time

Therapist reliability with gold-standard scores across four 8-week intervals is presented in Table 3. Based on ICCs, therapist reliability in each of the first three intervals (weeks 1–8, 9–16, 17–24) was Poor (ICC < .40); in interval 4 (weeks 25–32) reliability was Good (ICC = .64). In an effort to diagnose whether this substantial increase in reliability during interval 4 might be an artifact of differential therapist retention—that is, more adept coders simply survived longer in training, and/or less adept coders dropped out earlier—we re-examined ICC data for the subgroup of training completers who provided data during all four intervals (n = 24; 50% of the full coding sample). The pattern of ICCs for this completer subgroup replicated the pattern observed for the full sample: Reliability during the first three intervals (ICCs = − .06, − .19, and .27, respectively) was Poor; reliability during interval 4 (ICC = .64) was Good. Table 3 also depicts reliability based on kappa (any vs. none judgements); reliability of this kind was moderate in the first, second, and fourth time intervals and fair in the third time interval. Also, over the course of 32 exercise weeks therapists scored individual items on 4305 occasions collectively across all exercises; therapist score differed from the gold-standard score by > 1 point (i.e., low-accuracy scores) on 1310 (30%) of scoring occasions. The number and percentage of low-accuracy scores decreased noticeably over time: 515 (39%) in weeks 1–8, 335 (26%) in weeks 9–16, 313 (24%) in weeks 17–24, and 147 (11%) in weeks 24–32.

Results of multilevel modeling examining change in therapist accuracy are presented in Table 4. The trajectory of the Discrepancy Total score was not statistically significant; that is, there was no evidence of positive training effects in the form of shrinking discrepancy scores. As depicted in Fig. 1, analyses of Module scores showed significant linear change with a downward slope for Interactional Change techniques (B = − .01, *SE* = .002, *pseudo z* = − 3.70, $p < .001$, $\beta = -.14$); this indicates that therapists became more accurate over time in scoring items related to arranging and coaching family interactions in session. No significant changes were detected for Family Engagement (B = .002, *SE* = .001, *pseudo z* = 1.58, *p* = .11, $\beta = .03$) or Relational Orientation (B = − .002, *SE* = .001, *pseudo z* = − 1.40, *p* = 15, $\beta = -.05$) techniques.

Predictors of Training Effects

Results of predictor variable screening are presented in Table 4; we report variables that exceeded the pre-established $p < .10$ threshold here. The β coefficients used as effect sizes can be interpreted as follows (Cohen, 1988): Small = .10, Medium = .30, Large = .50. One time-invariant variable significantly predicted Discrepancy Total score: digital literacy (b = − .11, *SE* = .06, *pseudo z* = − 1.92, *p* = .05, $\beta = -.29$). There were also two significant time-varying predictors: vignette utility (b = − .06, *SE* = .03, *pseudo z* = − 2.04, *p* = .04, $\beta = -.05$) and number of submitted activity checklists (b = − .01, *SE* = .01, *pseudo z* = − 1.61, *p* = .10, $\beta = -.02$). Interactions between study condition and time-varying predictors were not significant. Results of multilevel modeling including these three predictors are presented also in Table 4. Digital literacy (b = − .13, *SE* = .06, *pseudo z* = − 2.34, *p* = .02, $\beta = -.36$) and number of submitted checklists (b = − .02, *SE* = .18, *pseudo z* = − 2.18, *p* = .03, $\beta = -.04$) remained significant, showing a medium and small effect size, respectively; vignette utility did not remain significant. Significant effects were in the expected direction: Higher self-ratings on digital literacy and greater number of submitted checklists submitted were associated with decreases in discrepancy score. More specifically: For every 1 unit superiority in digital literacy there was a .36 unit decrease in discrepancy score; for every 1 unit increase in submitted checklists there was a .04 unit decrease in discrepancy score.

