Published in final edited form as:

Prev Sci. 2016 January ; 17(1): 62–70. doi:10.1007/s11121-015-0589-0.

An Experimental Study of Procedures to Enhance Ratings of Fidelity to an Evidence-Based Family Intervention

Justin D. Smith,

Center for Prevention Implementation Methodology, Northwestern University Feinberg School of Medicine

Thomas J. Dishion,

REACH Institute, Arizona State University & Oregon Research Institute

Kimbree Brown,

Prevention Science Institute, University of Oregon

Karina Ramos,

Prevention Science Institute, University of Oregon

Naomi B. Knoble,

Prevention Science Institute, University of Oregon

Daniel S. Shaw, and

University of Pittsburgh

Melvin N. Wilson

University of Virginia

Abstract

The valid and reliable assessment of fidelity is critical at all stages of intervention research and is particularly germane to interpreting the results of efficacy and implementation trials. Ratings of protocol adherence typically are reliable, but ratings of therapist competence are plagued by low reliability. Because family context and case conceptualization guide the therapist's delivery of interventions, the reliability of fidelity ratings might be improved if the coder is privy to client context in the form of an ecological assessment. We conducted a randomized experiment to test this hypothesis. A subsample of 46 families with 5-year-old children from a multisite randomized trial who participated in the feedback session of the Family Check-Up (FCU) intervention were selected. We randomly assigned FCU feedback sessions to be rated for fidelity to the protocol using the COACH rating system either after the coder reviewed the results of a recent ecological assessment or had not. Inter-rater reliability estimates of fidelity ratings were meaningfully higher

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

Kimbree Brown is now affiliated with the Oregon Social Learning Center, Eugene, Oregon. Karina Ramos is now affiliated with the University of California at Berkeley. Naomi Knoble is now affiliated with Nemours/duPont Children's Hospital, Wilmington, Delaware.

Conflict of Interest. Thomas Dishion is the developer of the Family Check-Up program. Justin Smith, Kimbree Brown, Karina Ramos, Naomi Knoble, Daniel Shaw, and Melvin Wilson have no conflicts of interest or financial relationships relevant to this article to disclose.

for the assessment information condition compared to the no-information condition. Importantly, the reliability of the COACH mean score was found to be statistically significantly higher in the information condition. These findings suggest that the reliability of observational ratings of fidelity, particularly when the competence or quality of delivery is considered, could be improved by providing assessment data to the coders. Our findings might be most applicable to assessment-driven interventions, where assessment data explicitly guides therapist's selection of intervention strategies tailored to the family's context and needs, but they could also apply to other intervention programs and observational coding of context-dependent therapy processes, such as the working alliance.

Keywords

competence; Family Check-Up; fidelity; implementation; observational coding

Treatment fidelity encompasses the extent to which the therapist delivers prescribed treatment strategies of an intervention protocol (i.e., adherence) while maintaining an adequate degree of competence or quality. One could argue that an intervention cannot be delivered with adequate adherence sans competence (Perepletchikova & Kazdin, 2005). Accurate assessment of fidelity has critical implications across all stages of intervention research. A breakdown in fidelity impedes drawing valid inferences regarding the nature of a putative treatment effect, making it impossible to determine whether the results indicate problems with the intervention itself or with the quality of its delivery (Schoenwald & Henggeler, 2004). Treatment effects generally vary with the degree of intervention fidelity (Webb, DeRubeis, & Barber, 2010) and studies of family-centered interventions have consistently linked variability in fidelity to parenting and both child and program-level outcomes (e.g., Chiapa et al., in press; Forgatch & DeGarmo, 2011; Forgatch, DeGarmo, & Beldavs, 2005; Hogue et al., 2008; Huey, Henggeler, Brondino, & Pickrel, 2000; Prado, Pantin, Schwartz, Lupei, & Szapocznik, 2006; Smith, Dishion, Shaw, & Wilson, 2013).

A number of evidence-based family-centered prevention programs are now focused on implementation in real-world service delivery systems. Fidelity itself can also be a primary indicator of implementation success, assuming fidelity is well defined, is deemed to meet or exceed minimal standards, is linked to program outcomes, and can be sustained (Landsverk et al., 2012). Yet, training providers to deliver evidence-based interventions with fidelity is recognized as one of the greatest challenges to providing effective healthcare in the United States in the twenty-first century (Institute of Medicine, 2001), which has led to a greater emphasis and scrutiny on the assessment of fidelity (Schoenwald et al., 2011).

Observational assessment is the gold standard for assessing fidelity. Best practices involve ratings by therapists who have expertise in the intervention protocol and training in the rating system (Perepletchikova & Kazdin, 2005). Although ratings of adherence to a protocol are typically reliable, ratings of competence have shown consistently low reliability across treatment modalities within individual therapy (Muse & McManus, 2013; Webb et al., 2010) as well as family-centered intervention (Cross & West, 2011). Although there are only a handful of studies, the reliability of observationally-rated adherence is typically

higher ($r = .80$ Brody et al., 2004; $ICC = .64$ to $.79$ Hogue et al., 2008; $.79$ Hogue, Liddle, Becker, & Johnson-Leckrone, 2002) than observational ratings of competence ($ICC = .15$ to $.48$ Hogue et al., 2008; $ICC = .32$ to $.81$; Hogue, Liddle, Singer, & Leckrone, 2005; $r = .62$; Lee et al., 2008). Of note, we based our approach to assessing fidelity on the Fidelity of Implementation Rating System (FIMP), an observational system that jointly assesses adherence and the quality of delivery of the Parent Management Training Oregon Model (Knutson, Forgatch, & Rains, 2003). The FIMP has demonstrated good to excellent inter-rater reliability using this rating strategy: $ICC = .71$ to $.93$ (Forgatch, Patterson, & DeGarmo, 2005); $K = .67$ (Hukkelberg & Ogden, 2013); $ICC = .76$ and $.82$ (Forgatch & DeGarmo, 2011).

