Published in final edited form as:
J Clin Child Adolesc Psychol. 2019 ; 48(1): 16–28. doi:10.1080/15374416.2018.1561295.

Effectiveness and Efficiency of Observationally Assessing Fidelity to a Family-Centered Child Intervention: A Quasi-Experimental Study

Justin D. Smith,
Northwestern University Feinberg School of Medicine

Jenna Rudo-Stern,
REACH Institute, Arizona State University

Thomas J. Dishion,
REACH Institute, Arizona State University & Oregon Research Institute

Elizabeth A. Stormshak,
Prevention Science Institute University of Oregon

Samantha Montag,
Northwestern University Feinberg School of Medicine

Kimbree Brown,
Oregon Social Learning Center

Karina Ramos,
University of California Irvine Counseling Center

Daniel S. Shaw, and
University of Pittsburgh

Melvin N. Wilson
University of Virginia

Abstract

Objective.—Assessment of fidelity that is effective, efficient, and differentiates from usual practices is critical for effectively implementing evidence-based programs for families. This quasi-experiemntal study sought to determine whether observational ratings of fidelity to the Family Check-Up (FCU) could differentiate between levels of clinician training in the model, and from services as usual, and whether rating segments of sessions could be equivalent to rating complete sessions.

Method.—Coders rated 75 videotaped sessions—complete and 20-minute segments—for fidelity, using a valid and reliable rating system across three groups: (1) highly trained in FCU with universal, routine monitoring; (2) minimally trained in FCU with optional, variable monitoring;

Address correspondence to Justin D. Smith, Center for Prevention Implementation Methodology, Department of Psychiatry and Behavioral Sciences, Department of Preventive Medicine, Department of Pediatrics, Northwestern University Feinberg School of Medicine, 750 N Lake Shore Drive, Chicago, IL 60657. jd.smith@northwestern.edu.

and (3) services as usual with no training in the FCU. We hypothesized that certain dimensions of fidelity would differ by training, while others would not.

Results.—The results indicated that, as expected, one dimension of fidelity to the FCU, Conceptually accurate to the FCU, was reliably different between the groups ($\chi^2 = 44.63$, $p<0.001$). The differences observed were in the expected direction, showing higher scores for therapists with more training. The rating magnitude of session segments largely did not differ from those of complete session ratings; however, reliabilities were low for the segments.

Conclusions.—Although observational ratings were shown to be sensitive to the degree of training in the FCU on a unique and theoretically critical dimension, observational coding of complete sessions is resource-intensive and limits scalability. Additional work is needed to reduce the burden of assessing fidelity to family-centered programs.

Keywords

competent adherence; Family Check–Up; fidelity; implementation; observational coding

One of the greatest challenges implementers encounter when taking evidence-based programs (EBPs) to community settings is maintaining fidelity to the protocol (McHugh & Barlow, 2010). It is estimated that about 10% of the EBPs delivered in settings that serve children and families are done so with the fidelity intended by the original program developer (Biglan, 2015). Fidelity itself can be a primary indicator of implementation success when it is well defined, is deemed to meet or exceed minimal standards, and is linked to program outcomes (Landsverk et al., 2012). Because of the scarcity of resources in community service delivery systems, safeguards need to be in place to reduce the likelihood of implementation drift and the associated waning of the potential benefits of the EBP. Valid, reliable, and feasible systems for assessing fidelity are essential. There are two interrelated issues when considering the viability of a fidelity-rating system: effectiveness and efficiency (Schoenwald et al., 2011). If the scale up of EBPs is to be successful, feasible and effective assessment and monitoring strategies are needed to ensure interventions can be delivered with fidelity across settings and skill level and experience of practitioners (Perepletchikova, Treat, & Kazdin, 2007).

Effectiveness.

Effectiveness refers to ratings of fidelity that predict meaningful clinical and implementation outcomes, such as individual client outcomes, therapeutic processes, and sustainability (Berkel, Mauricio, Schoenfelder, & Sandler, 2011). Effectiveness ratings also should differentiate between high and low fidelity within delivery of the EBP and from other services. Thus, fidelity-rating systems ought to distinguish among differing levels of skill in delivering the EBP, so as to demonstrate attainment of minimum standards and to identify providers in need of remediation. There is strong empirical evidence indicating that fidelity to EPBs is linked to level of training and that training alone, without ongoing monitoring, is not sufficient to maintain loyal delivery (see Dusenbury, Brannigan, Falco, & Hansen, 2003 for a review). In a randomized study comparing the relations between fidelity and three different levels of training, Sholomskas et al. (2005) found incremental and statistically significantly higher ratings of fidelity attributable to the level of training received.

A related challenge when assessing fidelity in community-based trials is that the EBP is often compared to services as usual, which may or may not be evidence-based. Therefore, fidelity-rating systems need to be sufficiently sensitive to differentiate not only a clinician's expertise within a specific EBP protocol but also their behaviors specific to the target EBP from those prescribed to other EBP protocols (and thus proscribed to the target EBP). Differentiation is a core aspect of fidelity and refers to the extent to which an intervention(s) under study differ along appropriate lines and whether (and "to where") deviations from the protocol occur (Southam-Gerow & McLeod, 2013). In theory, it is possible for two EBPs (or an EBP and usual practice) to be discriminable but also to have significant overlap in therapist behaviors. Unfortunately, differentiation is typically overlooked even when EBPs are being implemented because few community settings adequately assess fidelity (Garland, Hurlburt, & Hawley, 2006).

Efficiency.

Efficiency refers to assessment procedures that are of minimal cost and low burden on clinicians, supervisors, and the service delivery system. Assessing and monitoring implementation fidelity is impeded by a multitude of factors in community settings, including a lack of training in EBPs, delivery of eclectic approaches, clients that differ from those in efficacy trials, limited expertise with assessing and monitoring fidelity to EBP protocols, and limited resources (Hanson et al., 2013). Fidelity assessment systems are considered optimal when based on direct observation (Gearing et al., 2011); however, because of the burden associated with carrying out assessments of fidelity, few studies have carried out observationally-based studies, particularly in low-resource community service settings. Hence, because of the burden of observational assessment, few studies have examined ways to increase efficiency and none have been done with parent training interventions specifically.

The most common hypothesized remedy is to rate portions of sessions rather than their entirety. In a pair of studies, Weck et al. (Weck, Bohn, Ginzburg, & Stangier, 2011; Weck, Grikscheit, Höfling, & Stangier, 2014) compared ratings of adherence and competence from complete sessions versus session segments (i.e., middle third of a session). In both studies, ratings from both groups were highly correlated with one another and with intervention outcomes; acceptable reliabilities were found. Despite some differences in relations between fidelity and outcome by type of client being treated and some lower reliabilities for the ratings of segments on global and specific dimensions of fidelity compared to entire sessions, the authors concluded that rating segments was an adequate, albeit nonequivalent, alternative.

EBPs in the Community

There is a nearly universal emphasis on EBPs in psychology and mental health services. Thus, it could be difficult to differentiate between a specific EBP protocol and services as usual, Research on community practice has indeed found that “usual care” contains some elements of EBPs, often at a low dose (intensity), but that there is wide variation (Garland et al., 2010). However, the different treatment situations are a germane empirical and practical issue based on the need to demonstrate that differences in intervention effects between an EBP and services as usual are due to the EBP and not to common factors or other evidence-based practice elements. Relatedly, when evaluating the effects of a community-based trial, a non-significant effect between the EBP and services as usual could be attributable to poor fidelity to the EBP protocol or due to the EBP being indistinct from the services delivered in the comparison arm.