Table 3 Reliability statistics averaged across four time intervals representing the 32-week training system for all items

Time interval	Training weeks	Low-accuracy scores[a] N (%)	ICC $_{(1,2)}$	Cohen's kappa
1	1–8	515 (39.3%)	.00	.47
2	9–16	335 (25.6%)	− .06	.42
3	17–24	313 (23.9%)	.26	.34
4	25–32	147 (11.2%)	.64	.42

[a] A count of the number of scoring occasions on which therapist score differed from the corresponding gold-standard score by 2, 3, or 4 points; over the course of all four time intervals there were 1310 total low-accuracy scores submitted by participants

Table 4 Results of multilevel models predicting change in discrepancy score over the 32-week training system

	β	B (SE)	*p*-value	95% CI
Mean discrepancy score, all items[a]	− 0.36	− 0.00 (0.00)	.15	[− .01, .00]
Univariate models[b]				
Time-invariant predictors (baseline)				
Family therapy allegiance	− .01	− .00 (.05)	.94	[− .08, .07]
Experimental condition	− .01	.01 (.08)	.93	[− .12, .14]
Digital literacy	− .29	− .11 (.06)	.05	[− .20, − .02]
Time-varying predictors				
Video vignette utility	− .05	− .06 (.03)	.04	[− .11, − .01]
Percent weeks completed	− .06	− .00 (.00)	.40	[− .00, .00]
Number of submitted checklists	− .02	− .01 (.01)	.10	[− .02, .00]
Final conditional model				
Time-invariant predictors (baseline)				
Digital literacy	− .36	− .13 (.06)	.02	[− .22, .04]
Time-varying predictors				
Video vignette utility	− .02	− .04 (.03)	.28	[− .09, − .02]
Number of submitted checklists	− .04	− .02 (.18)	.03	[− .03, − .00]

Negative estimates indicate a decrease in discrepancy score over time

SE standard error

[a] The unconditional model tested the average discrepancy score slope

[b] Variables were entered one at a time to screen for potential significant effects. Variables exceeding a threshold of $p < .10$ were included in the final model. Interactions between experimental condition and the three time-varying predictors were also screened for inclusion in the final conditional model and did not exceed the $p < .10$ threshold

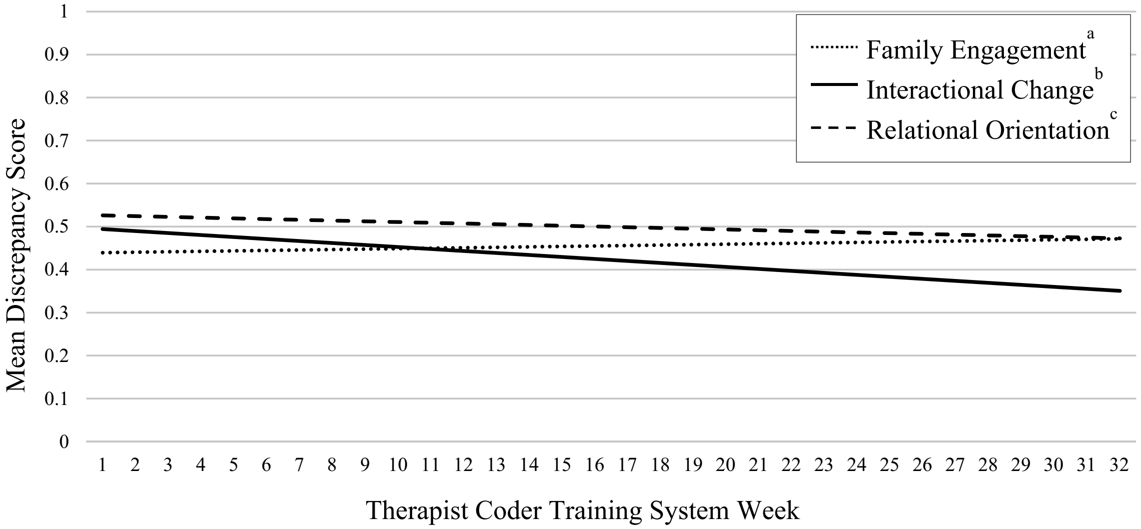

Fig. 1 Mean Discrepancy Score trend line by family therapy subgroup averaged across therapists. *Note*: Negative slope indicates a decrease in discrepancy score over time. [a]Average of the following items: Parent Collaboration, Love and Commitment, Parent Ecosystem, Adolescent Goal Collaboration. [b]Average of the following items: Prepare for Future Interactions, Stimulate Dialogue, Coach and Process, Teach Family Skills. [c]Average of the following items: Relational Focus, Focus on Process, Reframe, Relational Reframe, Family-Focused Rationale