One reason for poor reliability in ratings of competence or quality could be that program developers do not have a well-defined theory of intervention process. A theory of process provides a context from which the rater is able to evaluate the appropriateness (quality) of delivering the program under investigation. It stands to reason that multiple raters working from the same theory of intervention process will arrive at similar conclusions when rating therapists' competence in delivering a protocol; thus, improving inter-rater reliability. Another reason for low reliability might be the coders' awareness (or lack thereof) of the context in which the intervention is being delivered. Observational ratings of competence, and to some extent adherence, are based on their appropriateness given the case conceptualization (e.g., Silberschatz, Fretter, & Curtis, 1986). Client context is critical to formulating an individualized case conceptualization that guides intervention (Meier, 1999). Rating therapist competence without an understanding of client context could affect reliability and validity. These two factors, a theory-driven intervention process and a context-dependent case conceptualization, are intertwined and their respective, unique contributions to reliability ratings are not easily disentangled. They operate in tandem for both the therapist and the fidelity rater.

Fidelity to the Family Check-Up

In this study, assessment of fidelity to the Family Check-Up (FCU; Dishion & Stormshak, 2007) program is examined. The FCU is a brief, family-centered intervention that is assessment driven and uses Motivational Interviewing (MI; Miller & Rollnick, 2002) procedures to improve engagement, foster a therapeutic relationship, and enhance motivation to change (Dishion & Stormshak, 2007). Results from an ecological assessment form the foundation for motivating caregivers to change parenting practices. Following an initial session with the caregiver(s), the therapist visits the family's home to videotape a series of brief, semi-structured interaction tasks. The videotapes are then coded for developmentally appropriate parenting behaviors (e.g., positive behavior support) and family management practices (e.g., parental monitoring). These results, and caregiver(s) and teacher responses to questionnaires about child and family risk and protective factors, are then discussed in a collaborative feedback session. The feedback session becomes a pivotal point in the FCU, as the therapist's skills in delivering the intervention and engaging the caregivers are predictive of behavior change in parenting and the child's behavior (Smith et al., 2013). The FCU was designed to identify and motivate change in parent management practices through intervention strategies tailored and adapted to the parent's specific needs,

as revealed in the ecological family assessment. A more detailed description of the FCU model can be found in Dishion and Stormshak (2007).

In a previous study, Smith et al. (2013) found that observational ratings of fidelity to the FCU protocol were related to changes in observed positive behavior support practices one year later, which was in turn predictive of reduced child problem behaviors the following year. The effect of fidelity on outcomes occurred through ratings of observed caregiver engagement, which was positively associated with ratings of fidelity and positive parenting. The reliability of the fidelity rating, which consists of a mean of five items assessed via the COACH rating system (Dishion, Knutson, Brauer, Gill, & Risso, 2010), was acceptable ($ICC = .74$). However, two of the single-item reliability estimates were less than .60 (range .57–.76). Later, in a study of therapists in community agencies delivering the FCU, Smith, Stormshak, and Kavanagh (2014) obtained a COACH reliability of $ICC = .73$ and individual item reliabilities in the fair to excellent range ($ICC = .61$ – $.78$). Although the mean COACH score reliabilities could be considered acceptable, the low single-item reliabilities suggested potential specification issues of the individual items and/or issues with the coding procedure.

The Current Study

The aim of the present study was to evaluate a procedural refinement for rating fidelity to the FCU: providing trained coders with the family's ecological assessment data prior to rating videotaped feedback sessions for fidelity to the protocol. Previous studies reporting COACH results had not provided the coders with the family's assessment data at any point. Thus, they were rating context-dependent therapist skills and decisions without the context of the family that the therapist was using to guide the delivery of FCU feedback. We hypothesized that the reliability of fidelity ratings would be enhanced if coders were provided these data, which would give context to the therapists' clinical decision-making processes, and thus FCU skills, observed during the feedback session. With data to illuminate both the client context and clarify the lens through which the therapist bases their clinical decision-making, the reliability of fidelity ratings are likely to improve. To test this hypothesis, we randomly assigned the sessions to two conditions, one where the coders were provided with the assessment data and the other condition where they were not. We then compared the reliability of the ratings across conditions. This is highly consistent with the assessment- and theory-driven nature of the FCU, in that assessment data are the foundation of the case conceptualization, which guides the clinical decisions and intervention strategies delivered in the feedback session.

Methods

Participants

Coders—Coders were three female graduate students in a counseling psychology doctoral program. They each had achieved a master's degree in counseling psychology. Each coder had participated in a 12-month practicum during the previous academic year, which focused on delivery of the FCU. The coders had no previous experience or exposure to the COACH rating system before this study, and they were blind to the study hypotheses.