Adding to the challenge is that well-established EBPs are more likely to be taught in training programs and through continuing education and then translated to everyday practice. Relevant to this study, for example, are the broad class of parent training interventions and Motivational Interviewing (Miller & Rollnick, 2012). The empirical literature concerning the quality of EBP delivery in the community is sparse (Hoagwood & Kolko, 2009; Weisz, Jensen-Doss, & Hawley, 2006). When delivery of an EBP was evaluated in both research and community settings, there was a high degree of similarity but a lower dose of EBP elements and levels were more likely to wane over time in the community (M. Smith et al., 2017). Relatedly, adherence and competence ratings were lower for community therapists compared to therapists in research settings even with the same training and supervision protocols for both groups (McLeod et al., 2017). While clinicians and program directors tend to report moderate to high use of EBPs in standard practice, Santa Ana et al.’s (2008) observations of community-based mental health services indicated low use of EBPs. It will become increasingly important to develop fidelity-rating systems that can effectively distinguish nuanced skills in an EBP protocol from services as usual, especially as interventions delivered by mental health professionals in the real world more closely resemble common EBPs and their component elements.

Fidelity to Evidence-Based Parent Training Interventions

Research indicates that outcomes of parent training interventions consistently vary as a function of fidelity (e.g., Chiapa et al., 2015; Forgatch, Patterson, & DeGarmo, 2005; Hogue & Liddle, 2009; Smith, Dishion, Shaw, & Wilson, 2013). There are currently efforts to implement evidence-based family interventions on a wide scale in diverse contexts, such as schools (Smolkowski et al., 2017), social services, community mental health (Dishion, Forgatch, Chamberlain, & Pelham III, 2016), and primary care (Leslie et al., 2016). The complexity of community settings attenuates fidelity and in turn inhibits the positive outcomes of EBPs for children and families (Weisz et al., 2006). Given that EBPs tend to be more effective with youth and families than services as usual (e.g., Dulcan, 2000), greater attention must be given to the evaluation of fidelity, and to the fidelity assessment tools themselves, to ensure preservation of the behavior change mechanisms that make them effective. Evaluation and measurement of fidelity to family therapy for youth externalizing and substance use is well represented (see Hogue et al., 2017). Unfortunately, few trials of *parent training* interventions adequately measure fidelity, leaving a dearth of validated measures (Perepletchikova et al., 2007; Weisz, Doss, & Hawley, 2005). Notable outliers in the parent training literature are the Oregon model, which has decades of research on community translation rooted in the maintenance of fidelity to the program (Dishion, Forgatch, Chamberlain, & Pelham III, 2016), and the Family Check-Up (FCU), which has a valid and reliable observational measure (Smith, Dishion, Shaw, & Wilson, 2013).

The Current Study

The assessment of fidelity to the FCU is examined. The FCU is a brief, assessment driven, and family-centered intervention that uses motivational interviewing to improve engagement and enhance motivation to change parenting. Two published studies support a relation between ratings of fidelity to the FCU protocol and outcomes (Chiapa et al., 2015; Smith et al., 2013). Both papers reported on a subsample of 79 families who had toddler-age children in the clinical range of caregiver-reported problem behaviors at entry into a randomized trial of the FCU for ethnically diverse, indigent families (Dishion et al., 2008). Fidelity to the FCU was rated using the COACH observational rating system (Dishion, Smith, Gill, Shaw, & Knutson, 2014), which assesses five dimensions of observable therapist skill prescribed to the FCU: Conceptual accuracy to the FCU; Observant and responsive to client needs; Actively structures sessions; Careful and appropriate teaching; Hope and motivation are generated. Detailed information about each dimension can be found in Smith et al. (2013). Families randomized to the intervention arm were offered the FCU each year in a health maintenance framework (Dishion, Brennan, et al., 2014; Smith, Berkel, et al., 2016). Smith et al. (2013) found a relation between ratings of fidelity on the COACH to the FCU feedback session, at child age 2, and changes in observed positive behavior support practices of caregivers one year later (age 3), which was in turn predictive of caregiver-reported problem behaviors assessed at age 4. The effect of fidelity on child behavior occurred through observational ratings of caregiver in-session engagement, which was positively associated with ratings of fidelity and parenting practices. In a follow-up study with the same sample of 79 families, latent growth modeling of fidelity ratings over the first four years of the trial (age 2 to 5 years) indicated that variation in the trajectory was significantly related to caregiver and teacher reports of child problem behaviors assessed at ages 7.5/8.5 (Chiapa et al., 2015).

The aim of the current quasi-experimental study was to evaluate multiple aspects of fidelity assessment germane to translating EBPs to the community. First, the effectiveness of the COACH fidelity-rating system was evaluated. Hypothesis 1: Ratings could reliably distinguish between groups of clinicians who delivered FCU with differing levels of training, fidelity monitoring, and consultation, and clinicians who had received no training in the FCU and delivered services as usual. Hypothesis 2: Two dimensions of the COACH rating system would differ significantly between conditions—*Conceptual accuracy to the FCU*, which is both unique and essential to this program, and *Hope and motivation*, which is also central to effective delivery of FCU and is emphasized during training—and the conditions would not differ significantly on the other three dimensions of the COACH rating system, as these dimensions are necessary but not specific to the FCU. The *Hope and motivation* dimension of the COACH captures motivational interviewing skills, which are important as the Drinker's Check-Up (Miller, Sovereign, & Krege, 1988)—the precursor to contemporary motivational interviewing—is the basis of FCU. Hypothesis 3: Concerning the efficiency of observational ratings, we hypothesized that fidelity scores based on review of a 20-minute segment of a session would be equivalent in magnitude to ratings of the complete session and that they would similarly distinguish between the level of training conditions.

To test these related hypotheses, 75 family intervention sessions that had not previously been rated for fidelity were randomly drawn from archival datasets and observationally rated using the COACH. Twenty-five sessions from each of three conditions were selected: (1) highly trained therapists with universal and routine monitoring of fidelity, (2) minimally trained therapists with optional and variable fidelity monitoring, and (3) therapists with no training in the FCU and delivering services as usual. An attempt was made to select sessions across conditions with similar child characteristics (see section titled “Selection of FCU sessions”). Therapist characteristics across conditions were largely similar in that the majority of therapists were master’s level clinicians. Unfortunately, more detailed information about them was not available. Sessions ranged in length, from about 40 minutes to about 75 minutes in this sample. To address the issue of efficiency, a 20-minute segment from all sessions was also coded. An additional consideration became evident during the study. Coders reported difficulty rating the session segments with confidence because, inevitably, some key therapist skills did not occur during the 20-minute segment that was rated. Because of coders’ difficulty, we hypothesized that interrater reliability of fidelity ratings for 20-minute segments would be significantly lower compared to those from complete sessions (Hypothesis 4). Based on our findings, a post hoc reanalysis of two published studies (Chiapa et al., 2015; Smith et al., 2013) is reported to support the conclusions.

Method

Overview of Study Design

The conditions in this study were derived from two completed randomized trials of the FCU. Trial 1, an efficacy trial of the FCU, provided data for the highly trained condition. Trial 2, an effectiveness trial of the FCU conducted in community mental health agencies, was used for the FCU with minimal training, monitoring, and consultation and for the no training (services as usual) conditions. The different FCU trials allowed us to leverage existing data to conduct a study that would be challenging otherwise. Experts have called for such designs to make mental health services and implementation research more efficient (Chambers, Wang, & Insel, 2010).