Discussion

This study produced mixed findings in support of an online therapist coder training system in core techniques of family therapy for adolescent conduct and substance use problems. Therapist reliability in rating the presence or absence of techniques (i.e., technique recognition) remained primarily in the Moderate range across 32 exercise weeks of training. However, their reliability in rating the extensiveness of technique delivery (i.e., technique judgment) improved sharply, lingering in the Poor range for 24 weeks before climbing to the Good range in the final 8 weeks. With regard to coding accuracy, therapist sensitivity in identifying when techniques were present was robust for all items, whereas specificity in identifying absent techniques was highly variable across items. With regard to training therapists to recognize the extent to which given treatment techniques are used, a false positive, which results in lower sensitivity values, is arguably the graver error. Also, whereas therapists overestimated the mean extensiveness (i.e., dose) levels for almost every technique, their tendency to give low-accuracy scores (differing by more than two points from gold-standard scores) decreased steadily over time. The training did not produce growth in accuracy for the averaged set of all techniques, though therapists did show increased accuracy for one subset of core techniques, Interactional Change, which is considered by many to be the *sine qua non* of family therapy (Minuchin & Fishman, 1981).

A bright spot in study findings is the sharp increase documented for reliability in coding the extensiveness of observed techniques in the final quarter of training. Therapists progressed from having literally zero reliability during the first sixteen exercise weeks, to having merely poor reliability (ICC = .26) during the next 8 weeks, to achieving functional reliability (ICC = .64) during the final 8 weeks. A ruler that unpredictably changes length at each measurement occasion is a useless tool; by the conclusion of training, study therapists who completed training graduated to being potentially useful rulers. Therapists who learn to reliably judge the extensiveness of specific treatment techniques might also gain an edge in learning to deliver EBIs with fidelity and learning to self-report on EBI use in their own practices, two main pathways for enhancing quality of care (Stirman, 2020)—though links between EBI coding acuity and EBI utilization have yet to be tested. Note that results did not show uniform gains in reliability: The index for simply recognizing the presence of specific techniques (i.e., all-or-nothing agreement, versus magnitude of agreement; Hallgren, 2012) was steadfastly modest throughout training. Also, therapists were poorly reliable across training in judging almost one-third of the techniques, displaying wide variability in item-level ratings similar to the spread reported in other therapist coder studies (e.g., Caron & Dozier, 2019).

Most disappointing about study results was the lack of growth in therapist accuracy across the full set of techniques. As reported consistently (though not universally) in research on therapist self-ratings of EBI delivery (e.g., Caron et al., 2020; Martino et al., 2009), the current sample significantly overestimated the amount of family therapy techniques on display in vignettes, and their aim did not draw closer to the target as training progressed. It has been suggested that if this EBI overestimation bias proves resistant to correction despite efforts at remedial training—such as the current online course—the pragmatic course of action would be to develop inflation indices to adjust automatically for fixed biases in judgment (e.g., Hogue et al., 2015); of course for such corrections to be useful, therapist reliability needs to be robust enough such that the accuracy biases are known and (relatively) stable. Yet there were bright spots in accuracy effects as well. Therapists showed a steady decline in low-accuracy scores, dropping from 39% at beginning to 11% at end, meaning they were well-off-the-mark less and less often. These findings are congruent with observed increases in reliability. Taken together, the overall reliability and accuracy trends suggest that coder training can help therapists be both more consistent and closer to the mark in recognizing family therapy techniques, though with seemingly persistent bias toward overestimating how much they observe for most techniques.

Another bright spot for accuracy: Therapists showed significant accuracy gains in rating Interactional Change techniques. Therapists using these techniques act as coaches on the floor as they explicitly invite shifts in family behaviors that are observationally evident. Such "overtness" characteristics may render these techniques relatively more accessible to therapists as they gain coding experience, in contrast to Family Engagement and Relational Orientation techniques that target often implicit relational states and beliefs of family members (see Grotevant & Carlson, 1987). To improve therapist accuracy in family therapy across the board, the next version of the online training system may need to attend deliberately to strategies for coding interventions linked to more implicit relational and attitudinal changes (e.g., Moran et al., 2005).

There were two main therapist-level predictors of coding accuracy: baseline digital literacy (medium effect), and over-time submission of therapist-report checklists on use of the 13 techniques in their own therapy sessions (small effect). Both of these predictors are logically connected to participation in a technology-based online training and, importantly, both are highly malleable. Therapist digital literacy is certain to increase considerably as telehealth transitions to becoming a first-line intervention modality in the COVID-19 era and beyond (US Dept Health and Human Services, 2020). Mandates for therapists to document use of specific EBIs for given disorders are equally likely to strengthen as regional and national regulations for quality improvement and value-based purchasing expand (Damberg et al., 2014; Stirman, 2020) and as providers endeavor to replicate EBI documentation methods in practice contexts that have proven beneficial in training contexts.

