Assessing Fidelity of Treatment Delivery in Group and Individual 12-Step Facilitation

Barbara K. Campbell, PhD[1], **Jennifer K. Manuel, PhD**[2], **Sarah Turcotte Manser, MA**[3], **K. Michelle Peavy, PhD**[4], **Julija Stelmokas, MS**[5], **Dennis McCarty, PhD**[1], and **Joseph R. Guydish, PhD**[3]

[1]Oregon Health & Science University, Department of Public Health & Preventive Medicine, Portland, OR 97239, United States

[2]University of California, San Francisco, Department of Psychiatry, San Francisco, CA 94118, United States

[3]University of California, San Francisco, Philip R. Lee Institute for Health Policy Studies, San Francisco, CA 94118, United States

[4]University of Washington, Alcohol & Drug Abuse Institute, Seattle, WA 98105, United States

[5]Pacific University, School of Professional Psychology, Hillsboro, OR, 97123, United States

Abstract

Twelve Step Facilitation (TSF) is an emerging, empirically supported treatment, the study of which will be strengthened by rigorous fidelity assessment. This report describes the development, reliability and concurrent validity of the Twelve Step Facilitation Adherence Competence Empathy Scale (TSF ACES), a comprehensive fidelity rating scale for group and individual TSF treatment developed for the National Drug Abuse Treatment Clinical Trials Network study, Stimulant Abuser Groups to Engage in 12-Step. Independent raters used TSF ACES to rate treatment delivery fidelity of 966 (97% of total) TSF group and individual sessions. TSF ACES summary measures assessed therapist treatment adherence, competence, proscribed behaviors, empathy and overall session performance. TSF ACES showed fair to good overall reliability; weighted kappa coefficients for 59 co-rated sessions ranged from .31–1.00, with a mean of .69. Reliability ratings for session summary measures were good to excellent (.69–.91). Internal consistency for the instrument was variable (.47–.71). Relationships of the TSF ACES summary measures with each other, as well as relationships of the summary measures with a measure of therapeutic alliance provided support for concurrent and convergent validity. Implications and future directions for use of TSF ACES in clinical trials and community treatment implementation are discussed.

Keywords

12-step facilitation; treatment fidelity; coding system

Corresponding author: Barbara K. Campbell, PhD, 1942 NW Kearney St. #24, Portland, Oregon 97209, Phone: 503.227.3932, Fax: 503.636.8784, drbarbaracampbell@earthlink.net.

1. Introduction

Treatment fidelity, also known as treatment integrity, is essential for conducting meaningful clinical trials of behavioral interventions. Gearing et al. (2011) define treatment fidelity as "the extent to which core components of interventions are delivered as intended by the protocols (p. 79)." Ensuring treatment fidelity increases a study's internal validity; it is necessary for accurately measuring treatment effects (Borrelli, 2011; Gearing et al., 2011) and is useful for identifying the active ingredients of an effective treatment. In addition, treatment fidelity enhances external validity; fidelity procedures can facilitate accurate replication in a range of treatment settings (Bellg et al., 2004). This is particularly important when conducting multisite trials in community treatment settings with diverse samples of providers and participants (Baer et al., 2007; Campbell, 2011). Finally, fidelity procedures can be used for implementing evidence-based treatments (EBTs) in clinical practice. For example, a treatment such as Twelve Step Facilitation (TSF) (Baker, 1998; Nowinski, Baker, & Carroll, 1992) requires rigorous implementation fidelity in order to distinguish it from similar, but untested 12-step oriented treatments routinely practiced in community treatment (Carroll et al., 2000). In short, treatment fidelity is a foundation of behavioral research from intervention development through clinical implementation.

1.1. Components of Treatment Fidelity

Monitoring the delivery of treatment across interventionists, participants, and sessions is an essential fidelity procedure (Carroll et al., 2007; Gearing et al., 2011). Components of treatment delivery that should be monitored include: (a) therapist adherence to prescribed content, (b) therapist competence or skill, both specific to the prescribed treatment and non-specific (e.g., therapist empathy, timing and alliance-facilitating behavior) and (c) differentiation from other treatments (Borrelli, 2011; Gearing et al., 2011; Perepletchikova, Treat & Kazdin, 2007). Adherence, competence and treatment differentiation can be measured by monitoring both prescribed and proscribed therapist behaviors. Waltz, Addis, Koerner and Jacobson (1993) recommend inclusion of proscribed therapist behaviors that are unique treatment elements from a comparison treatment to assess potential contamination of one treatment by another. Proscribed therapist behaviors can also be demonstrations of poor general skill that detract from any therapy (Carroll et al., 2000). Measuring both specific (e.g., intervention adherence) and non-specific (e.g., empathy, alliance-facilitating behaviors) components of treatment delivery provides comprehensive performance information for supervision during a trial. It also allows for identification of elements that contribute independently to patient outcomes. In a meta-analytic review, Webb, DeRubeis, and Barber (2010) found significant heterogeneity in the contributions of therapist adherence and competence to outcomes across studies, and that controlling for therapeutic alliance resulted in significantly smaller competence-outcomes effect sizes. Yet, Borrelli et al. (2005) found that nonspecific therapist skills were rarely measured in health behavior studies. Thus, treatment outcomes may be confounded by unmeasured and unknown intervention components. Developing methods that accurately assess these components is necessary for identifying the active ingredients of a particular treatment. There are a number of ways to monitor treatment delivery, including exit interviews with patients, subjective ratings by interventionists of their own performance, and objective ratings by observers of recorded or live sessions (Borrelli et al., 2005). Monitoring across time is important to ensure therapists achieve initial criterion performance and maintain it, preventing deviation from the prescribed intervention over the course of treatment. Evaluation of sessions by independent raters has been shown to be the most reliable method (Borrelli, 2011), primarily because providers' self-report of adherence often differs from objective evaluation of adherence (e.g., Miller & Mount, 2001). Review of recorded sessions across the span of treatment by independent raters using scales with demonstrated reliability

and validity is considered the gold standard of measurement of treatment delivery (Baer et al., 2007; Gearing et al., 2011).