FCU sessions and participating families—Videotapes were drawn from the Early Steps Multisite Project, a randomized prevention 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; and Pittsburgh, PA. They were randomly assigned to either a group that would be offered the FCU each year, in a health maintenance framework, or to a services-as-usual condition. All therapists in the parent study possessed a master's or doctoral degree and received group supervision and ongoing fidelity monitoring after completing initial training in the FCU. The full recruitment and randomization procedure can be found in Dishion et al. (2008).

We examined a subset of families assigned to the FCU condition who had a toddler rated by caregivers as having clinically elevated problem behaviors¹ at entry into the study (i.e., age 2), and received the FCU feedback – 79 families met these criteria at age 2. Although this subsample reported clinically elevated problem behaviors, the FCU was delivered in a selective prevention context in the Early Steps trial. These families represent the subsample of the Smith et al. (2013) study, in which the COACH system was first reported. Families of this original 79 who opted to receive the FCU feedback at child age 5 were included in the current study: 53 families met this criterion. Videotapes of 46 feedback sessions were available and were included in the randomization and coding procedures. It was discovered during coding that a technical problem had occurred during one of the sessions, rendering it uncodable. Thus, 45 tapes were included in the final analysis. Families were equally distributed across the three treatment sites: Charlottesville, 15; Eugene, 15; and Pittsburgh, 16. Females represented 40% of the children in this final sample. The average age of the child was 5.39 years ($SD = 3.19$ months). Caregivers reported that the children were European American (51%), African American (29%), Hispanic/Latino (5%), and multiple ethnicities (16%). Nearly all the primary caregivers attending the feedback sessions at this age were biological mothers (96%).

Procedures

Randomization strategy for fidelity rating—The 46 FCU feedback sessions were randomly assigned to either the assessment information or no information condition and were then randomly assigned to one of the coders. In the assessment information condition, coders were provided with the results of the family's ecological assessment prior to rating the session. We controlled for coder assignment in this randomization so that each coder rated an equal number of sessions in each condition. Next, to calculate inter-rater reliability, 23 sessions (50%) were randomly selected for double coding by a different member of the coding team. The treatment site and gender of the child were proportionally distributed across conditions and coders to control for potential bias. Analyses after randomization indicated that the two conditions did not differ in terms of child gender or caregiver-reported problem behaviors assessed via the Externalizing Scale of the Child Behavior Checklist (Achenbach, 1991).

¹*T*-scores 70 on the Externalizing Scale of the parent-report version of the Child Behavior Checklist (Achenbach, 1991).

Coder training—Coders received approximately 20 hours of training in the COACH. The reliability criterion at the conclusion of training required scoring three sessions in a row with 85% agreement and a kappa greater than .65. A reliable code is one in which the two raters are within one rating point on each dimension. This standard was applied to both the no information and the information condition procedures. All coders achieved reliability on the no information procedures before they were trained to use the family assessment information. Raters attended biweekly meetings to maintain reliability and minimize rater drift after the criterion was achieved. Coders provided ratings based on the entire FCU feedback session, which ranged in length from roughly 30 to 90 minutes.

Ecological assessment data—For the information condition of the study, coders were provided with the complete questionnaire results from the Early Steps trial assessment battery. The primary caregiver (96% biological mothers in the subsample used for this study) and an alternate caregiver, when applicable, completed the questionnaires. The caregivers reported on aspects of the family environment (e.g., household chaos, number of people living in the home, family income), caregiver functioning (e.g., depression, substance use, social/partner support in parenting), child functioning (e.g., behavior problems, anxiety and depression, inhibitory control), and the use of parent management skills relevant to 5-year-old children (e.g., safeguarding, limit setting, positive behavior support). The coders were provided with a print out of the assessment results for each family that included the mean or T score for each construct assessed. A color-coded qualitative description (red = area of concern; yellow = some concern, need to monitor; green = family strength) accompanied each score. The color-coded descriptors are based on normative data for this age group and are used in the FCU feedback session to communicate the findings to caregivers. The FCU therapists are given this printout when they are preparing to provide feedback to the caregivers.

Measures

Fidelity—Therapist's fidelity to the FCU was assessed using the COACH rating system. The COACH assesses adherence to the core content components of the program and competent execution by using accomplished clinical and teaching practices, which is referred to as *competent adherence*. Competent adherence has been found to have adequate reliability, and variations in ratings were found to be predictive of the degree of change in child and family outcomes (e.g., Forgatch & DeGarmo, 2011; Smith et al., 2013). The FCU is theory based, meaning that rigid adherence to a treatment manual is neither necessary nor desired, as long as the core interventional components of the model are implemented in a manner that is tailored to the needs of each family. Thus, evaluation of fidelity to the FCU requires assessment of both content delivery and process. Competent adherence reflects a combination of constructs that form the superordinate construct of treatment fidelity. The majority of fidelity rating systems currently in use heavily favor adherence, even though it is only one component of fidelity of implementation (Berkel, Mauricio, Schoenfelder, & Sandler, 2011; Carroll et al., 2007).