Strengths and Limitations

The training system was designed to be pragmatic, featuring an online delivery platform that was asynchronous (i.e., could be accessed at user convenience) and required only 15–20 min per week to complete an exercise. Training uptake was at least moderately successful: Discounting those who left clinic employment prior to course activation, 72% of therapists from the parent trial completed at least 3 training exercises, and almost three-quarters of this enrolled group completed at least 17 of 32 available exercises. Study results are generalizable in some respects to therapists recruited in the parent trial: There were no differences in baseline characteristics between therapists in the parent trial who were included in the current study versus those who were not. Also, two-part sample diagnostics contraindicated the possibility that therapist reliability increased mainly because superior coders participated longer in training: There were no baseline differences between training completers (those who completed at least one exercise during the final training interval) versus training dropouts (those who did not); and the subsample of training completers showed a pattern of reliability gains that was nearly identical to that showed by the full sample. Thus, the sharp increase in interrater reliability observed during the final 8 weeks seems more likely due to bona fide training effects than to characteristics of the therapists who managed to finish training. Testing strategies to boost workforce training participation remains a top priority in EBI implementation science (Jensen-Doss et al., 2020). It was not an aim of this study to isolate variance components for reliability or accuracy data; if training effects are robust in future studies, it would be valuable to learn which therapist and vignette characteristics account for what proportions of observed variance. Patterns in the raw data for overall discrepancy scores suggested that we test for linear change over time, rather than equally plausible quadratic (e.g., discrepancies decrease and then stabilize over time) or loglinear (e.g., rapid decrease early in training that fades over time) change.

The study sample fairly represented the workforce demographics of participating sites, meaning that sex (88% female) and race/ethnicity (79% White Non-Latinx) characteristics were relatively homogenous. Because coding activities were completed by study participants in community settings rather than by research staff in controlled settings, we applied more lenient interrater reliability standards recommended by Cicchetti (1994) for ICC and Landis and Koch (1977) for kappa; had we instead applied more stringent standards (see Hallgren, 2012; Koo & Li, 2016), observed therapist reliability levels would be deemed somewhat less favorable. To minimize participant burden, the training system required that therapists code only 5 of the possible 13 treatment techniques during each exercise, thereby creating a high volume of item-level "missing data" for each exercise; as a result, we did not have sufficient power to calculate therapist reliability at weekly intervals and instead collapsed reliability data into four 8-week intervals. This prevented us from subjecting the reliability data (which are based on coder score variance) to the more informative multilevel modeling techniques that were used for the accuracy data (which are based on raw discrepancy scores). Finally, it was beyond the scope of this study to test for predictive effects of the training system, that is, whether coder training can yield ultimate clinical benefit by prompting increased delivery of the focal techniques.

Potential Next Steps in Therapist Coder Training

It remains an unfortunate truth that methods for training therapists to implement EBIs in usual care have not produced substantial impacts on EBI adoption, delivery, or quality (Beidas et al., 2019; Valenstein-Mah et al., 2020). Perhaps teaching therapists to be more reliable and accurate coders of EBI techniques can yield benefits that are cognitive (e.g., increased knowledge of EBI techniques and when they can be used), attitudinal (e.g., stronger confidence and motivation to use EBIs), and/or behavioral (e.g., enhanced capability to deliver EBIs via observational modeling and instrumental learning) in nature (McLeod et al., 2018). As discussed above, these benefits could manifest in increased EBI self-report acumen and/or increased EBI use in daily practice—though this downstream transferal of training effects has not been studied to date. Yet, therapist coder training effects may need to improve from those reported here to avail meaningful impact on EBI implementation. Future research on the current or similar training systems can examine other EBI approaches, vary training load with regard to session length and frequency, and vary training parameters with regard to number of techniques captured in vignettes and scored by coders and/or inclusion of coding aids to scaffold rating concordance as training progresses. Another promising approach is training therapists to code their own treatment sessions, rather than mock session vignettes, as part of a standardized EBI fidelity feedback process; this intuitive method has not been systematically investigated (but see Caron & Dozier, 2019), and it remains to be seen whether "participant coding" of this kind would introduce strong biases and/or be an amenable practice in the general workforce. Also, it will be important to test whether therapists need to have already acquired a certain level (i.e., "critical mass") of therapeutic proficiency, in general and/or in the specific treatment approach being trained, in order for coder training gains to translate consistently into enhanced treatment delivery.