1.2. Monitoring Fidelity in Substance Abuse Treatments

The implementation of evidence-based treatments in substance abuse treatment relies on treatment studies with strong internal and external validity; treatment fidelity is a primary component of validity. In fact, the Substance Abuse and Mental Health Service Administration's Registry of Evidence-Based Programs and Practices lists intervention fidelity as one of six criteria used to evaluate the quality of treatment research (National Registry of Evidence-based Programs and Practices, 2012). Despite the importance of treatment fidelity monitoring in substance abuse treatment research (Borelli et al., 2011; Carroll & Rounsaville, 2003; Gearing et al., 2011), few rating systems have reported psychometric properties (Baer et al., 2007). Reliability and validity data have been reported for individual drug counseling (Barber, Mercer, Krakauer, & Calvo, 1996), supportive-expressive psychotherapy (Barber, Krakauer, Calvo, Badgio, & Faude, 1997), motivational interviewing (see Madson & Campbell, 2006 for a review), and two models of family therapy (Hogue et al., 2008; Robbins et al., 2011). Carroll et al. (2000) developed the Yale Adherence and Competence Scale (YACS) to measure treatment delivery fidelity for three, individual therapies: (a) cognitive behavioral therapy, (b) clinical management, and (c) TSF. The YACS demonstrated good reliability, concurrent and discriminant validity in a sample of 576 sessions from 122 participants with comorbid cocaine and alcohol substance use disorders.

1.3. Measuring Fidelity of Group and Individual TSF

TSF is a manual guided treatment that seeks to increase clients' engagement in 12-step activities outside of formal treatment sessions. TSF has a developing base of empirical support (Brown, Seraganian, Tremblay, & Annis, 2002; Carroll, Nich, Ball, McCance, & Rounsaville, 1998; Kaskutas, Subbaraman, Witbrodt, & Zemore, 2009; Project Match Research Group, 1997; Project Match Research Group, 1998), yet the study of fidelity in TSF is an under-investigated area of research. Rigorous fidelity assessment of TSF interventions can strengthen its evidence base by monitoring the internal validity of trials, particularly regarding the delivery of TSF, as distinct from routine, but untested, 12-step oriented treatment. Information from fidelity assessment can also inform the identification of mechanisms of action in TSF, including both specific (e.g., education about 12-step principles, facilitating contact with other 12-step members) and non-specific (e.g., empathic vs. confrontational delivery) treatment factors. Of particular interest may be mechanisms of action in TSF that generate different outcomes for different subgroups of participants (e.g., Brown et al., 2002). Finally, a reliable and valid fidelity instrument and defined fidelity monitoring procedures can assist TSF implementation in community treatment settings.

This report describes the development of a comprehensive fidelity rating scale for group and individual TSF tested in Stimulant Abuser Groups to Engage in 12-Step (STAGE-12; Daley, Baker, Donovan, Hodgkins, & Perl, 2011), a multisite study conducted within the National Drug Abuse Treatment Clinical Trials Network (CTN). The STAGE-12 TSF study presented certain advantages for examining fidelity and developing a ratings instrument. First, STAGE-12 took place in community treatment programs, providing the opportunity to assess fidelity to a manualized intervention conducted by community-based counselors. Second, STAGE-12 included individual and group sessions, capitalizing on the established efficacy of TSF for individual treatment (Project MATCH Research Group, 1997), as well as group-based variations (Brown et al., 2002). As a combined individual and group treatment, it matched formats typically used within community treatment settings, increasing its adoption potential. Following completion of the STAGE-12 clinical trial, we developed

the Twelve Step Facilitation Adherence Competence Empathy Scale (TSF ACES) to comprehensively evaluate treatment fidelity and examine its relationship to patient outcomes. To our knowledge, this is the only TSF fidelity rating scale that measures both group and individual sessions. This report focuses on scale composition, reliability, concurrent and convergent validity, as well as on methods used to train, certify, monitor and enhance the reliability of raters.

2. Method

2.1. STAGE-12 Study

STAGE-12 was a randomized, clinical trial conducted at 10 community-based, outpatient, addiction treatment centers across the United States. The study compared the effectiveness of combined group and individual TSF integrated into standard treatment with standard treatment alone for stimulant abusers.

2.1.1. Client Participants—Research staff explained the study to potential participants, obtained informed consent and conducted a screening for eligibility. Inclusion criteria for participants were: (a) 18 years of age or older; (b) admitted to outpatient treatment with at least five weeks of outpatient treatment remaining and a minimum of five hours of treatment per week; (c) used a stimulant drug within the past 60 days (or within the past 90 days, if incarcerated during the past 60 days); (d) met DSM-IV diagnostic criteria for current stimulant abuse or dependence as a primary or secondary drug of abuse; and (e) able to provide consent (including willingness to provide substance use information, accept random assignment to treatment, and be audio-recorded during all treatment sessions). The final study sample of 471 adults included 59% women, 48% Caucasians, 36% African Americans, 9% reporting more than one race, and 6% reporting Latino or Hispanic ethnicity. The mean age of participants was 38 (SD=9.7) years, and mean years of education was 12.2 (SD=1.6). Just over 35% of the sample reported being unemployed and 21 % were court mandated to treatment. Participants' substance dependence diagnoses included cocaine dependence (71.8%), alcohol dependence (45.2%) and methamphetamine dependence (36.1%).

2.1.2. Counselor Participants—Counselor eligibility criteria included (a) credentialed to provide substance abuse services, (b) willingness to participate in the study protocol and procedures, (c) recommended by the treatment program's supervisory staff, (d) willingness to be randomized, and (e) familiarity with the 12-step orientation. There were 106 counselors at the 10 sites; 39 (37 %) met all criteria, including willingness to participate, and were included in the study pool. Two counselors per site from the pool were chosen at random to conduct the TSF treatment. Those not randomized were available for training to serve as replacement interventionists. Over the course of the study, there were four replacement counselors. Supervisors were also trained and certified as back-up interventionists. We obtained demographic information from 33 of the 34 counselors and supervisors who delivered TSF sessions; they were predominantly Caucasian (70%) women (67%) with a mean age of 52 years (SD= 9.2). Most (82%) had at least five years of counseling experience and 55% had masters degrees or above. Counselors were monitored for adherence by on-site supervisors and expert raters (4 clinicians experienced in substance abuse treatment and trained in the TSF intervention; 1 masters level, 1 doctoral candidate, and 2 doctoral level). All TSF treatment sessions were audio recorded.