The COACH system assesses five dimensions of observable therapist skill prescribed to the FCU: Conceptual accuracy and adherence (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); Observant and responsive to client needs (the therapist observes the client's immediate and pressing concerns and contextual and cultural factors and feedback is modified accordingly); Actively structures sessions (the therapist actively structures the change process using actions that encourage caregiver involvement such as constructing useful questions, conducting role-plays, and redirecting discussions to motivate and empower the caregivers to behave differently in their interactions with children) ; Careful and appropriate teaching (the therapists sensitively provides feedback to increase caregiver's accurate self-appraisals and increase motivation to either build on existing strengths or take corrective action in one or more areas); Hope and motivation are generated (the therapist incorporates specific therapeutic techniques from MI to promote caregiver's hope, motivation, and change). Observed client engagement in the session is also rated (see next section). Detailed information about each dimension is provided in Smith et al. (2013) and the COACH manual (Dishion et al., 2010). The five dimensions are rated separately on a 9-point scale: 1–3 (*needs work*), 4–6 (*acceptable work*), 7–9 (*good work*). Scores in the “needs work” range indicate minimal knowledge and skill implementing the FCU. Scores in the “acceptable work” range indicate an acceptable level of process skill and conceptual understanding of the model as evidenced by a demonstration of key basic skills. Scores in the “good work” category indicate clear understanding of the principles and mastery of the process skills of the model: Execution is relatively fluid, barriers to change are tactfully addressed, feedback is effective and realistic, motivation enhancement techniques are skillfully applied, and the therapist capitalizes on opportunities to proactively enhance behavior change and motivation.

Client engagement—Assessment of the client's level of engagement with the therapist was accomplished at the same time and by the same coders who had used the COACH and consisted of the same training and reliability procedures. High scores (7–9) indicate a client's active participation in the session as exemplified by engaging verbally and nonverbally in conversation with the therapist and staying on topic; giving complete, thoughtful responses to open-ended questions; using change talk; being willing to discuss personal information; asking questions and elaborating on the therapist's comments; showing initiative; and actively participating in session tasks. Moderate (4–6) scores are exemplified by signs of engagement in similar areas, such as occasional verbal and nonverbal signs of interest and understanding; modest elaboration of therapist comments; short responses to questions; and reluctance to disclose and share personal information and problems. Low (1–3) scores indicate that the client appeared inattentive or disengaged as exemplified by shutting out the therapist, lacking nonverbal engagement (e.g., eye contact, head nodding), displaying signs of boredom, providing very brief responses to the therapist's open-ended questions, repeatedly attempting to steer the conversation away from parent management topics, and seeming dishonest or disingenuous.

Data Analysis

First, we computed the mean and internal consistency of a score comprised of the five COACH dimensions related to therapist skills. Analysis of variance was used to compare the mean COACH scores between the information and no information conditions of the study. Next we calculated inter-rater reliability of the COACH (individual items and the mean score) using a one-way random effects model inter-rater correlation coefficient, or ICC(1,1), (Shrout & Fleiss, 1979). This was done for the full sample as well as separately for the two conditions. We used Cicchetti's (1994) interpretative guidelines to describe reliability and compare the two conditions for meaningful differences: *poor* ($< .40$), *fair* (.40–.59), *good* (.60–.74), *excellent* ($\geq .75$). Finally, we used the Delta method (Rao, 2009) to estimate the standard error of the ICCs from which we were able to perform a z -test and obtain a corresponding probability (p -value) to determine the statistical significance of the difference between the ICCs in the two conditions. Analyses were conducted in SPSS (2014) and R (2012).

Results

Intercorrelations and descriptives of COACH ratings are provided in Table 1. The COACH items are highly intercorrelated (range: $r = .74 - .88$) and internally consistent ($\alpha = .95$). We have argued that the COACH mean score, and not single-item scores, accurately represents fidelity to the FCU (Smith et al., 2013). The mean level of COACH scores did not differ by condition ($F(1,43) = .068, p = .80$) or coder ($F(1,43) = 1.143, p = .33$). Inter-rater reliability estimates in this study are presented in Table 2 alongside the results of the two previous studies reporting COACH score reliability. The 95% confidence intervals are provided for the ICC estimates from this study. Individual-item ICCs for the full sample were in the fair to good range (.52 – .67), client engagement was .59, and the COACH mean score had an ICC of .67. To address our primary hypothesis, we calculated reliabilities for the two assessment information conditions separately. The ICCs for the information condition ranged from good to excellent (.69 – .88) and the COACH mean was .81. ICCs for the no-information condition ranged from poor to fair (.15 – .42) and the COACH mean score had an ICC of .33. After applying the Delta method, the results of z -tests indicated that two individual-item ICCs (Actively structures, Hope and Motivation) and the ICC of the COACH mean score were reliably different between conditions, while two others (Conceptually accurate, client engagement) approached significance: C: $z = 1.90, p = .06$; O: $z = 0.86, p = ns$; A: $z = 2.58, p = .01$; C: $z = 1.22, p = ns$; H: $z = 2.32, p = .02$; COACH mean: $z = 2.30, p = .02$; client engagement: $z = 1.91, p = .06$.

Discussion

In this study we randomly assigned FCU feedback sessions to be rated for fidelity to the model with or without the coder having reviewed the family's ecological assessment findings and compared inter-rater reliabilities. We found support for our prediction that providing the coders with the families' assessment data would improve the reliability of their fidelity ratings. The results indicated pronounced differences by condition. The reliability of COACH ratings in the assessment information condition were in the good to

excellent range. Conversely, reliability of the no information condition ratings was in the poor to fair range. Further, the confidence intervals indicate more trustworthy reliability estimates were achieved in the information condition compared to the no information condition. Statistically significant differences in two individual-item ICCs, and most importantly the COACH mean score ICC, provide further evidence for our hypothesis that providing raters with families' assessment data contributes to meaningful differences in reliability. However, interpretation of these results should be cautious since confidence intervals, which are integral in the Delta method, are directly related to the size of the sample. Given the power limitations, our results are quite promising.