Participants

Trial 1.—For the highly trained/universal and routine monitoring of fidelity condition, sessions were drawn from a randomized efficacy trial of the FCU for indigent families with young children. Mothers with a 2-year-old child were recruited from the Women, Infants, and Children Nutritional Supplement Program in three geographically diverse regions in the United States (Charlottesville, VA; Eugene, OR; Pittsburgh, PA) and were randomly assigned to either the intervention or services-as-usual. In the intervention arm, families were offered the FCU each year up to child age 10.5, with a total of eight opportunities for services (see REMOVED FOR MASKED REVIEW]). Sessions were delivered in the family home. The sample was culturally diverse and included European American (71%), African

American (14%), Hispanic/Latino (5%), and multiple ethnicities (10%). Children selected for inclusion in the current study (see section titled "Selection of sessions") had a mean age of 9.79 (*SD* = 1.50) years and were 43% female. The therapists had a master's or doctoral degree and received intensive front-end training along with weekly group supervision and cross-site supervision. For complete trial procedures see, REMOVED FOR MASKED REVIEW].

Trial 2.—The minimally trained/optional and variable fidelity monitoring and no training/services-as-usual conditions were drawn from an effectiveness trial of the FCU where 40 master's level therapists in community mental health centers were randomized either to be trained in the FCU at the beginning of the trial (intervention condition) or to be in the services-as-usual arm and receive training in FCU at the end. The therapists were independently licensed in marriage and family therapy or social work. Families included children age 5 to 17 years (M = 11.82, *SD* = 2.13, 49% female) seeking services for a variety of mental health concerns. The ethnic backgrounds of the children in the FCU/services-as-usual subsamples used in the current study were European American (65%/65%), African American (13%/17%), Hispanic/Latino (0%/4%), Native American/American Indian/Alaska Native (4%/4%), and multiple (17%/9%). The complete trial procedures can be found in REMOVED FOR MASKED REVIEW].

Procedures

Selection of sessions.—The child characteristics of age, gender, racial/ethnic background, and caregiver-reported problem behaviors[1] were considered in the selection of sessions to rate so as to reduce variation in fidelity not attributable to the study condition. Due to the small sample available from Trial 2, which had a total of 73 families but only 33 that completed an FCU feedback session and only 26 families who received at least 2 sessions of services as usual, we used this trial as the basis for session selection. It was important for there to be at least 2 sessions of services as usual as this corresponds to when the feedback occurs in FCU and comparison to an initial session or a later session could introduce bias. Thus, session 2 was coded when available and session 3 was used when no videotape of session 2 was available (n = 2). First, we selected the 25 cases from Trial 2 with caregiver reports of elevated child conduct problems. (The remaining 8 families from Trial 2 were not included in any of the results presented). This criterion was included to align with previous research on FCU fidelity ratings (Smith et al., 2013). Next, we identified a pool of potential sessions available from Trial 1. To best match the demographics of families in Trial 2, which occurred in the greater Portland, OR metropolitan area, we limited our pool from Trial 1 to families recruited in the Eugene, OR area (rather than including families from Pittsburgh, PA and Charlottesville, VA), limited the ages from 5 to 10.5 years, and oversampled from the last wave (age 10.5) of Trial 1 so as to approach the mean age of children in Trial 2. A data manager who did not know the hypotheses of this study randomly

[1] The two trials administered different caregiver reports of child problem behaviors. Study 1 administered the Child Behavior Checklist (Atkins, Steyvers, Imel, & Smyth, 2014), a multi-scale questionnaire used to assess behavioral problems in youth ages 1.5 to 18. Study 2 used the 5-item conduct problems subscale of the Strengths and Difficulties Questionnaire (Gallo et al., 2014). Five items from the CBCL were selected that matched those of the SDQ (i.e., fighting, lying, stealing, noncompliance, losing one's temper). Internal consistencies were acceptable for the SDQ (α = .80) and the 5-items of the CBCL (α = .77).

selected 25 cases from Trial 1 that were similar in age, race/ethnicity, and gender to those families drawn from Trial 2 and that also had equivalently high parent reports of child conduct problems. Next, the selected sessions were rated using the COACH. For FCU conditions, the feedback session was rated. For the services-as-usual condition, the second or third session was rated to align the timeframes.

Coders and coder training.—Three coders (two graduate students in a psychology doctoral program, one Bachelor’s level staff) were used. Each had been previously trained (approximately 20 hours) to reliability in the COACH and had rated sessions using the system for at least one year before beginning this study. Coders had each rated at least 50 complete FCU feedback sessions prior to this study. Raters attended biweekly meetings during this study to maintain reliability and minimize rater drift. Raters were masked to the study hypotheses, including that there was even a condition in which the therapists had no training in FCU to reduce potential bias in fidelity ratings.

Coding procedures.—Coding procedures followed those established in a study by Smith, Dishion, et al. (2016), where it was determined that reliability is significantly improved when raters review the results of the ecological assessment beforehand. Assignment of sessions to rate followed a multistep strategy to ensure that different coders rated the complete and segment conditions as well as the sessions selected for double coding to calculate reliability. First, each of the 75 sessions was randomly assigned to one of the three coders for coding of the complete session. Second, the sessions were randomly assigned to a different coder to rate the segment. We controlled for coder assignment in this randomization so that each coder rated equal numbers of sessions in each of the three conditions. Third, 20% of the sessions in each condition, evenly distributed, were randomly selected for double coding to calculate reliability. The reliability coder was always the remaining coder who was not assigned to code the complete or segment of that session to reduce potential confounding. For consistency in the coding of segments, raters coded 20 minutes of the session between minutes 10 and 30. This time segment was selected in part because the total length of the sessions varied with the low end of the range being about 40 minutes. The order of coding was randomized within rater (i.e., the sequence of the sessions any given rater reviewed was random between complete sessions and segments and across conditions).

Measures

Fidelity.—Clinician’s fidelity to the FCU feedback session protocol was assessed using the COACH rating system (Dishion, Smith, et al., 2014). The COACH assesses adherence to and competent execution of the core dimensions of the FCU to arrive at a single metric referred to as *competent adherence*. Competent adherence has been found to be predictive of the effects of multiple parent training programs on child and family outcomes (Forgatch et al., 2005). The FCU is theory-based, meaning that rigid adherence to a manual is neither necessary nor desired, as long as the core components of the model are tailored to the needs of each family. Thus, evaluating fidelity to the FCU requires rating both delivery of the content and the process.

The COACH assesses five dimensions of observable therapist skill in the FCU: Conceptually accurate to the FCU; Observant and responsive to client needs; Actively structures sessions; Careful and appropriate teaching; Hope and motivation. The five dimensions are rated separately on a 9-point scale: 1–3 (*needs work*), 4–6 (*acceptable work*), 7–9 (*good work*). Reliability of the mean score of the five COACH dimensions has been acceptable (intraclass correlation coefficients [ICC] range from .67 to .77) in previously published studies (Chiapa et al., 2015; Smith, Dishion, et al., 2016; Smith et al., 2013; Smith, Stormshak, & Kavanagh, 2015).

Assessment of caregiver's in-session engagement occurs at the same time and is rated on the same scale but with appropriate anchors: 1–3 (*low*, caregiver is inattentive or disengaged), 4–6 (*medium,* modest signs of engagement), and 7–9 (*high,* caregiver actively participates and is attentive and responsive). ICCs for the caregiver engagement item have been fair to excellent (.59 to .87) in previous studies (see Smith et al., 2015).