2.1.3. Treatments—Participants were randomly assigned to: (a) treatment-as-usual (TAU; $n = 237$), 5–15 hours of weekly treatment as it was typically delivered by program counselors; or (b) TAU plus STAGE-12 Substitution ($n = 234$), in which manualized,

STAGE-12 TSF (Baker, Daley, Donovan, & Floyd, 2009) replaced five group and three individual TAU sessions to equate treatment hours with TAU. The five STAGE-12 group sessions shared a similar format, beginning with check-in, a review of participants' engagement in 12-step activities and recovery tasks during the prior week. Check-in was followed by presentation and interactive discussion of a different 12-step concept for each group: (a) acceptance (Step 1); (b) people, places and things; (c) surrender (Steps 2 and 3); (d) getting active in 12-step; and (e) managing emotions using 12-step tools. Materials used for STAGE-12 groups were adapted from TSF for drug abuse (Baker, 1998), and groups (Brown et al., 2002). Group sessions were augmented by three individual sessions focused on engaging participants in 12-step activities (e.g., finding a sponsor, attending meetings), encouraging use of the STAGE-12 journal, and connecting participants with volunteers active in 12-step (i.e., intensive 12-step referral; Timko, DeBenedetti, & Billow, 2006).

2.2. TSF ACES Rating Scale

The TSF ACES fidelity rating scale was based upon the STAGE-12 adherence scales which the STAGE-12 study team adapted from previous TSF fidelity measures (e.g., Project MATCH rating scale; YACS; Carroll et al., 2000) to operationalize key elements of the manualized treatment. TSF ACES has four content specific rating scales (one for groups and three for individual sessions 1–3). The items for individual sessions 2–3 vary depending upon whether or not participants reported attending 12-Step meetings in the previous week. TSF ACES items are rated on 6-point Likert scales (i.e., 1 = not at all to 6 = extensively for adherence and proscribed behavior; 1 = unsatisfactory to 6 = excellent for competence, global score and empathy). Items in TSF ACES assess adherence, competence, proscribed behaviors, overall empathy and global session performance. (See Appendix 1 for a sample of TSF items and ratings format):

Adherence Items—Content specific, adherence items assess the degree to which therapists delivered each session's prescribed content; adherence items are summarized in Table 1.

Competence items—Competence ratings are given for all adherence items (see Table 1) corresponding to a specific treatment session, with the exception of group check-in items, which are rated for competence using an overall rating, ("Overall, how well did the counselor conduct group check in?").

Proscribed therapist behaviors—We chose to identify proscribed therapist behaviors that detracted from general skill using information from the STAGE-12 Treatment Manual (Baker, Daley, Donovan & Floyd, 2009) and from STAGE-12 expert raters who had completed fidelity ratings of approximately 20% of the STAGE-12 sessions (K.M. Peavy & M. Hatch-Maillette, personal communication, December 3, 2009). We identified three behaviors that had convergence from these two sources: (a) presented material an overly structured, non-interactive manner; (b) used excessive, inappropriate or irrelevant self-disclosure; and (c) allowed the focus of the session to shift to irrelevant topics. These items are rated for every session.

Global empathy rating—The STAGE-12 treatment manual described "empathic delivery of knowledge about recovery tools" (Baker et al., 2009, p.39), as an important, non-specific, therapist behavior for the STAGE-12 intervention. TSF ACES adopted the global empathy item developed for the Motivational Interviewing Treatment Integrity (MITI) scale (Moyers, Martin, Manuel, Hendrickson & Miller, 2005). The TSF ACES item asks, "Overall, how well did the counselor understand or make an effort to grasp the clients' perspectives?" A global empathy rating is given for every session.

Global session rating—TSF ACES retained the global session rating from the original STAGE-12 adherence scales (i.e., "Overall, how well did the counselor conduct this specific session?"). This rating is given for every session.

2.3. Helping Alliance Questionnaire-II

The patient version of the Helping Alliance questionnaire-II (HAq-II; Luborsky et al., 1996) is a self-report measure of therapeutic alliance that assesses the degree to which patients experience the counselor and the treatment as collaborative and helpful. The HAq-II was chosen by the STAGE-12 study team, in part, because it demonstrated good test-retest reliability (.78), excellent internal consistency (.9), and good convergent validity (correlations of .59–.69 with the California Psychotherapy Alliance Scale; Gaston & Marmar, 1994) on its normative sample of cocaine abusers (Luborsky et al., 1996). It is a commonly used measure in alliance research (Horvath, Del Re, Fluckiger, & Symonds, 2011), particularly in research with cocaine and other drug-abusing samples (see Meier, Barrowclough, & Donmall, 2005 for a review). The instrument contains 19 items measured on a 6-point Likert scale; the sum of the items (with negative items reverse scored) forms the total score. STAGE-12 study participants completed the HAq-II at week two of treatment and week eight, end of treatment. Both were used in our analysis.

2.4. Rater Selection, Training, and Certification

We recruited raters from local university graduate programs to conduct independent fidelity ratings of all STAGE-12, TSF sessions. Nine raters were employed over the course of the study; seven held masters degrees and two had doctoral degrees. Years of direct clinical experience in substance abuse or mental health treatment ranged from 0 to 11 (mean = 4.7), and years of research experience ranged from 2 to 20 (mean = 7.1). Prior ratings experience ranged from 0 to 9 years (mean = 1.2). The expert rater was a doctoral level psychologist with extensive clinical and research experience, including prior experience developing fidelity ratings scales, conducting fidelity monitoring, training and supervising raters (e.g., Campbell et al., 2009).

Raters took part in an initial, one-day, in-person, group training session lead by the study expert rater and an expert rater from the STAGE-12 study. Prior to the training, raters reviewed the STAGE-12 intervention manual and viewed 12 hours of the video recorded, STAGE-12 clinician training. During the training, raters reviewed the STAGE-12 Adherence Manual, the TSF ACES Ratings Manual and procedures, and practiced rating audio recorded, mock sessions using the TSF ACES instrument. Raters also completed online, IRB, HIPAA and, Good Clinical Practices trainings.