Acknowledgements Preparation of this article was supported by the National Institute on Drug Abuse (R34DA044740; PI: Hogue). The authors acknowledge the contributions of our dedicated colleagues Cheryl Ching, Guy Diamond, Frances Ferayorni, Genoveva Garcia, Cori Hammond, Suzanne Levy, Bryce McLeod, and Michael Southam-Gerow.

References

Author. (2019a). 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.

Author. (2019b). 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*(25), 1–12.

Author. (2020). 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*. https://doi.org/10.1186/s13063-019-3783-8

Baldwin, S. A., Christian, S., Berkeljon, A., & Shadish, W. R. (2012). The effects of family therapies for adolescent delinquency and substance abuse: A meta-analysis. *Journal of Marital and Family Therapy, 38*, 281–304.

Becker, E. M., & Jensen-Doss, A. (2014). Therapist attitudes towards computer-based trainings. *Administration and Policy in Mental Health and Mental Health Services Research, 41*, 845–854.

Beidas, R. S., 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.

Beidas, R. S., & Kendall, P. C. (2010). Training therapists in evidence-based practice: A critical review of studies from a systems-contextual perspective. *Clinical Psychology: Science and Practice, 17*, 1–30.

Beidas, R. S., Williams, N. J., Becker-Haimes, E. M., Aarons, G. A., Barg, F. K., Evans, A. C., Jackson, K., Jones, D., Hadley, T., Hoagwood, K., & Marcus, S. C. (2019). A repeated cross-sectional study of clinicians' use of psychotherapy techniques during 5 years of a system-wide effort to implement evidence-based practices in Philadelphia. *Implementation Science, 14*, 67.

Caron, E. B., & 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*, 445–457.

Caron, E. B., Muggeo, M. A., Souer, H. R., Pella, J. E., & Ginsburg, G. S. (2020). Concordance between clinician, supervisor and observer ratings of therapeutic competence in CBT and treatment as usual: Does clinician competence or supervisor session observation improve agreement? *Behavioural and Cognitive Psychotherapy, 48*, 350–363.

Carroll, K. M., Nich, C., Sifry, R., Nuro, K. F., Frankforter, T. L., Ball, S. A., Fenton, L., & Rounsaville, B. J. (2000). A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug & Alcohol Dependence, 57*, 225–238.

Chorpita, B. F., & Daleiden, E. L. (2009). Mapping evidence-based treatments for children and adolescents: Application of the distillation and matching model to 615 treatments from 322 randomized trials. *Journal of Consulting and Clinical Psychology, 77*, 566–579.

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

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Damberg, C. L., Sorbero, M. E., Lovejoy, S. L., Martsolf, G. R., Raaen, L., & Mandel, D. (2014). Measuring success in health care value-based purchasing programs: Findings from an environmental scan, literature review, and expert panel discussions. *Rand Health Quarterly, 4*(3), 9.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*, 319–340.

Diggle, P., Heagerty, P., Liang, K., & Zeger, S. (2002). *Analysis of longitudinal data (2nd edition)*. Oxford University Press.

Ehrenreich-May, J., Dimeff, L. A., Woodcock, E. A., Queen, A. H., Kelly, T., Contreras, I. S., Rash, B., Kelley-Brimer, A., Hauschildt, J., Danner, S. M., & Kennedy, S. M. (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.

Gonder, J., Metlay, W., & Shapiro, T. (2018). Testing assumptions: Can performance rating feedback result in objective performance improvements? *Journal of Management and Innovation*. https://doi.org/10.18059/jmi.v4i2.100

Grotevant, H. D., & Carlson, C. I. (1987). Family interaction coding systems: A descriptive review. *Family Process, 26*, 49–74.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*, 23.

Henderson, C. E., Hogue, A., & Dauber, S. (2019). Family therapy techniques and one-year clinical outcomes among adolescents in usual care for behavior problems. *Journal of Consulting and Clinical Psychology, 87*, 308–312.

Hogue, A., Dauber, S., Chinchilla, P., Fried, A., Henderson, C. E., Inclan, J., Reiner, R., & Liddle, H. A. (2008). Assessing fidelity in individual and family therapy for adolescent substance abuse. *Journal of Substance Abuse Treatment, 35*, 137–147.