Data Analysis

First, the mean and internal consistency of a mean score comprised of the five COACH dimensions were computed. Next reliability of the COACH (individual items and the mean score) were calculated using a one-way random effects model inter-rater correlation coefficient, or ICC(1,1). Cicchetti's (1994) interpretative guidelines will be used to describe reliability: *poor* (< .40), *fair* (.40–.59), *good* (.60–.74), *excellent* (≥ .75). Because the subscales of the COACH are not normally distributed, nonparametric statistical tests were used to compare differences in the subscales among the three conditions and two session length ratings. In addition to testing for differences among the three conditions, we also compared the two conditions where FCU was delivered ($n = 50$) to the services-as-usual condition ($n = 25$) to increase power and to provide a more relevant comparison for public health purposes (i.e., EBP vs. usual practice). Kruskal-Wallis tests were used for comparisons among three conditions (H1), Wilcoxon Rank-Sum tests were used for comparisons between two conditions (H1, H2), and Wilcoxon Signed-Rank tests were used for comparisons between the two session-rating-length conditions (H3). Tests were adjusted for multiple comparisons using the Bonferroni method. Finally, multivariate analysis of variance (MANOVA) was used to determine which dimensions of the COACH might account for differences among the three conditions, and Roy's Maximum Root test was used to determine the maximum F statistic for all linear combinations of COACH item-level ratings (H2). The Delta method (Rao, 2009) was used to estimate the standard error of the ICCs to perform a z test to determine the statistical significance (p value) of the difference between the ICCs of the complete session and segment conditions (H4). Analyses were conducted using SAS 9.4 (SAS Institute Inc., 2014) or R (R Core Team, 2012).

Results

Descriptive Statistics.—Correlations among study variables are provided in Table S1 in the Supplemental Materials. The COACH items were significantly intercorrelated (range: $r = .48–.78$). Conceptual accuracy and caregiver engagement ratings were correlated with study condition such that higher levels of training and fidelity monitoring associated with higher levels of fidelity and caregiver engagement. There were no significant correlations between the study variables and the segment vs. complete session conditions. Descriptives, reliabilities, and the internal consistencies of COACH ratings by study condition are provided in Table 1. Within condition, ICCs were in the good range for the complete session ratings (.67–.82), the poor to good range for ratings of segments (.34–.76) and had high internal consistency (α = .88–.97).

Preliminary analyses.—Omnibus tests (e.g., ANOVA, chi-square statistic) were conducted to demonstrate condition equivalence on key child characteristics. The conditions did not significantly differ by child gender, $\chi^2(2) = .046$, $p = .977$; child race/ethnicity $\chi^2(8) = 7.283$, $p = .506$; or parent reports of child conduct problems, $F(2, 74) = .085$, $p = .919$. However, child age significantly differed across groups, $F(2, 74) = 8.036$, $p = .001$. Tukey's post hoc test indicated the services-as-usual condition in Trial 2 had older children (Trial 1: M = 9.79, $SD = 1.50$; Trial 2 FCU: M = 12.12, $SD = 1.46$, Trial 2 services as usual: M = 11.54, $SD = 2.63$). We were unable to test for differences in therapist characteristics because the trials we drew from did not include it.

Primary analyses.—The chi-square value and *p*-value resulting from the Kruskal-Wallis tests of differences among the three conditions and the S scores and *p*-values resulting from Wilcoxon Rank-Sum Tests of differences between the session length conditions are reported in Table 2 (H1, H2). Differences were assessed for the complete session and segment ratings separately. Accordingly, a Bonferonni correction was applied separately (significance = *p*-value < 0.00125). Among the ratings of complete sessions, Conceptual accuracy was statistically different overall and between the pooled FCU conditions and the services-as-usual condition. This difference in the FCU conditions was also evident among the segment ratings. In addition, among the complete session ratings, a significant difference was found in ratings of caregiver engagement. However, the difference was only significant overall ($p <$ 0.00119[2]) and not between the pooled FCU conditions compared to no training ($p = 0.054$).

F scores and *p*-values of the Roy's Maximum Root Test resulting from the MANOVA analyses are reported in Table 3 (H2). Two models were created to test the hypotheses that (a) the Conceptual accuracy to the FCU and Hope and motivation would differ between the conditions (Model 1) and (b) there would be no differences between conditions on the other COACH dimension ratings (Model 2). Model 1 showed a statistically significant difference between the three conditions ($F = 17.37$, $p < 0.001$) and the pooled FCU to no training conditions ($F = 16.42$, $p < 0.001$). By examining the first eigenvector for each test, we found that the differences were almost exclusively due to Conceptual accuracy in both tests. Further, differences between the conditions appeared to be almost exclusively accounted for by lower scores in the services-as-usual/no training condition compared to the two FCU conditions; the services-as-usual condition has much lower scores on COACH dimensions compared to those of the other two conditions. There were no statistically significant differences found for Model 2.

S scores and *p*-values resulting from the Wilcoxon Signed-Rank Tests comparing the complete session and segment ratings are reported in Table 4 (H3). There were no significant

[2] P-values are given to the 5th decimal point in text because of the Bonferroni correction.

differences between ratings of the complete session and segment conditions among the minimal training and services-as-usual groups. However, caregiver engagement was significantly higher ($p = 0.004$) in the highly trained condition. Further, when all conditions were examined concurrently, we observed that Hope and motivation ($p = 0.013$) favored the FCU highly trained condition.

The Delta method was applied for H4 (complete results are in Table S2 in the Supplemental Materials). The results (z tests) indicated that within the minimal training condition, the COACH mean score ICC was significantly different between the complete and segment conditions, $z = 1.74$, $p = .041$. ICCs were not statistically different for the highly trained and services-as-usual condition, $z = .31$, $p = .379$ and $z = .59$, $p = .299$, respectively.

Post hoc Analysis.—To demonstrate that one of the primary findings of this study—Conceptually accurate to the FCU dimension meaningfully varies by training and skill level —we re-analyzed the results of the Smith et al. (2013) and Chiapa et al. (2015) studies, which both found that ratings on the COACH (from complete sessions) were significantly related to intervention outcomes. The re-analysis involved substituting the single item, Conceptually accurate, to the FCU score for the COACH mean score used in the original analyses and comparing the results. In both cases, the significance of the paths in the model were identical. For the indirect effect of fidelity on child outcomes, through caregiver engagement and parenting, reported in Smith et al. (2013), the original and re-analysis were $B = -.24$, SD = .19, 95% CI = –.664 | –.019 and $B = -.11$, SD = .11, 95% CI = –.342 | –.004, respectively. From Chiapa et al. (2015), the effect of the slope of fidelity on child outcomes were $B = .66$, SD = .46, 95% CI = .060 | 1.887 in the original analysis and $B = 2.35$, SD = 1.14, 95% CI = .490 | 4.611 in the re-analysis.[3] All models provided good fit to the data. Although the models are non-nested, thus precluding formal significance testing of differences in fit, comparing the Bayesian Information Criterion suggests only a modest decline in fit when comparing the models: Smith et al. (2013) 1998 vs. 2010; Chiapa et al. (2015) 7433 vs. 7440.[4] The modest decline in fit is likely due to measurement reliability by which a mean score is typically greater than a single-item score.

Discussion

Accurate assessment of fidelity is important across the translational research spectrum; however, a number of challenges emerge as EBPs move to the community. This study sought to evaluate three aspects of assessing fidelity to family-centered EBPs: effectiveness in distinguishing variable levels of training, differentiation from usual practice, and efficiency. Data were drawn from two completed trials of the FCU to obtain a sample of three conditions that varied by the level of training in the EBP and level of training and ongoing fidelity monitoring. A services-as-usual condition with no training in the FCU was among these. Additionally, ratings from review of complete sessions were compared to those of the 20-minute segments.

[3] Unstandardized results are presented due to Mplus only calculating unstandardized path estimates for indirect effects when using Bayesian estimation.

[4] Complete results of the re-analysis are available by request from the first author.

Effectiveness and Differentiation.