After training, each rater coded one group and one individual, STAGE-12 counselor certification session that had also been rated by the study expert. Item level, inter-rater reliability (rater and expert) was calculated using a weighted Cohen's kappa coefficient (Cohen, 1968). A minimum kappa of .70 for one group and one individual session was required to achieve full certification. If the co-rated certification session fell below .70 kappa, the rater received additional training and then completed a second certification session. These procedures were repeated using subsequent certification sessions until a .70 kappa was achieved. Raters achieved group certification within 1 to 5 sessions (mean = 3.22) and individual session certification within 1 to 4 sessions (mean =3.67).

2.5. Ratings Procedures

Digitally recorded audio files of all STAGE-12 group (n=512) and individual (n=487) sessions were randomly assigned to certified raters for review. Of the 999 recorded sessions, 33 audio files were either incomplete or of poor audio quality, leaving a total of 966 rated

sessions. Bi-weekly group, conference calls were conducted by the expert with other raters, during which ratings guidelines were reviewed and clarified to improve ratings consistency. Sessions were assigned to each rater in sets of 20; within each set, one session was randomly assigned for co-rating. The sessions assigned for co-rating were blinded to both the rater and the expert; raters did not know whether a session would be co-rated, and the expert did not know the assigned rater. A weighted kappa coefficient was calculated for each co-rated session to determine agreement between rater and expert. Sessions with kappa values below .70 were jointly reviewed by the rater and expert. Rater de-certification occurred if three consecutive co-ratings fell below .70. After de-certification, rater and expert co-rated future sessions, with expert's ratings used as the study fidelity rating until rater re-certification. Raters were re-certified after achieving a .70 kappa for a co-rated session. During the study, three raters were de-certified; two achieved recertification after co-rating one session and one did not attempt recertification because the study was nearing completion.

2.6. Human Subjects' Protections

The University of California, San Francisco and Oregon Health and Science University Institutional Review Boards (IRBs) reviewed and approved the study procedures for the fidelity study. The STAGE-12 study was reviewed and approved by the University of Washington IRB, as well as IRBs of all academic institutions affiliated with participating sites.

2.7. Statistical Analysis

Cohen's weighted kappas coefficients (Cohen, 1968) were calculated for each co-rated session to measure agreement between rater and expert rater, using the expert's ratings as the standard for rater certification and performance monitoring. Cohen's kappa is considered to be a conservative measure that takes into account agreement that occurs by chance, and weighted kappas are applied when using ordinal scales to account for relative distance between scale points. Mean kappas were then calculated for each session type (i.e., group, individual 1–3, both versions).

Five TSF ACES summary measures (mean adherence, mean competence, mean proscribed behaviors, global empathy and global session rating) were calculated from ratings of each session. Global empathy and global session ratings were based on single ratings given for the whole session. Mean adherence and mean competence were calculated using all adherence /competence items for that session type. The proscribed behaviors summary measure used the mean of the three proscribed therapist behaviors rated for every session. Cronbach's alpha coefficients (Cronbach, 1951) were calculated for the multiple-item, summary measures to determine internal consistency. Because the number of adherence and competence items differs by session type, alpha coefficients for these summary measures were calculated separately for each session type; weighted, mean alphas of those summary measures for all session types were then calculated. Because the number of proscribed items (3) was consistent across sessions, one alpha coefficient was calculated for the proscribed summary measure and includes all session types. To examine inter-rater reliabilities of the five summary measures, intraclass correlation coefficients using a 2-way random mixed model (McGraw & Wong, 1996) were calculated using pairs of ratings from each of the 59 randomly selected, co-rated sessions. Pearson product-moment correlations were calculated to examine the relationships of summary measure scores with each other. Finally, we used Pearson partial correlations to assess the relationship of TSF ACES scores with HAq-II scores; average TSF ACES summary measures scores for each participant were correlated with participants' HAq-II scores collected at weeks two and eight of treatment. TSF ACES scores from the "no 12-Step meetings attended" version of individual session two (N =19)

were omitted from these correlational analyses due to anomalous alpha coefficients obtained for that version of the session.

3. Results

3.1. Reliability

Weighted kappa coefficients were calculated for 59 (31 group and 28 individual) randomly selected, co-rated sessions to certify raters and monitor raters' performance; kappas for the 59 co-rated sessions ranged from a minimum of .31 to a maximum of 1.0. The mean kappa coefficient for all co-rated sessions was .69 and ranged from .66 for ratings of group sessions to .76 for ratings of individual session two. Reliabilities for session summary measures (i.e., mean adherence, mean competence, mean proscribed behaviors, global empathy and global session rating) were calculated using intraclass correlation coefficients (ICCs). ICCs for the single-item summary measures were .69 for global empathy and .80 for the global session rating. ICCs for adherence, competence and proscribed behaviors based on mean scores per session were: adherence (.91), competence, (. 90), and proscribed behaviors (.83). Overall, ICCs indicated high reliability for the session summary measures. Cronbach's alphas assessed the internal consistency of the summary measures that were based on means of multiple items. As shown in Table 2, the weighted mean alpha for adherence ratings across session types was .69 and that for competence ratings was .71, indicating an acceptable degree of internal consistency of these items. One session type, individual session two, (version for those who had not attended a 12-step meeting since the prior session, N=19) had anomalous alpha coefficients for both adherence (−.58) and competence (.06). Adherence and competence ratings on the four items in this session (i.e., addressing barriers to treatment, encouraging a participant to attend a meeting, talking with an AA "buddy" by phone, and encouraging STAGE-12 journaling) were not consistent. In 11 of the 19 sessions, counselors covered at least one item satisfactorily while failing to cover another item or addressing it in cursory fashion, resulting in low or even negative alphas. The alpha for proscribed behaviors was also low, (.47). This modest internal consistency suggests that proscribed items may have measured different types of poor skill that do not tend to co-occur (e.g., allowing session to drift off topic versus overly structured presentation).