The absolute differences in the magnitudes between the conditions were larger than expected, as were the differences in the range of the confidence intervals, which indicate the trustworthiness of the ICC estimates. The information condition not only resulted in better reliability estimates but they were the most trustworthy compared to the no information condition. Finding differences in reliability between the two conditions might seem self-evident in the context of the FCU. The therapist's clinical decisions and selection of intervention techniques in the feedback session are directly informed by the results of the ecological assessment. Without access to the raw data from which the therapist is working, coders are at a distinct disadvantage while evaluating the appropriateness of the therapist's clinical skills in delivering the model. It seems that raters were better able to agree upon the extent to which the therapist provided appropriate feedback and guided delivery of the intervention when assessment data was available.

The reliability of caregiver engagement ratings was also meaningfully better in the information condition. Caregiver engagement has been identified as an important factor in fidelity to the FCU as it relates to outcome (Smith et al., 2013). Although this dimension of the COACH primarily concerns caregiver rather than therapist behaviors, there is certainly an interactional process occurring between therapists and caregivers. Coder's understanding of the family's context is helpful to appreciating a therapist's skill in delivering FCU techniques that are explicitly aimed at engaging the caregiver in the session. For example, knowing a caregiver is clinically depressed is helpful when appreciating a therapists efforts and success with client engagement, as well as the client's effort to overcome depression in the context of the session. The ecological assessment data helped the coders develop a more accurate case conceptualization, which clarified how the FCU could be helpful for the family and whether the caregiver was able to articulate the problem, offer effective solutions, demonstrate understanding, and to develop a shared understanding of the family story with the therapist. It seems clear that the context of the family is important for reliably evaluating the in-session behaviors of both the clinician and the caregiver(s). In the context of the FCU, which is a theory-driven intervention program (Dishion & Stormshak, 2007), the accuracy of the therapist's case conceptualization is critical for delivering the model and this also appears to be true when observationally evaluating the quality of FCU skills.

Conversely, the coder's agreement in the no information condition was much lower than expected. It is worth considering how findings in this study compare to previous reliability estimates achieved for ratings with the COACH system, which differed slightly in procedural and coder characteristics, as well as how the coding procedures of this study

might have affected reliability when no information was provided. For ease of comparison, Table 2 includes the reliability estimates of previously published studies (Smith et al., 2013; Smith et al., 2014). In general, ICCs from the information condition in this study are quite comparable to those from the two previous studies, with the exception of a fairly large improvement in the dimension of “Actively structures.” There was modest improvement from previous studies for the dimensions of “Conceptually accurate,” “Hope and Motivation,” and the COACH mean score. Following Cicchetti's (1994) interpretative guidelines, these modest improvements are likely to be somewhat meaningful (e.g., moving from good to excellent). The observed differences could be attributable to providing assessment data to the coder, to the composition of the coding team, or to a combination of these two factors. This study included only coders trained in the FCU, whereas previous teams included coders with varied FCU knowledge, including no formal training. Rating fidelity to an evidence-based intervention by coders trained in the program adheres to previous suggestions (Kazdin, 2003) and also more closely resembles the real-world use of the COACH system in which supervisors with training and experience in the FCU serve as primary coders. Additionally, real-world application of the COACH system will also naturally include use of the ecological assessment data. Supervisors and trainers will review the ecological assessment data with the clinician prior to the feedback session and then rate the session for fidelity on the COACH. Thus, the findings of this study arguably have greater external validity compared to prior studies. However, because comparison of these results with other studies is quasi-experimental, any conclusions drawn must be tentative.

The study of reliability in previous studies was confounded by having trained coders who were not clinicians. The COACH measure is intended for use with clinicians in general and supervising clinicians in particular. The use of coders trained in the FCU could explain the large differences in agreement found in this study. Despite the coders being trained to an a priori reliability on both conditions, the coders reported that they instinctively developed a case conceptualization for the family regardless of whether they were provided assessment data as a basis for making these inferences. Therefore, it seems that when assessment data are provided, two coders are better able to develop a shared conceptualization. As fidelity ratings are based in large part on the application of the FCU techniques and procedures that correspond to the case conceptualization, failing to develop a shared and accurate conceptualization would likely affect agreement. This unexpected finding raises a very important question concerning how clinicians ought to be trained to use observational fidelity rating systems. That is, training to the rating system itself, as was done in previous studies of the FCU and achieved passable reliability estimates, might prove more challenging when coders with training in the model almost inevitably apply clinical judgment. If this is the case, then it seems that providing assessment data might mitigate the effect. Our study was not designed to test this hypothesis, hence our interpretations are put forth cautiously. However, this matter is highly relevant to implementation and deserves further evaluation.