First, the effectiveness of ratings for detecting differences by training was evaluated. We hypothesized that fidelity ratings would be sensitive to the level of training and ongoing consultation in the FCU that clinicians received, and that we could reliably differentiate FCU sessions from services as usual delivered in the community (H1). For both hypotheses, we found some support, which was limited to specific dimensions of the rating system, as expected. The only dimension that was significantly different for both the complete sessions and segments was Conceptually accurate to the FCU. This was found when comparing all three training conditions and when comparing the pooled FCU conditions with services as usual. Caregiver engagement ratings differed significantly across the three conditions, but only for the ratings of complete sessions. Both analyses indicated declines between therapist groups, with the highly trained group being the highest and services as usual being the lowest. One possible explanation for this is due to the trial from which these sessions were drawn. In this trial (Trial 1), families were offered the FCU annually. Thus, there is potential that therapist and caregiver have had prior sessions, which could relate to ratings of caregiver engagement.

For Hypothesis 2, we had expected that two dimensions of the COACH rating system—Conceptually accurate to the FCU and Hope and motivation—would differ significantly between conditions. Although the results supported this hypothesis when both dimensions were included in the model simultaneously, probing the contributions of the individual dimensions clearly indicated that Conceptually accurate to the FCU was accounting for the differences found. The other dimensions of the COACH that concern therapist behaviors were not distinguishing between the three conditions or FCU sessions (pooled) compared to services as usual.

It is challenging to compare the findings of this study to the existing literature because most studies examine ratings of fidelity within only one training condition—typically either the intervention group in an efficacy or effectiveness trial or a cohort of interventionists trained to deliver an EBP in a community setting. There are a few examples of fidelity evaluations in both effectiveness trials (e.g., McLeod et al., 2017; Smith et al., 2017; Southam-Gerow et al., 2010) and community practice. Some studies of community practice have examined only whether an EBP was used, but not to what extent it was delivered with fidelity (Santa Ana et al., 2008; Weisz et al., 2013), while others have sought to characterize the extent to which elements of EBPs exist in usual care, which has required a more detailed assessment of fidelity (e.g., Brookman-Frazee et al., 2010; Garland et al., 2010; McLeod & Weisz, 2010; Smith et al., 2017). There is a need for research on fidelity to EPBs when they are intentionally translated to real-worlds settings and how these protocols relate to services as usual (i.e., are they demonstrably and meaningfully different). As EBPs become more commonplace, it should be expected that at least some aspects of EBPs will be detectable in usual care conditions. We expected and found as such in this study on certain domains of the COACH fidelity rating system. Relatedly, better understanding what clinicians routinely do could help explain the null effects of trials comparing implementation of an EBP to usual care (e.g., Southam-Gerow et al., 2010).

The findings of the current study are both theoretically and practically important. Conceptually accurate to the FCU—the therapist demonstrates an accurate understanding of the FCU model in terms of its emphasis on family-centered change, caregiver leadership in the change process, support of specific skills that define family management, and that the model is assessment-driven and tailored to the specific needs of children and families—is necessary for fidelity to this program. We have argued in previous studies of the COACH system that a composite score of the five dimensions is more appropriate than examining single-item ratings when considering the validity of those ratings. The findings presented here, however, suggest that although the composite score presents a more comprehensive and reliable index of competent adherence to the FCU, the Conceptually accurate dimension is sensitive to level of training in the FCU, differentiates FCU sessions from services as usual, and validly predicts clinical change. This is germane to initial and ongoing training in the model and for ongoing monitoring of fidelity. We had also hypothesized that the dimension of Hope and motivation would distinguish between conditions. However, this was not the case. Overall, these findings are consistent with the body of research supporting a positive association between training and fidelity (e.g., Dusenbury et al., 2003); however, the Dusenbury et al. review was not specific to family-centered interventions.

Efficiency.

The third aim of this study concerned the efficiency of observational ratings. We hypothesized that fidelity scores would be equivalent in magnitude to the ratings of complete sessions and that they would similarly differentiate between conditions (H3). The results indicated that the magnitude of the ratings did not differ by length of the videotape rated, but the reliability of the segment ratings must be considered. Based on coding challenges, we hypothesized that the interrater reliability of fidelity scores of segments would be significantly lower than scores from complete sessions (H4). Tests of COACH composite scores comparing the session length conditions indicated that reliability was significantly lower only in the minimal training condition. However, based on the findings showing that Conceptual accuracy was the dimension that differentiated between training conditions, we thought it prudent to test this variable specifically in terms of reliability using the Delta method. As expected, the difference in reliability between session length conditions was significant.

Additionally, caregiver engagement ratings were significantly higher among complete sessions compared to segments within the highly trained group. This is conceptually and empirically relevant based on previous findings using the COACH and FCU that indicated the mediating role of caregiver engagement in the link between competent adherence and clinical outcomes (Smith et al., 2013). The link from fidelity to engagement to outcomes is also specified in conceptual models of family-centered prevention program implementation (Berkel et al., 2011). The inability to accurately and reliably capture this dimension of implementation would be disadvantageous for monitoring fidelity.

The results of these analyses need to be interpreted cautiously based on the small sample in each condition, which led to wide confidence intervals for the ICCs. The likelihood of failing to detect a significant effect because of power is much higher than the failure to reject the null hypothesis. Additionally, the overall ICC was only above the acceptable threshold for the highly trained group for both the complete session and segment ratings, whereas they were in the poor range for segment ratings in the other groups. Based on these findings, we hesitate to suggest using ratings segments of FCU sessions in the manner used in this study. One potential remedy could be to rate “meaningful” 10-minute segments that are selected by the clinician. This approach has yielded valid and reliable ratings of fidelity to the Parent Management Training–Oregon model (e.g., Forgatch & DeGarmo, 2011; Forgatch et al., 2005).

Implications

This study begins to address some of the concerns regarding assessment of fidelity that have implications for implementation and translation, but more work is needed. The failure to effectively implement an EBP is the primary reason for the “dilution effect” or “voltage drop” that is sometimes seen when programs move from the lab to the real world (Chambers, Glasgow, & Stange, 2013). Fidelity is just one of the elements that contributes to implementation success. The field needs prospective data on fidelity to determine the relative contribution compared with other variables (e.g., common therapeutic factors and other elements that are not unique to the EBP), and we need monitoring systems that provide rapid, yet accurate and cost-efficient, feedback to counter drift. As our results show, rating segments as opposed to complete sessions may not be a solution for the resource-intensive nature of observational coding of fidelity, particularly given that the most important variable in the rating system could not be reliably rated with session segments. The results are promising in terms of ratings of fidelity to the FCU being sensitive to skill level. This is necessary for effectively gauging clinician’s ongoing need for consultation or remediation as the FCU is being delivered.

Strengths and Limitations

The current study sought to capitalize on existing data to answer basic questions regarding fidelity assessment using the COACH rating system. This strategy has been promoted as an efficient way of conducting implementation research (Chambers et al., 2010) but it also presents challenges. In this study, a few restrictions contributed to the design. First, we were limited in the number of sessions in the minimal training and no training conditions, which limited power. Second, more granular data concerning amount, quality, and type of training and consultation each group of clinicians received was unavailable. Thus, there is some degree of heterogeneity within each group and we were unable to examine direct relations between these variables and fidelity ratings. Relatedly, while training and working with the therapists in Trial 2, it became apparent that some community clinicians had received training in evidence-based parent training or family management programs, such as Triple P, Parent Management Training–Oregon Model, and Parent-Child Interaction Therapy. Unfortunately, this information was not collected, nor was use of these proscribed interventions assessed as part of differentiation. Both of these limitations can be addressed in future research. Last, the services-as-usual condition with no training in the FCU model had an older sample of youth compared to the other two conditions. This cannot be addressed empirically by controlling for age of the child and needs to be explored further. It could be argued that youth exhibiting problem behaviors later in adolescence are more likely to have enduring issues that are more challenging to clinicians and thus might introduce a confound in the assessment of fidelity. Future research is needed to prospectively evaluate this.