3.2. Relationships Between TSF ACES Summary Measures

Pearson product moment correlations were calculated to examine the relationships between the summary measures (Table 3). All correlations were statistically significant. As expected, proscribed therapist behaviors correlated negatively with every other category, and all other categories were positively correlated. Global empathy correlated significantly, but modestly (i.e., .35 to .53) with other summary measures, suggesting a moderate degree of independent information provided by this measure. The correlation for adherence and competence items was .82, indicating that these measures provided only a small amount of independent information.

3.3 Relationships of TSF ACES Summary Measures with the HAq-II

Table 4 presents correlations of HAq-II total scores at weeks two (N = 168) and eight (N=143) with TSF ACES summary measures. There were no significant correlations for HAq-II scores collected at week two. At week eight, treatment end, HAq-II scores correlated significantly with mean adherence (.31), mean competence (.28), and global session rating (.21). The correlation of HAq-II scores with global empathy (.15) approached significance (p = .0796) and the correlation with proscribed behavior, while not significant, was negative.

4. Discussion

This report describes the development of the Twelve Step Facilitation Adherence, Competence, Empathy Scale (TSF ACES), a fidelity ratings instrument for group and individual TSF. TSF ACES was developed and tested in a large, effectiveness trial conducted by community-based clinicians. Ratings were collected on approximately 97% of all group and individual TSF sessions. Initial examination of TSF ACES showed that it has good reliability. ICCs for mean adherence and competence were .91 and . 90, respectively. The summary measure with the lowest ICC was global empathy (ICC = .69). This finding is similar to the reliability of empathy on other scales such as the MITI (Moyers et al., 2005). As suggested by Madson and Campbell (2006) for the MITI, refinements that include more explicit behavioral anchors for Likert scale ratings may improve reliability of global empathy in TSF ACES. Internal consistency of the summary measures in TSF ACES that use mean scores were acceptable for adherence (.69) and competence (.71) and low for proscribed items (.47). The inconsistent adherence/competence ratings that occurred in individual session two for participants who had not attended a 12-step meeting may need to be addressed by fidelity scale modification. However, they may also reflect: (a) insufficient intervention training for that session (a version of session two that did not occur often), (b) difficulty maintaining adherence when faced with lack of participant follow through, and/or (c) session content that was not clinically coherent and was difficult to deliver. This is an example of the value of obtaining thorough treatment delivery information. It can be used to refine training and/or intervention content to improve fidelity and, ultimately, outcomes. The low internal consistency of the proscribed behavior summary measure makes clinical sense (i.e., showing poor skill by being overly didactic may run counter to showing poor skill by allowing the session to drift off topic). Perhaps proscribed behaviors should be considered individually, rather than averaged.

Initial evaluation of concurrent validity of TSF ACES was promising. Relationships of summary measures were in expected directions (e.g., proscribed items correlated negatively with all other categories) and strength; adherence and competence, which focused on the completeness and quality of content delivery, were more highly correlated with each other than they were with empathy, which focused on a nonspecific, therapeutic skill. The correlation of adherence and competence items was high (.82), although not out of range with other studies, examples of which include .62 for individual TSF (Carroll et al., 2000) and .90 for encouraging 12-step participation in individual drug counseling (Barber et al., 1996). The TSF ACES ratings manual provided guidelines for rating competence (e.g., appropriate timing; clear language when presenting specific content), but behavioral anchors for numerical ratings were not identified. In addition, raters were taught to begin with adherence ratings and move up the scale for demonstrations of high competence and down the scale for demonstrations of poor competence. This limitation may have influenced raters to rely implicitly on adherence when rating competence.

Relationships of summary measures with a participant self-report measure of therapeutic alliance (HAq-II) provided support for convergent validity. Alliance scores at treatment end correlated with TSF ACES summary measures in the expected directions (i.e., all positive except for proscribed behaviors). The significant correlations of HAq-II scores with adherence and competence replicates findings for the individual TSF scale in validation work for the YACS (Carroll et al., 2000). The positive correlation of HAq-II scores with TSF ACES global empathy approached significance, lending some support for the convergent validity of global empathy; the finding contributes to research showing a positive relationship between empathy and therapeutic alliance (see Ackerman & Hilsenroth, 2003 for a review). Conclusions about the relationship of TSF ACES ratings with therapeutic alliance may be limited by use of the HAq-II. The HAq-II was a revision of the

HAq-I that eliminated items regarding early symptom improvement, but may still reflect some confounding of alliance with improvement-based, patient satisfaction.

TSF ACES is a comprehensive scale, assessing dimensions of adherence, competence, and non-specific skill (i.e., global empathy and proscribed behaviors). As such, it can be used in clinical trials for training, monitoring, and identification of active treatment ingredients including both specific and nonspecific treatment factors. Fidelity monitoring of all these dimensions is particularly important in multisite, community-based trials, where diverse groups of counselors may implement TSF differently in the absence of training and monitoring of general skills as well as intervention content.

Community-based trials may benefit from the procedures we report for training, supervising and monitoring the performance of fidelity raters to minimize ratings drift and enhance inter-rater reliability. Factors that contributed to rater de-certification appeared to range from characteristics of some treatment sessions that made them more difficult to rate, (e.g., groups with a high number of participants or sessions conducted with very inconsistent adherence) to behavior of raters, such as attendance and participation in supervision. The procedures we employed for ratings supervision (post hoc, joint review of co-ratings and group discussion of ratings challenges) provided a rich source of information that led to refinement and clarification of ratings criteria. Effectiveness trials often use on-site supervisors to conduct ratings, a method that has been endorsed as a way to promote post-trial, clinical implementation of effective treatments (Guydish, Tajima, Manser & Jessup, 2007). Our ratings procedures can serve as a model for community-based, effectiveness trials. It is important to monitor and minimize drift among raters and supervisors in order to maintain the fidelity of treatment delivery by interventionists.