Hogue, A., Dauber, S., & Henderson, C. E. (2014). Therapist self-report of evidence-based practices in usual care for adolescent behavior problems: Factor and construct validity. *Administration and Policy in Mental Health and Mental Health Services Research, 41*, 126–139.

Hogue, A., Dauber, S., & Henderson, C. E. (2017). Benchmarking family therapy for adolescent behavior problems in usual care: Fidelity, outcomes, and therapist performance differences. *Administration and Policy in Mental Health and Mental Health Services Research, 44*, 626–641.

Hogue, A., Dauber, S., Henderson, C. E., Bobek, M., Johnson, C., Lichvar, E., & Morgenstern, J. (2015). Randomized trial of family therapy versus non-family treatment for adolescent behavior problems in usual care. *Journal of Clinical Child and Adolescent Psychology, 44*, 954–969.

Hogue, A., Henderson, C. E., Becker, S. J., & Knight, D. K. (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.

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

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.

Jensen-Doss, A., Smith, A. M., Walsh, L. M., Ringle, V. M., Casline, E., Patel, Z., Shaw, A. M., Maxwell, C., Hanson, R., & 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.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155–163.

Kreft, I., & DeLeeuw, J. (1998). Introducing multilevel modeling. Sage Publications.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

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

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

McCart, M. R., & Sheidow, A. J. (2016). Evidence-based psychosocial treatments for adolescents with disruptive behavior. *Journal of Clinical Child & Adolescent Psychology, 45*, 529–563.

McLeod, B. D., Cox, J. R., Jensen-Doss, A., Herschell, A., Ehrenreich-May, J., & Wood, J. J. (2018). Proposing a mechanistic model of clinician training and consultation. *Clinical Psychology: Science and Practice, 25*(3), e12260.

Minuchin, S., & Fishman, H. C. (1981). *Family therapy techniques*. Harvard University Press.

Moran, G., Diamond, G. M., & Diamond, G. S. (2005). The relational reframe and parents' problem constructions in attachment-based family therapy. *Psychotherapy Research, 15*, 226–235.

Muthén, L. K., & Muthén, B. O. (1998–2021). *Mplus user's guide* (Seventh ed.). Muthen & Muthen.

Riedinger, V., Pinquart, M., & Teubert, D. (2017). Effects of systemic therapy on mental health of children and adolescents: A meta-analysis. *Journal of Clinical Child and Adolescent Psychology, 46*, 880–894.

Robbins, M., Feaster, D., Horigian, V., Puccinelli, M., Henderson, C. E., & Szapocznik, J. (2011). Therapist adherence in Brief Strategic Family Therapy for adolescent drug abusers. *Journal of Consulting and Clinical Psychology, 79*, 43–53.

Stirman, S. W. (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.

Stirman, S. W., Bhar, S. S., Spokas, M., Brown, G. K., Creed, T. A., Perivoliotis, D., Farabaugh, D. T., Grant, P. M., & Beck, A. T. (2010). Training and consultation in evidence-based psychosocial treatments in public mental health settings: The access model. *Professional Psychology: Research and Practice, 41*, 48.

Tanner-Smith, E. E., Wilson, S. J., & Lipsey, M. W. (2013). The comparative effectiveness of outpatient treatment for adolescent substance abuse: A meta-analysis. *Journal of Substance Abuse Treatment, 44*, 145–158.

U.S. Department of Health and Human Services. (2020). Telehealth: Delivering care safely during COVID-19. Website: https://www.hhs.gov/coronavirus/telehealth/index.html

Valenstein-Mah, H., Greer, N., McKenzie, L., Hansen, L., Strom, T. Q., Wiltsey Stirman, S., Wilt, T. J., & Kehle-Forbes, S. M. (2020). Effectiveness of training methods for delivery of evidence-based psychotherapies: A systematic review. *Implementation Science, 15*, 1–17.

Weingardt, K. R. (2004). The role of instructional design and technology in the dissemination of empirically supported, manual-based therapies. *Clinical Psychology: Science and Practice, 11*, 313–331.

Weingardt, K. R., Cucciare, M. A., Bellotti, C., & Lai, W. P. (2009). A randomized trial comparing two models of web-based training in cognitive–behavioral therapy for substance abuse counselors. *Journal of Substance Abuse Treatment, 37*, 219–227.

Weisz, J. R., 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.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.