Future Directions

With the myriad demands on community practitioners, feasible and effective fidelity-rating systems are crucial. Observational assessment is resource-intensive, limiting the quantity and frequency of fidelity monitoring. One promising development to address this problem is the recent emergence of technology-assisted methods for fidelity assessment and monitoring (Brown et al., 2015). These machine learning based methods use semantic and vocal acoustic information from audio recordings of sessions to reliably assess fidelity to different aspects of EBP delivery, such as therapists' use of empathic statements (Xiao, Imel, Georgiou, Atkins, & Narayanan, 2015) and open-ended questions and complex reflections (Lord et al., 2015). Machine learning methods for coding FCU fidelity are currently being developed and tested (Smith et al., 2018); however, their widespread use in the real-world is still a few years away and the FCU and other similar parent training programs continue to go to scale in the meantime. Not only will it need to be demonstrated that these methods are feasible, acceptable, and accurate, but evidence of cost savings and downstream benefits for families is vital for uptake. Shortening existing fidelity rating scales, using such approaches as item-response theory, is another viable method.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments.

Justin Smith was supported by National Institute on Drug Abuse grant DA027828 (PI Hendricks Brown); National Institute of Mental Health research training grant MH20012 (PI Elizaebeth Stormshak); and by the Implementation Research Institute (IRI) at the George Warren Brown School of Social Work, Washington University in St. Louis through an award from the National Institute of Mental Health (5R25MH08091607) and the Department of Veterans Affairs, Health Services Research & Development Service, Quality Enhancement Research Initiative (QUERI) (PI Enola Proctor). Thomas Dishion, Daniel Shaw, and Melvin Wilson were supported by National Institute on Drug Abuse grant DA016110. Samantha Montag was supported by National Institute on Drug Abuse grant DA027828. Elizabeth Stormshak was supported by Centers for Disease Control and Prevention grant CE001389. The authors wish to thank the mebers of the research team in Eugene, Phoenix, Pittsburgh, Portland, and Charlottesville, and the families who participated in the two trials. Specifically, we appreciate the contribution of Lisa Brauer for her work in developing the COACH fidelity coding system and coding of videotaped intervention sessions. We are also grateful to C. Hendricks Brown for statistical consultation and Enola Proctor for feedback on early drafts of this paper.

References

Atkins D, Steyvers M, Imel Z, & Smyth P (2014). Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. Implementation Science, 9(1), 49. doi:10.1186/1748-5908-9-49 [PubMed: 24758152]

Berkel C, Mauricio AM, Schoenfelder EN, & Sandler IN (2011). Putting the pieces together: An integrated model of program implementation. Prevention Science, 12(1), 23–33. doi:10.1007/s11121-010-0186-1 [PubMed: 20890725]

Biglan A (2015). The nurture effect: How the science of human behavior can improve our lives and our world Oakland, CA: New Harbinger Publications.

Brookman-Frazee L, Haine RA, Baker-Ericzén M, Zoffness R, & Garland AF (2010). Factors Associated with Use of Evidence-Based Practice Strategies in Usual Care Youth Psychotherapy.

Administration and Policy in Mental Health and Mental Health Services Research, 37(3), 254–269. doi:10.1007/s10488-009-0244-9 [PubMed: 19795204]

Brown CH, PoVey C, Hjorth A, Gallo CG, Wilensky U, & Villamar J (2015). Computational and technical approaches to improve the implementation of prevention programs. Implementation Science, 10(Suppl 1), A28. doi:10.1186/1748-5908-10-S1-A28

Chambers DA, Glasgow R, & Stange K (2013). The dynamic sustainability framework: Addressing the paradox of sustainment amid ongoing change. Implementation Science, 8(1), 117. doi: 10.1186/1748-5908-8-117 [PubMed: 24088228]

Chambers DA, Wang PS, & Insel TR (2010). Maximizing efficiency and impact in effectiveness and services research. General Hospital Psychiatry, 32(5), 453–455. doi:10.1016/j.genhosppsych. 2010.07.011 [PubMed: 20851264]

Chiapa A, Smith JD, Kim H, Dishion TJ, Shaw DS, & Wilson MN (2015). The trajectory of fidelity in a multiyear trial of the Family Check-Up predicts change in child problem behavior. Journal of Consulting and Clinical Psychology, 83(5), 1006–1011. doi:10.1037/ccp0000034 [PubMed: 26121303]

Cicchetti DV (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6, 284–290. doi: 10.1037/1040-3590.6.4.284

Dishion TJ, Brennan LM, Shaw DS, McEachern AD, Wilson MN, & Jo B (2014). Prevention of problem behavior through annual Family Check-Ups in early childhood: Intervention effects from home to early elementary school. Journal of Abnormal Child Psychology, 42(3), 343–354. doi: 10.1007/s10802-013-9768-2 [PubMed: 24022677]

Dishion TJ, Forgatch M, Chamberlain P, & Pelham WE, III (2016). The Oregon model of behavior family therapy: From intervention design to promoting large-scale system change. Behavior Therapy doi:10.1016/j.beth.2016.02.002

Dishion TJ, Shaw DS, Connell A, Gardner FEM, Weaver C, & Wilson M (2008). The Family Check-Up with high-risk indigent families: Preventing problem behavior by increasing parents' positive behavior support in early childhood. Child Development, 79(5), 1395–1414. doi:10.1111/j. 1467-8624.2008.01195.x [PubMed: 18826532]

Dishion TJ, Smith JD, Gill AM, Shaw DS, & Knutson N (2014). Family Check-Up & Everyday Parenting Fidelity COACH Rating Manual: V.4.0 Arizona State University.

Dulcan MK (2000). Does community mental health treatment of children and adolescents help? Journal of the American Academy of Child and Adolescent Psychiatry, 39(2), 153–153. doi: 10.1097/00004583-200002000-00012

Dusenbury L, Brannigan R, Falco M, & Hansen WB (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. Health Education Research, 18(2), 237–256. doi:10.1093/her/18.2.237 [PubMed: 12729182]

Forgatch MS, & DeGarmo DS (2011). Sustaining fidelity following the nationwide PMTO™ implementation in Norway. Prevention Science, 12(3), 235–246. doi:10.1007/s11121-011-0225-6 [PubMed: 21671090]

Forgatch MS, Patterson GR, & DeGarmo DS (2005). Evaluating fidelity: Predictive validity for a measure of competent adherence to the Oregon Model of Parent Management Training. Behavior Therapy, 36(1), 3–13. doi:10.1016/S0005-7894(05)80049-8 [PubMed: 16718302]

Gallo C, Pantin H, Villamar J, Prado G, Tapia M, Ogihara M, … Brown CH (2014). Blending qualitative and computational linguistics methods for fidelity assessment: Experience with the Familias Unidas preventive intervention. Administration and Policy in Mental Health and Mental Health Services Research, 42(5), 574–585. doi:10.1007/s10488-014-0538-4

Garland AF, Brookman-Frazee L, Hurlburt M, Accurso EC, Zoffness RJ, Haine-Schlagel R, & Ganger W (2010). Mental health care for children with disruptive behavior problems: A view inside therapists' offices. Psychiatric Services, 61(8), 788–795. [PubMed: 20675837]

Garland AF, Hurlburt MS, & Hawley KM (2006). Examining psychotherapy processes in a services research context. *Clinical Psychology*: Science and Practice, 13(1), 30–46. doi:10.1111/j. 1468-2850.2006.00004.x