Implications

Our current results have implications for the evaluation of fidelity in a variety of research and applied contexts, from clinical trials to monitoring of fidelity in the community during

scale up. This implication likely extends beyond the FCU model to other intervention programs, but this needs to be tested. Our findings are also broadly informative to the fields of intervention and implementation science. For clinicians and community-based agencies adopting the FCU, initial training to acceptable levels of fidelity, followed by ongoing fidelity monitoring, will be imperative for evaluating implementation outcomes and clinical effectiveness. On the basis of our findings in this study, a review of the family's assessment data should precede rating FCU feedback sessions for fidelity. In that coders in this study reported that reviewing the results required less than 15 minutes, and supervisors in the community likely would have reviewed the results with the therapist prior to the feedback session, this procedural modification seems to have a cost benefit.

The results might be most applicable to intervention program that are explicitly assessment-driven and directly link assessment findings to case conceptualization, clinical decisions, and intervention procedures. However, our findings beg the question of whether access to even basic assessment data would improve the reliability of coder's ratings of therapists' clinical behaviors in the context of other intervention models and when rating therapeutic processes beyond fidelity. For example, the reliability of observational ratings of therapy processes such as the therapeutic alliance, where case conceptualization informs the therapist's navigation of alliance ruptures and repairs (Safran, Muran, Samstag, & Stevens, 2001), might be improved with access to assessment data.

Last, the operationalization of fidelity in the COACH rating system as *competent adherence*, or a combination of adherence to core skills coupled with the quality level of delivery, addresses certain shortcomings in the literature. Competence or quality of program delivery has received less attention than adherence in the context of assessing fidelity to evidence-based prevention programs, despite being inherently interconnected such that program adherence is unachievable in the absence of competent delivery (Pereplechikova & Kazdin, 2005). Although jointly assessing adherence and competence in this unidimensional manner disallows interpreting suboptimal scores as the result of competent nonadherence or incompetent adherence, because it could be either, the inclusion of a quality dimension is a step forward. Intervention developers ought to design implementation rating systems that include both adherence and competence and incorporate them in a meaningful way. Second, the results of this study suggest that competence ratings might be at least in part dependent upon the context of the client or family. The development of fidelity assessment instruments that have a competence component should incorporate procedures to provide coders with the client or family context. The results of a recent assessment, which includes constructs and data aligned with the theory and targets of the intervention protocol, are likely to be the most relevant for these purposes. The training of implementation raters would thus include instructions on how to use the assessment data to rate therapists' skill in delivering program components. As mentioned previously, this procedure might have the largest impact on fidelity ratings in the context of intervention approaches that explicitly link in-session intervention decisions on assessment data, such as the FCU.

Limitations

A number of limitations of this study must be addressed. First is the small sample size, which affects power and trustworthiness of the reliability estimates and limits the generalizability of the findings. The sample characteristics could also be a limiting factor. These families reported clinically significant behavior problems in their 2-year-old child 3 years prior to receiving the FCU feedback that was rated for fidelity. Thus, the findings might not apply to families with children displaying less or more severe behavior problems. A second aspect of this design is the health maintenance approach: These families were offered the FCU each year between child ages 2 and 5, and all participated at least at ages 2 and 5, which was our criterion for inclusion in this study. Engagement and fidelity are likely to be somewhat related to the existing relationship these families have with the FCU therapist.

Conclusions and Future Directions

This study begins to address some of the concerns regarding assessment of intervention fidelity that have implications for implementation (Carroll et al., 2007; Schoenwald et al., 2011), but more work is needed. Accurately assessing and monitoring fidelity is imperative for successful uptake outside of carefully controlled laboratory settings (e.g., Henggeler, Schoenwald, Liao, Letourneau, & Edwards, 2002). Our findings, which closely emulate application in community settings, indicate that rating fidelity to the FCU can be carried out with reliability and that reviewing family assessment data improves inter-rater agreement. Future research in this area might evaluate the benefits of using assessment data when rating fidelity and other clinical processes across intervention models. Further, the COACH system needs to be systematically tested within a multifaceted fidelity of implementation monitoring system intended to prevent drift and ensure proper delivery of the FCU in community uptake efforts. With the myriad demands on community practitioners, feasible and effective fidelity rating systems are crucial.

Acknowledgments

This work was supported by the National Institute on Drug Abuse (NIDA) through a pilot study grant awarded to Justin Smith by the Center for Prevention Implementation Methodology for Drug Abuse and Sex Risk Behavior (DA027828) and DA016110 to Thomas Dishion, Daniel Shaw, and Melvin Wilson. Justin Smith and Naomi Knoble received support from National Institute of Mental Health research training grant T32 MH20012, awarded to Elizabeth Stormshak. The authors wish to thank Cady Berkel, Amanda Chiapa, Guillermo Prado, and Juan Villamar for their helpful comments on earlier drafts of this manuscript; Hendricks Brown and Carlos Gallo for their assistance with the statistical analyses; the Early Steps team in Eugene, Pittsburgh, and Charlottesville; and the families who participated in the study.