TSF is emerging as an evidence-based treatment (Manuel, Hagedorn, & Finney, 2011) that may be quite transportable because it is compatible in theoretical orientation (12-step) with treatment provided in many community programs. Transportability will be increased further by identifying "core versus adaptable" treatment components (Damschroder & Hagedorn, 2011, p. 200). This may be particularly important when implementing evidence based TSF to distinguish it from untested or idiosyncratic 12-step treatments. As a comprehensive fidelity instrument, TSF ACES can be used to verify delivery of specific treatment components in studies that seek to identify core TSF components. Moreover, the comprehensive fidelity information collected by TSF ACES will enable analysis of fidelity as a predictor of treatment retention, as a mediator of outcomes, and as a way to refine the treatment intervention to enhance its effectiveness.

Fidelity instruments such as TSF ACES can be used in clinical implementation to train and supervise counselor adherence and skill. The labor intensive ratings procedures necessary for trial integrity, however, may be impractical for monitoring implementation in treatment programs. Research supporting implementation should identify the core elements of group and individual TSF, simplify TSF ACES to capture those elements, and identify practical ways for supervisors to use the instrument.

TSF ACES showed promising reliability, concurrent and convergent validity. A major limitation of the current study was the absence of fidelity monitoring of the TAU only condition, prohibiting an evaluation of the discriminant validity of TSF ACES. In addition, there were limitations regarding the measurement of competence and non-specific skills in TSF ACES. Ratings criteria for competence items should be clarified by the addition of specific behavioral anchors for numerical ratings. This should also occur for ratings of global empathy. Proscribed items, which were included as reverse measures of non-specific, counselor skill, should be studied further to determine their meaningfulness for scale

inclusion. For example, proscribed behaviors were negatively, but not significantly, correlated with therapeutic alliance. Their relationship with patient retention and outcomes should be examined. Finally, initial validation work of TSF ACES was conducted with counselors in community treatment programs participating in the NIDA CTN. We were unable to determine how many counselors from study sites met selection criteria and were approached to participate as TSF therapists but declined, although STAGE-12 study staff believed that number to be small (S. Garrett, personal communication, March 21, 2012). Furthermore, treatment programs participating in the CTN possess some distinct features (Roman, Ducharme, & Knudsen, 2006). Generalizability of findings is limited as a result.

The exhaustive TSF treatment delivery fidelity information obtained in our study, using a reliable measure, provides a comprehensive data source for additional investigations. Future studies should examine how counselor characteristics impact TSF treatment fidelity to provide information for selection, training, and supervision of counselors. For example, does familiarity with 12-step oriented treatment facilitate learning and accurate delivery of manualized TSF? Additionally, examining the extent to which adherence, competence, general skills and their interactions with therapeutic alliance predict outcomes may illuminate otherwise mixed results regarding the relationship between treatment fidelity and treatment outcomes (Webb et al., 2010) and provide useful information regarding TSF treatment's active ingredients. This is important information for TSF implementation in community settings. Understanding the degree to which treatment components can be adapted to local settings and the need for strict versus flexible adherence can influence a treatment's appeal to programs considering adoption (Damschroder & Hagedorn, 2011). As the evidence base for group and individual TSF develops, information about factors that facilitate adoption (e.g., characteristics of counselors who learn it well; degree to which it can be adapted and still remain effective; ways to train and monitor general skills required for effective delivery) can help speed the translation of research into practice.

Acknowledgments

References

Ackerman SJ, Hilsenroth MJ. A review of therapist characteristics and techniques positively impacting the therapeutic alliance. Clinical Psychology Review. 2003; 23:1–33. [PubMed: 12559992]

Baer JS, Ball SA, Campbell BK, Miele GM, Schoener EP, Tracy K. Training and fidelity monitoring of behavioral interventions in multi-site addictions research: A review. Drug and Alcohol Dependence. 2007; 82:107–118. [PubMed: 17023123]

Baker, S. Twelve Step Facilitation for Drug Dependence. New Haven, CT: Psychotherapy Development Center, Department of Psychiatry, Yale University; 1998.

Baker S, Daley DS, Donovan DM, Floyd A. Stimulant Abuser Groups to Engage in 12 Step Programs: Evaluation of a combined group and individual treatment program. 2009 Unpublished treatment manual.

Barber JP, Krakauer I, Calvo I, Badgio PC, Faude J. Measuring adherence and competence of dynamic therapists in the treatment of cocaine dependence. Journal of Psychotherapy Practice and Research. 1997; 6:12–24. [PubMed: 9058557]

Barber JP, Mercer D, Krakauer I, Calvo N. Development of an adherence/competence rating scale for individual drug counseling. Drug and Alcohol Dependence. 1996; 43:125–132. [PubMed: 9023068]

Bellg AJ, Borrelli B, Resnick B, Hecht J, Minicucci DS, Ory M, Czajkowski S. Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. Health Psychology. 2004; 23:443–451. [PubMed: 15367063]

Borrelli B, Sepinwall D, Ernst D, Bellg AJ, Czajkowski S, Breger R, Orwig D. A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. Journal of Consulting and Clinical Psychology. 2005; 73:852–860. [PubMed: 16287385]

Borrelli B. The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials. Journal of Public Health Dentistry. 2011; 71:S52–S63.

Brown TG, Seraganian P, Tremblay J, Annis H. Process and outcome changes with relapse prevention versus 12-Step aftercare programs for substance abusers. Addiction. 2002; 97:677–689. [PubMed: 12084137]

Campbell BK. Fidelity in public health clinical trials: Considering provider–participant relationship factors in community treatment settings. Journal of Public Health Dentistry. 2011; 71:S64–S65. [PubMed: 21656955]

Campbell BK, Fuller B, Lee ES, Woelfel T, Jenkins L, Robinson J, Booth R, McCarty D. Facilitating outpatient treatment entry following detoxification for injection drug use: A multi-site test of three interventions. Psychology of Addictive Behaviors. 2009; 23:260–270. [PubMed: 19586142]

Carroll KM, Rounsaville BJ. Bridging the gap: A hybrid model to link efficacy and effectiveness research in substance abuse treatment. Psychiatric Services. 2003; 53(3):333–339. [PubMed: 12610240]

Carroll KM, Nich C, Ball SA, McCance E, Rounsaville BJ. Treatment of cocaine and alcohol dependence with psychotherapy and disulfiram. Addiction. 1998; 93(5):713–727. [PubMed: 9692270]

Carroll KM, Nich C, Sifry RL, Nuro KF, Frankforter TL, Ball SA, Rounsaville BJ. A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. Drug and Alcohol Dependence. 2000; 57:225–238. [PubMed: 10661673]

Carroll C, Patterson M, Wood S, Booth A, Rick J, Balain S. A conceptual framework for implementation fidelity. Implementation Science. 2007; 2 Retrieved from http://www.implementationscience.com.