Gearing RE, El-Bassel N, Ghesquiere A, Baldwin S, Gillies J, & Ngeow E (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. Clinical Psychology Review, 31(1), 79–88. doi:10.1016/j.cpr.2010.09.007 [PubMed: 21130938]

Hanson RF, Gros KS, Davidson TM, Barr S, Cohen J, Deblinger E, … Ruggiero KJ (2013). National trainers' perspectives on challenges to implementation of an empirically-supported mental health treatment. Administration and Policy in Mental Health and Mental Health Services Research, 41(4), 522–534. doi:10.1007/s10488-013-0492-6

Hoagwood K, & Kolko DJ (2009). Introduction to the special section on practice contexts: A glimpse into the nether world of public mental health services for children and families. Administration and Policy in Mental Health and Mental Health Services Research, 36(1), 35–36. doi:10.1007/s10488-008-0201-z [PubMed: 19115103]

Hogue A, Bobek M, Dauber S, Henderson CE, McLeod BD, & Southam-Gerow MA (2017). Distilling the Core Elements of Family Therapy for Adolescent Substance Use: Conceptual and Empirical Solutions. Journal of Child & Adolescent Substance Abuse, 26(6), 437–453. doi: 10.1080/1067828X.2017.1322020 [PubMed: 30705581]

Hogue A, & Liddle HA (2009). Family-based treatment for adolescent substance abuse: Controlled trials and new horizons in services research. Journal of family therapy, 31(2), 126–154. doi: 10.1111/j.1467-6427.2009.00459.x [PubMed: 21113237]

Landsverk JA, Brown CH, Chamberlain P, Palinkas LA, Ogihara M, Czaja S, … Horwitz SM (2012). Design and analysis in dissemination and implementation research In Brownson RC, Colditz GA, & Proctor EK (Eds.), Dissemination and implementation research in health: Translating research to practice (pp. 225–260). New York, NY: Oxford University Press.

Leslie LK, Mehus CJ, Hawkins JD, Boat T, McCabe M, Barkin SL, … Beardslee W (2016). Primary health care: Potential home for family-focused preventive interventions. American Journal of Preventive Medicine, 51(4 (supp 2)), S106–S118. doi:10.1016/j.amepre.2016.05.014 [PubMed: 27498167]

Lord SP, Can D, Yi M, Marin R, Dunn CW, Imel ZE, … Atkins DC (2015). Advancing methods for reliably assessing motivational interviewing fidelity using the motivational interviewing skills code. Journal of Substance Abuse Treatment, 49, 50–57. [PubMed: 25242192]

McHugh RK, & Barlow DH (2010). The dissemination and implementation of evidence-based psychological treatments: A review of current efforts. American Psychologist, 65(2), 73–84. doi: 10.1037/a0018121 [PubMed: 20141263]

McLeod BD, Southam-Gerow MA, Jensen-Doss A, Hogue A, Kendall PC & Weisz JR (2017). Benchmarking treatment adherence and therapist competence in individual cognitive-behavioral treatment for youth anxiety disorders, Journal of Clinical Child & Adolescent Psychology, DOI: 10.1080/15374416.2017.1381914

McLeod BD, & Weisz JR (2010). The Therapy Process Observational Coding System for Child Psychotherapy Strategies Scale. Journal of Clinical Child & Adolescent Psychology, 39(3), 436–443. doi:10.1080/15374411003691750 [PubMed: 20419583]

Miller WR, & Rollnick S (2012). Motivational interviewing: Helping people prepare for change (3rd ed.). New York, NY: Guilford Press.

Miller WR, Sovereign RG, & Krege B (1988). Motivational interviewing with problem drinkers: II. The Drinker's Check-up as a preventive intervention. Behavioural and Cognitive Psychotherapy, 16(04), 251–268. doi:doi:10.1017/S0141347300014129

Perepletchikova F, Treat TA, & Kazdin AE (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. Journal of Consulting and Clinical Psychology, 75(6), 829–841. doi:10.1037/0022-006X.75.6.829 [PubMed: 18085901]

R Core Team. (2012). R: A language and environment for statistical computing Vienna, Austria: R Foundation for Statistical Computing Retrieved from http://www.R-project.org/

Rao CR (2009). Linear statistical inference and its applications (Vol. 22): John Wiley & Sons.

Santa Ana EJ, Martino S, Ball SA, Nich C, Frankforter TL, & Carroll KM (2008). What is usual about "treatment-as-usual"? Data from two multisite effectiveness trials. Journal of Substance Abuse Treatment, 35(4), 369–379. doi:10.1016/j.jsat.2008.01.003 [PubMed: 18337053]

SAS Institute Inc. (2014). The SAS system, Version 9.4. Cary, NC: SAS Institute Inc.

Schoenwald SK, Garland AF, Chapman J, Frazier S, Sheidow A, & Southam-Gerow M (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. doi:10.1007/s10488-010-0321-0 [PubMed: 20957425]

Sholomskas DE, Syracuse-Siewert G, Rounsaville BJ, Ball SA, Nuro KF, & Carroll KM (2005). We don't train in vain: A dissemination trial of three strategies of training clinicians in cognitive–behavioral therapy. Journal of Consulting and Clinical Psychology, 73(1), 106–115. doi:10.1037/0022-006X.73.1.106 [PubMed: 15709837]

Smith JD, Berkel C, Hails KA, Dishion TJ, Shaw DS, & Wilson MN (2018). Predictors of participation in the Family Check-Up program: A randomized trial of yearly services from age 2 to 10 years. Prevention Science, 19(5), 652–662, doi:10.1007/s11121-016-0679-7 [PubMed: 27405512]

Smith JD, Berkel C, Jordan N, Atkins DC, Narayanan SS, Gallo C, … Bruening MM (2018). An individually tailored family-centered intervention for pediatric obesity in primary care: Study protocol of a randomized type II hybrid implementation-effectiveness trial (Raising Healthy Children study). Implementation Science, 13(11), 1–15. doi:10.1186/s13012-017-0697-2 [PubMed: 29301543]

Smith JD, Dishion TJ, Brown K, Ramos K, Knoble NB, Shaw DS, & Wilson MN (2016). An experimental study of procedures to enhance ratings of fidelity to an evidence-based family intervention. Prevention Science, 17(1), 62–70. doi:10.1007/s11121-015-0589-0 [PubMed: 26271300]

Smith JD, Dishion TJ, Shaw DS, & Wilson MN (2013). Indirect effects of fidelity to the Family Check-Up on changes in parenting and early childhood problem behaviors. Journal of Consulting and Clinical Psychology, 81(6), 962–974. doi:10.1037/a0033950 [PubMed: 23895087]

Smith JD, Stormshak EA, & Kavanagh K (2015). Results of a pragmatic effectiveness-implementation hybrid trial of the Family Check–Up in community mental health agencies. Administration and Policy in Mental Health and Mental Health Services Research, 42(3), 265–278. doi:10.1007/s10488-014-0566-0 [PubMed: 24927926]

Smith MM, McLeod BD, Southam-Gerow MA, Jensen-Doss A, Kendall PC, & Weisz JR (2017). Does the delivery of CBT for youth anxiety differ across research and practice settings? Behavior Therapy, 48(4), 501–516. doi:10.1016/j.beth.2016.07.004 [PubMed: 28577586]

Smolkowski K, Seeley JR, Gau JM, Dishion TJ, Stormshak EA, Moore KJ, … Garbacz SA (2017). Effectiveness evaluation of the Positive Family Support intervention: A three-tiered public health delivery model for middle schools. Journal of School Psychology, 62, 103–125. doi:10.1016/j.jsp.2017.03.004 [PubMed: 28646972]