References

- Achenbach, TM. Manual for the Child Behavior Checklist/4-18 and 1991 Profile. University of Vermont, Department of Psychiatry; Burlington, VT: 1991.
- Berkel C, Mauricio AM, Schoenfelder EN, Sandler IN. Putting the pieces together: An integrated model of program implementation. *Prevention Science*. 2011; 12(1):23–33. doi: 10.1007/s11121-010-0186-1. [PubMed: 20890725]
- Brody GH, Murry VM, Gerrard M, Gibbons FX, Molgaard V, McNair L, Neubaum-Carlan E. The Strong African American Families Program: Translating Research into Prevention Programming. *Child Development*. 2004; 75(3):900–917. doi: 10.2307/3696600. [PubMed: 15144493]

- Carroll C, Patterson M, Wood S, Booth A, Rick J, Balain S. A conceptual framework for implementation fidelity. *Implementation Science*. 2007; 2(1):40. doi: 10.1186/1748-5908-2-40. [PubMed: 18053122]
- Chiapa A, Smith JD, Kim H, Dishion TJ, Shaw DS, Wilson MN. 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*. in press doi: 10.1037/ccp0000034.
- Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*. 1994; 6:284–290. doi: 10.1037/1040-3590.6.4.284.
- Cross WF, West JC. Examining implementer fidelity: Conceptualizing and measuring adherence and competence. *Journal of children's services*. 2011; 6(1):18–33. doi: 10.5042/jcs.2011.0123.
- Dishion TJ, Knutson N, Brauer L, Gill A, Risso J. Family Check-Up: COACH ratings manual. 2010 Unpublished coding manual. Available from the Child and Family Center, 6217 University of Oregon, Eugene, OR 97403.
- Dishion TJ, Shaw DS, Connell A, Gardner FEM, Weaver C, Wilson M. The Family Check-Up with high-risk indigent families: Preventing problem behavior by increasing parents' positive behavior support in early childhood. *Child Development*. 2008; 79(5):1395–1414. doi: 10.1111/j.1467-8624.2008.01195.x. [PubMed: 18826532]
- Dishion, TJ.; Stormshak, EA. Intervening in children's lives: An ecological, family-centered approach to mental health care. American Psychological Association; Washington, DC: 2007.
- Forgatch MS, DeGarmo DS. Sustaining fidelity following the nationwide PMTO™ implementation in Norway. *Prevention Science*. 2011; 12(3):235–246. doi: 10.1007/s11121-011-0225-6. [PubMed: 21671090]
- Forgatch MS, DeGarmo DS, Beldavs ZG. An efficacious theory-based intervention for stepfamilies. *Behavior Therapy*. 2005; 36(4):357–365. doi: 10.1016/S0005-7894(05)80117-0. [PubMed: 16718303]
- Forgatch MS, Patterson GR, DeGarmo DS. Evaluating fidelity: Predictive validity for a measure of competent adherence to the Oregon Model of Parent Management Training. *Behavior Therapy*. 2005; 36(1):3–13. doi: 10.1016/S0005-7894(05)80049-8. [PubMed: 16718302]
- Henggeler SW, Schoenwald SK, Liao JG, Letourneau EJ, Edwards DL. Transporting efficacious treatments to field settings: The link between supervisory practices and therapist fidelity in MST programs. *Journal of Clinical Child & Adolescent Psychology*. 2002; 31(2):155–167. doi: 10.1207/S15374424JCCP3102_02. [PubMed: 12056100]
- Hogue A, Henderson CE, Dauber S, Barajas PC, Fried A, Liddle HA. Treatment adherence, competence, and outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology*. 2008; 76(4):544–555. doi: 10.1037/0022-006x.76.4.544. [PubMed: 18665684]
- Hogue A, Liddle HA, Becker D, Johnson-Leckrone J. Family-based prevention counseling for high-risk young adolescents: Immediate outcomes. *Journal of Community Psychology*. 2002; 30(1):1–22. doi: 10.1002/jcop.1047.
- Hogue A, Liddle HA, Singer A, Leckrone J. Intervention fidelity in family - based prevention counseling for adolescent problem behaviors. *Journal of Community Psychology*. 2005; 33(2): 191–211. doi: 10.1002/jcop.20031.
- Huey SJ Jr. Henggeler SW, Brondino MJ, Pickrel SG. Mechanisms of change in multisystemic therapy: Reducing delinquent behavior through therapist adherence and improved family and peer functioning. *Journal of Consulting and Clinical Psychology*. 2000; 68(3):451–467. doi: 10.1037/0022-006x.68.3.451. [PubMed: 10883562]
- Hukkelberg SS, Ogden T. Working alliance and treatment fidelity as predictors of externalizing problem behaviors in parent management training. *Journal of Consulting and Clinical Psychology*. 2013; 81(6):1010–1020. doi: 10.1037/a0033825. [PubMed: 23895086]
- Institute of Medicine. Crossing the quality chasm: A new health system for the 21st century. National Academy Press; Washington, DC: 2001.
- Kazdin, AE. Research design in clinical psychology. 4th ed.. Allyn & Bacon; Boston, MA: 2003.