Cohen J. Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin. 1968; 70:213–220. [PubMed: 19673146]

Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951; 16:297–334.

Daley D, Baker S, Donovan D, Hodgkins C, Perl H. A combined group and individual 12-Step facilitative intervention targeting stimulant abuse in the NIDA Clinical Trials Network: STAGE-12. Journal of Groups in Addiction & Recovery. 2011; 6:228–244. [PubMed: 22859917]

Damschroder LJ, Hagedorn HJ. A guiding framework and approach for implementation research in substance use disorders treatment. Psychology of Addictive Behaviors. 2011; 25:194–205. [PubMed: 21443291]

Gearing RE, El-Bassel N, Ghesquiere A, Baldwin S, Gilles J, Ngeow E. Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. Clinical Psychology Review. 2011; 31:79–88. [PubMed: 21130938]

Gaston, L.; Marmar, CR. California Psychotherapy Alliance Scale. In: Horvath, A.; Greenberg, L., editors. The working alliance: Theory, research and practice. New York, NY: Wiley; 1994. p. 85-108.

Guydish JR, Tajima BM, Manser ST, Jessup MA. Strategies to encourage adoption in multisite clinical trials. Journal of Substance Abuse Treatment. 2007; 32:177–188. [PubMed: 17306726]

Hogue A, Dauber S, Chinchilla P, Fried A, Henderson C, Inclan J, Reiner RH, Liddle HA. Assessing fidelity in individual and family therapy for adolescent substance abuse. Journal of Substance Abuse Treatment. 2008; 35:137–147. [PubMed: 17997268]

Horvath AO, Del Re AC, Fluckiger C, Symonds D. Alliance in Individual Psychotherapy. Psychotherapy. 2011; 48:9–16. [PubMed: 21401269]

Kaskutas LA, Subbaraman MS, Witbrodt J, Zemore SE. Effectiveness of making Alcoholics Anonymous easier: A group format 12-step facilitation approach. Journal of Substance Abuse Treatment. 2009; 37:228–239. [PubMed: 19339148]

Luborsky L, Barber JP, Siqueland L, Johnson S, Najavits LM, Frank A, Daley D. The revised Helping Alliance questionnaire (HAq-II): Psychometric properties. Journal of Psychotherapy Practice and Research. 1996; 5:260–271. [PubMed: 22700294]

Madson MC, Campbell TC. Measures of fidelity in motivational enhancement: A systematic review. Journal of Substance Abuse Treatment. 2006; 31:67–73. [PubMed: 16814012]

Manuel JK, Hagedorn HJ, Finney JW. Implementing evidence-based psychosocial treatment in specialty substance use disorder care. Psychology of Addictive Behaviors. 2011; 25:225–237. [PubMed: 21668085]

McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996; 1:30–46. & corrections to that paper in the same journal, *1*, 390.

Meier PS, Barrowclough C, Donmall MC. The role of the therapeutic alliance in the treatment of substance misuse: a critical review of the literature. Addiction. 2005; 100:304–316. [PubMed: 15733244]

Miller WR, Mount KA. A small study of training in motivational interviewing: Does one workshop change clinician and client behavior? Behavioral and Cognitive Psychotherapy. 2001; 29:457–471.

Moyers TB, Martin T, Manuel JK, Hendrickson SML, Miller WR. Assessing competence in motivational interviewing. Journal of Substance Abuse Treatment. 2005; 28:19–26. [PubMed: 15723728]

National Registry of Evidence-based Programs and Practices. 2012. Retrieved from http://www.nrepp.samhsa.gov.

Nowinski, J.; Baker, S.; Carroll, KM. Twelve-step facilitation therapy manual: A clinical research guide for therapists treating individuals with alcohol abuse and dependence (NIAAA Project MATCH Monograph Series Vol.1, DHHS Publication No. [ADM] 92-1893). Rockville, MD: National Institute on Alcohol Abuse and Alcoholism; 1992.

Perepletchikova F, Treat TA, Kazdin AE. Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. Journal of Consulting and Clinical Psychology. 2007; 75:829–841. [PubMed: 18085901]

Project Match Research Group. Matching alcoholism treatments to client heterogeneity: Project MATCH post treatment drinking outcomes. Journal of Studies on Alcohol. 1997; 58:7–29. [PubMed: 8979210]

Project MATCH Research Group. Matching alcoholism treatments to client heterogeneity: Project MATCH three-year drinking outcomes. Alcoholism: Clinical and Experimental Research. 1998; 22:1300–1311.

Robbins MS, Feaster DJ, Horigian VE, Puccinelli MJ, Henderson C, Szapocznik J. Therapist adherence in brief strategic family therapy for adolescent drug abusers. Journal of Consulting and Clinical Psychology. 2011; 79:43–53. [PubMed: 21261433]

Roman PM, Ducharme LJ, Knudsen HK. Patterns of organization and management in private and public substance abuse treatment programs. Journal of Substance Abuse Treatment. 2006; 31:235–243. [PubMed: 16996386]

Timko C, DeBenedetti A, Billow R. Intensive referral to 12-Step self-help groups and 6-month substance use disorder outcomes. Addiction. 2006; 101:678–688. [PubMed: 16669901]

Waltz J, Addis ME, Koerner K, Jacobson NS. Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. Journal of Consulting and Clinical Psychology. 1993; 61:620–630. [PubMed: 8370857]

Webb CA, DeRubeis RJ, Barber J. Therapist Adherence/Competence and Treatment Outcome: A meta-analytic review. Journal of Consulting and Clinical Psychology. 2010; 78:200–211. [PubMed: 20350031]

Appendix 1. TSF ACES Format and Sample Itemsa

1.A. Adherence and Competence Items

1.A.1 Group Session Items

    a. *Group check -in, reviewing reaction to recovery tasks* : To what extent did the group counselor review members' reactions to last week's group session recovery tasks *(meetings, readings, sponsor, using telephone to contact 12-step peers, and completing written assignments)?*

    b. *Covering group objectives*: To what extent did the group counselor cover objectives and content of group session in an interactive manner with clients?