Southam-Gerow MA, & McLeod BD (2013). Advances in applying treatment integrity research for dissemination and implementation science: Introduction to special issue. Clinical Psychology: Science and Practice, 20(1), 1–13. [PubMed: 23970819]

Southam-Gerow MA, Weisz JR, Chu BC, McLeod BD, Gordis EB, & Connor-Smith JK (2010). Does cognitive behavioral therapy for youth anxiety outperform usual care in community clinics? An initial effectiveness test. Journal of the American Academy of Child & Adolescent Psychiatry, 49(10), 1043–1052. doi:10.1016/j.jaac.2010.06.009 [PubMed: 20855049]

Weck F, Bohn C, Ginzburg DM, & Stangier U (2011). Assessment of adherence and competence in cognitive therapy: Comparing session segments with entire sessions. Psychotherapy Research, 21(6), 658–669. doi:10.1080/10503307.2011.602751 [PubMed: 21793688]

Weck F, Grikscheit F, Höfling V, & Stangier U (2014). Assessing treatment integrity in cognitive-behavioral therapy: Comparing session segments with entire sessions. Behavior Therapy, 45(4), 541–552. doi:10.1016/j.beth.2014.03.003 [PubMed: 24912466]

Weisz JR, Doss AJ, & Hawley KM (2005). Youth psychotherapy outcome research: A review and critique of the evidence base. Annual Revew of Psychology, 56, 337–363. doi:10.1146/annurev.psych.55.090902.141449

Weisz JR, Jensen-Doss A, & Hawley KM (2006). Evidence-based youth psychotherapies versus usual clinical care: A meta-analysis of direct comparisons. American Psychologist, 61(7), 671–689. doi:10.1037/0003-066X.61.7.671 [PubMed: 17032068]

Weisz JR, Kuppens S, Eckshtain D, Ugueto AM, Hawley KM, & Jensen-Doss A (2013). Performance of evidence-based youth psychotherapies compared with usual clinical care: a multilevel meta-analysis. JAMA psychiatry, 70(7), 750–761. doi:10.1001/jamapsychiatry.2013.1176 [PubMed: 23754332]

Xiao B, Imel ZE, Georgiou PG, Atkins DC, & Narayanan SS (2015). “Rate My Therapist”: Automated detection of empathy in drug and alcohol counseling via speech and language processing. PLoS ONE, 10(12), e0143055. doi:10.1371/journal.pone.0143055 [PubMed: 26630392]

Table 1.

COACH mean score descriptive statistics and reliability by training condition

	Highly trained/routine and universal monitoring				Minimally trained/optional and variable monitoring				No training				Overall			
Session Length	**M**	**SD**	**ICC**	**α**	**M**	**SD**	**ICC**	**α**	**M**	**SD**	**ICC**	**α**	**M**	**SD**	**ICC**	**α**
Complete	5.14	1.36	.82	.97	4.46	.99	.71	.90	4.13	1.28	.67	.93	4.65	1.27	.75	.93
Segment	4.73	.90	.76	.89	4.44	.87	.34	.88	3.88	.97	.41	.92	4.37	.96	.61	.90
Overall	4.94	1.16	.76	.94	4.45	.92	.50	.89	4.06	1.14	.56	.92	4.50	1.13	.63	.92

Note. SD = standard deviation. ICC = one-way random effects model intraclass correlation coefficient. α = internal consistency.

Table 2.

Results of the Kruskal-Wallis Tests and the Wilcoxon Rank–Sum Tests comparing differences between the conditions

		Highly trained vs Minimally trained		Highly trained vs No training		Minimally trained vs No training		Overall			Pooled FCU vs No training	
		Z	***p***	**Z**	***p***	**Z**	***p***	χ^2	**DF**	***p***	**Z**	***p***
Complete Sessions	COACH composite score	−1.77	.077	−2.20	.028	0.89	.373	5.96	2	.051	−1.81	.070
	Conceptually accurate	−0.92	.355	−4.39	<.001	3.67	<.001	22.50	2	<.001	−4.66	<.001
	Observant and responsive	−1.69	.091	−0.83	.406	−0.87	.385	2.95	2	.229	−0.01	.995
	Actively structures	−2.07	.039	−1.50	.132	−0.08	.937	4.46	2	.108	−0.85	.397
	Corrective feedback	−1.62	.105	−1.81	.070	0.50	.614	4.22	2	.122	−1.37	.170
	Hope and motivation	−1.61	.106	−1.83	.067	0.71	.477	4.38	2	.111	−1.50	.135
	Caregiver Engagement	−3.22	.001	−3.09	.002	0.16	.874	13.47	2	.001	−1.92	.055
20-minute Segments	COACH composite score	−1.08	.278	−2.64	.008	−1.99	.047	8.07	2	.018	−2.67	.008
	Conceptually accurate	−0.35	.725	−4.54	<.0001	−3.63	<.001	22.44	2	<.001	−4.70	<.001
	Observant and responsive	−1.26	.206	−1.59	.112	−0.55	.584	3.06	2	.217	−1.24	.213
	Actively structures	−2.13	.034	−2.45	.014	−1.02	.308	7.75	2	.021	−2.00	.045
	Corrective feedback	−0.53	.597	−2.33	.020	−1.77	.077	5.90	2	.052	−2.36	.018
	Hope and motivation	−0.39	.693	−1.36	.174	−1.02	.307	2.05	2	.358	−1.37	.170
	Caregiver Engagement	1.02	.309	−1.60	.109	−2.05	.040	5.19	2	.075	−2.10	.035

Note.

[*] *p*-value significant after Bonferroni correction for multiple comparisons, < .00125.

Table 3.

Results of Roy's Maximum Root Test between the three conditions and the FCU conditions pooled compared with no training

	Three Conditions		**Pooled FCU vs No training**	
	F **(df)**	***p***	***F***	***p***
Model 1	17.37 (2, 68)	<0.001 *	16.42 (2, 68)	<0.001 *
Conceptually accurate	15.28 (2,68)	<0.001 *	27.85 (1,69)	<0.001 *
Hope and motivation	2.36 (2,68)	0.102	2.62 (1,69)	0.110
Model 2	1.88 (3,67)	0.141	1.47 (3,67)	0.231
Observant and responsive	1.55 (2,68)	0.219	0.01 (1,69)	0.939
Actively structures	2.70 (2,68)	0.075	0.47 (1,69)	0.493
Corrective feedback	2.40 (2,68)	0.098	1.82 (1,69)	0.182

Note.

* *p*-value significant after Bonferroni correction for multiple comparisons, < .00125.

Table 4.

Results of the Wilcoxon Signed-Rank Tests comparing scores for the complete sessions and segments

	Highly trained/routine and universal monitoring		Minimally trained/optional and variable monitoring		No training		Overall	
	S	***p***	**S**	***p***	**S**	***p***	**S**	***p***
COACH composite score	−48.5	.143	−2.5	.937	−25.5	.353	−252.5	.099
Conceptually accurate	−35.5	.065	−0.5	1.000	−3.0	.842	−110.0	.180
Observant and responsive	−16.5	.639	7.5	.796	−15.0	.534	−73.0	.595
Actively structures	−16.0	.609	3.0	.893	−24.0	.309	−100.5	.381
Corrective feedback	−41.5	.142	11.0	.654	−32.0	.195	−181.0	.152
Hope and motivation	−52.0	.052	−16.5	.365	−21.5	.223	−248.5	.013*
Caregiver Engagement	−82.0	.004*	22.5	.232	−18.5	.159	−228.5	.023

Note.

* *p*-value significant after Bonferroni adjustment for multiple comparisons, < .018.