- Knutson, NM.; Forgatch, MS.; Rains, LA. Fidelity of Implementation Rating System (FIMP): The training manual for PMTO. Oregon Social Learning Center; Eugene, OR: 2003.
- Landsverk, JA.; Brown, CH.; Chamberlain, P.; Palinkas, LA.; Ogiwara, M.; Czaja, S.; Horwitz, SM. Design and analysis in dissemination and implementation research. In: Brownson, RC.; Colditz, GA.; Proctor, EK., editors. *Dissemination and implementation research in health: Translating research to practice*. Oxford University Press; New York, NY: 2012. p. 225-260.
- Lee C-Y, August G, Realmuto G, Horowitz J, Bloomquist M, Klimes-Dougan B. Fidelity at a distance: Assessing implementation fidelity of the Early Risers prevention program in a going-to-scale intervention trial. *Prevention Science*. 2008; 9(3):215–229. doi: 10.1007/s11121-008-0097-6. [PubMed: 18648936]
- Meier ST. Training the Practitioner-Scientist: Bridging Case Conceptualization, Assessment, and Intervention. *The Counseling Psychologist*. 1999; 27(6):846–869. doi: 10.1177/0011000099276008.
- Miller, WR.; Rollnick, S. *Motivational interviewing: Preparing people for change*. 2nd ed.. Guilford Press; New York, NY: 2002.
- Muse K, McManus F. A systematic review of methods for assessing competence in cognitive-behavioural therapy. *Clinical Psychology Review*. 2013; 33(3):484–499. doi: 10.1016/j.cpr.2013.01.010. [PubMed: 23454222]
- Perepletchikova F, Kazdin AE. Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*. 2005; 12(4):365–383. doi: 10.1093/clipsy/Bpi045.
- Prado G, Pantin H, Schwartz SJ, Lupei NS, Szapocznik J. Predictors of engagement and retention into a parent-centered, ecodevelopmental HIV preventive intervention for Hispanic adolescents and their families. *Journal of Pediatric Psychology*. 2006; 31(9):874–890. doi: 10.1093/jpepsy/jsj046. [PubMed: 16049264]
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2012. Retrieved from <http://www.R-project.org/>
- Rao, CR. *Linear statistical inference and its applications*. Vol. 22. John Wiley & Sons; 2009.
- Safran JD, Muran JC, Samstag LW, Stevens C. Repairing alliance ruptures. *Psychotherapy: Theory, Research, Practice, Training*. 2001; 38(4):406–412. doi: 10.1037/0033-3204.38.4.406.
- Schoenwald SK, Garland AF, Chapman J, Frazier S, Sheidow A, Southam-Gerow M. Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*. 2011; 38(1):32–43. doi: 10.1007/s10488-010-0321-0. [PubMed: 20957425]
- Schoenwald SK, Henggeler SW. A public health perspective on the transport of evidence-based practices. *Clinical Psychology: Science and Practice*. 2004; 11(4):360–363. doi: 10.1093/clipsy.bph092.
- Shrout P, Fleiss J. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*. 1979; 86:420–428. doi: 10.1037/0033-2909.86.2.420. [PubMed: 18839484]
- Silberschatz G, Fretter PB, Curtis JT. How do interpretations influence the process of psychotherapy? *Journal of Consulting and Clinical Psychology*. 1986; 54(5):646–652. doi: 10.1037/0022-006X.54.5.646. [PubMed: 3771882]
- Smith JD, Dishion TJ, Shaw DS, Wilson MN. Indirect effects of fidelity to the Family Check-Up on changes in parenting and early childhood problem behaviors. *Journal of Consulting and Clinical Psychology*. 2013; 81(6):962–974. doi: 10.1037/a0033950. [PubMed: 23895087]
- Smith JD, Stormshak EA, Kavanagh K. 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*. 2014; 42(3):265–278. doi: 10.1007/s10488-014-0566-0. [PubMed: 24927926]
- SPSS Statistics Version 22.0.0. SPSS Inc.; Chicago, IL: 2014.
- Webb CA, DeRubeis RJ, Barber JP. Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology*. 2010; 78(2):200–211. doi: 10.1037/a0018912. [PubMed: 20350031]

Table 1

Intercorrelations between ratings on the COACH

Variable	1	2	3	4	5	6	7	8
1 <u>C</u> onceptually accurate	—	.75**	.80**	.77**	.74**	.87**	.58**	-.01
2 <u>O</u> bservant and responsive		—	.80**	.87**	.88**	.94**	.64**	-.03
3 <u>A</u> ctively structures			—	.79**	.85**	.92**	.71**	-.13
4 <u>C</u> orrective feedback				—	.87**	.94**	.53**	.00
5 <u>H</u> ope and motivation					—	.95**	.71**	-.03
6 COACH composite score						—	.69**	-.04
7 Engagement							—	-.11
8 Information Condition								—

Note.

**
 $p < .01$.

Table 2

Reliability of COACH Ratings

	From Smith, et al. (2013)			From Smith, Stormshak, Kavanagh (2014)		
	<i>M</i>	<i>SD</i>	Full Sample ICC [95% CI]	Information ICC [95% CI]	No information ICC [95% CI]	ICC
Conceptually accurate	5.18	1.08	.52 [.121 .798]	.71 [.557 .912]	.15 [-.143 .788]	.64
Observant and responsive	5.30	1.34	.58 [.314 .823]	.69 [.688 .905]	.42 [.321 .857]	.71
Actively structures	4.98	1.27	.66 [.186 .854]	.88 [.618 .964]	.24 [.066 .811]	.61
Corrective feedback	5.00	1.41	.66 [.201 .856]	.77 [.533 .929]	.41 [.189 .853]	.78
Hope and motivation	4.95	1.56	.61 [.175 .834]	.83 [.532 .947]	.30 [.146 .744]	.72
COACH composite score	5.08	1.23	.67 [.220 .860]	.82 [.699 .944]	.33 [.195 .834]	.73
Engagement	5.32	1.38	.59 [.229 .825]	.81 [.586 .943]	.32 [.140 .831]	.80

Note. ICC = random effects model intraclass correlation coefficient.