1.A.2. Individual session Items:

    a. *Session 1, 12-step philosophy*: To what extent did the individual counselor review and discuss the 12-step Programs' philosophy of recovery, structure and terminology of meetings, and any concerns of the participant regarding participation?

    b. *Session 2, Meeting attendance*: To what extent did the individual counselor determine 12-step meeting(s) attendance since last session, and review reactions to any meeting(s)?

    c. *Session 3, Understanding of 12-step*: To what extent did the individual counselor review and compare what the participant's understanding of NA/CA/AA/CMA was prior to treatment and what it is now?

Adherence Scale:

Not At All		Somewhat		Considerably		Extensively
1		2	3	4	5	6

Competence Scale How well did the counselor handle this item?

Unsatisfactory		Fair		Good		Excellent
1		2	3	4	5	6

1.B. Proscribed Behavior Item:

1.B.1. Overly structured

To what extent did the counselor present didactic material in an overly structured, non-interactive manner?

Not At All	Somewhat		Considerably		Extensively
1	2	3	4	5	6

1.C. Global Skills Items

1.C.1. Empathy

Overall, how well did the counselor understand or make an effort to grasp the client's perspectives?

1.C.2. Global Session Rating

Overall, how well did the counselor conduct this specific session?

Unsatisfactory	Fair	Good	Excellent
1	2 3	4 5	6

[a] The full TSF ACES Rating Scale and Manual can be downloaded from the CTN Dissemination Library at http://ctndisseminationlibrary.org/PDF/795 TSFACES pdf.

Summary of Adherence Items for Group and Individual Sessions*

Session Type									
Group Session	Discuss reaction to recovery tasks (e.g., written assignments, use phone to contact 12-step peers)	Encourage 12-step participation	Discuss reactions/responses to assigned readings	Discuss urges/cravings and how they were managed	Discuss slips since last session	Encourage participation in 12-step program activities (e.g., meetings, sponsorship)	Discuss 12-step concepts[a]	Plan 12-step activities/Use 12-step "tools" to manage addiction	Assign recovery tasks
Individual Session 1	Review 12-step philosophy	Encourage 12-step meeting attendance prior to next session	Review 12-step meeting list	Encourage goals for 12-step engagement	Encourage use of 12-step "buddy"	Instruct on journal use	--	--	--
Attended 12-step Meeting Since Last Session **Individual Session 2**	Discuss 12-step attendance and reactions	Review journal use	Encourage sponsorship	Encourage additional 12-step meeting attendance	Instruct on further journal use		--	--	--
Individual Session 3	Discuss 12-step attendance and reactions	Review journal use	Encourage sponsorship	Encourage additional 12-step meeting attendance	Discuss changes in views of addiction	Discuss changes in understanding of 12-step programs	Review plans for future meeting attendance -	Discuss willingness to continue keeping a journal	Obtain feedback about treatment
Did Not Attend **Individual Session 2**	Focus on barriers to 12-step meeting attendance	Encourage 12-step attendance	Encourage use of 12-step "buddy"	Instruct on further journal use	--		--	--	--
Individual Session 3	Focus on barriers to 12-step meeting attendance	Encourage 12-step attendance	Encourage use of 12-step "buddy"	-	Discuss changes in views of addiction -	Discuss changes in understanding of 12-step programs	Review plans for future meeting attendance	Discuss willingness to continue keeping a journal	Obtain feedback about treatment

*
All items were rated on a 1-6 Likert scale (1=not at all adherent; 6=extensive adherence)

[a] Each group session covered a different concept (Acceptance, Risky people, places & things, Surrender, Getting active in 12-step, Managing emotions).

Table 2

Alpha Coefficients and Number of Items for TSF ACES Multiple-Item Summary Measures

Summary Measure	Group N=490	Individual 1 N=208	Individual 2 N=161		Individual 3 N=107		Weighted Mean N=966
			Attended n=142	Not Attd[b] n=19	Attended n=100	Not Attd[b] n=7	
Adherence	0.84 (3)[a]	0.55 (6)[a]	0.49 (5)[a]	−0.58 (4)[a]	0.81 (9)[a]	0.53 (8)[a]	0.69
Competence	0.76 (3)[a]	0.68 (6)[a]	0.59 (5)[a]	0.06 (4)[a]	0.84 (9)[a]	0.63 (8)[a]	0.71
Proscribed Behaviors							0.47 (3)[a,b]

[a]Number of items

[b]Alpha coefficient for all session types

Table 3

Pearson Product-moment Correlations for TSF ACES Summary Measures n = 947 sessions

	Adherence	Competence	Proscribed	Empathy	Global
Adherence	1.00	0.82	−0.36	0.35	0.75
		<.0001	<.0001	<.0001	<.0001
Competence	0.82	1.00	−0.46	0.50	0.83
	<.0001		<.0001	<.0001	<.0001
Proscribed	−0.36	−0.46	1.00000	−0.45	−0.57
	<.0001	<.0001		<.0001	<.0001
Empathy	0.35	0.50	−0.5	1.00	0.53
	<.0001	<.0001	<.0001		<.0001
Global	0.75	0.83	−0.57	0.53	1.00
	<.0001	<.0001	<.0001	<.0001	

Table 4

Pearson Partial Correlations for TSF Summary Measures with HAq-II Scores

HAq-II Total Score	N		Adherence	Competence	Proscribed	Empathy	Global
Week 2	168	(r)[a]	0.12	0.07	0.12	−0.01	0.04
		(p-value)	0.1209	0.3927	0.1308	0.9265	0.6232
Week 8	143	(r)[a]	0.31	0.28	−0.09	0.15	0.21
		(p-value)	0.0002	0.0008	0.2681	0.0796	0.0118

[a]Pearson partial correlation