Published in final edited form as:
J Subst Abuse Treat. 2014 September ; 47(3): 222–228. doi:10.1016/j.jsat.2014.05.008.

Treatment adherence and competency ratings among therapists, supervisors, study-related raters and external raters in a clinical trial of 12-Step facilitation for stimulant users

K. Michelle Peavy[1,2], **Joseph Guydish**[3], **Jennifer K. Manuel**[4], **Barbara K. Campbell**[5], **Nadra Lisha**[6], **Thao Le**[3], **Kevin Delucchi**[6], and **Sharon Garrett**[1]

[1]University of Washington, Alcohol & Drug Abuse Institute, Seattle, WA 98105

[2]Evergreen Treatment Services,Seattle, WA 98134

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

[4]Department of Veterans Affairs, Office of Mental Health

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

[6]University of California, San Francisco, Department of Psychiatry, Box 0984-TRC, San Francisco, CA 94143

Abstract

This study investigated the correspondence among four groups of raters on adherence to STAGE-12, a manualized 12-step facilitation (TSF) group and individual treatment targeting stimulant abuse. The four rater groups included the study therapists, supervisors, study-related ("TSF expert") raters, and non-project related ("external") raters. Results indicated that external raters rated most critically Mean Adherence - the mean of all the adherence items - and global performance. External raters also demonstrated the highest degree of reliability with the designated expert. Therapists rated their own adherence lower, on average, than did supervisors and TSF expert raters, but therapist ratings also had the poorest reliability. Findings highlight the challenges in developing practical, but effective methods of fidelity monitoring for evidence based practice in clinical settings. Recommendations based on study findings are provided.

Keywords

Treatment adherence; fidelity; 12-step facilitation; substance abuse

Corresponding author: K. Michelle Peavy, Evergreen Treatment Services, 1700 Airport Way, South, Seattle, WA 98134. peavy@evergreentx.org Phone: 206.223.3644 Fax: 206.223.1482.

1. Introduction

Evidence-based behavioral treatments (EBTs) for substance use disorders are increasingly in demand, as healthcare policies and systems focus on infusing such practices into clinical settings (Glasner-Edwards & Rawson, 2010). Therapist adherence (the degree to which a treatment is being delivered as intended; Gearing et al., 2011), and therapist competency (skill demonstrated by therapists to deliver an intervention; Waltz et al., 1993) are components of treatment fidelity, also known as treatment integrity (Borrelli, 2011). These components are necessary to determine if evidence-based interventions are accurately implemented in clinical practice. The current study examines the correspondence of treatment adherence and competence ratings among therapists, supervisors, experts, and external raters in a clinical trial of a 12-step facilitation (TSF) intervention.

1.1 The role of treatment adherence in evidence-based treatments

Accurately measuring treatment adherence is crucial for experimentally validating the treatment delivered during studies testing an intervention. As EBTs are translated into clinical practice, adherence monitoring allows for verification that the EBT has been effectively implemented (Miller et al., 2005). As noted by Garner (2009) “one of the most significant barriers to implementation research may be the lack of objective criteria for what determines when an EBT implementation has or has not occurred in practice.” (p. 394). Among the various methods of measuring treatment adherence, some studies have relied on the agency or therapist self-report of implementation (Manuel, Hagedorn & Finney, 2011); other studies have questioned the degree to which therapists are able to measure their clinical skills when learning a new EBT (e.g., Miller & Mount, 2001). Finally, treatment adherence also has important clinical implications given findings that positively associate adherence with treatment outcomes (Henggeler et al., 1997).

1.2 Measures of treatment adherence

In recent years, there has been an increase in the use of rating scales as a way to quantify therapist adherence to an EBT’s fundamental techniques and skills. Often used in conjunction with EBT training studies (e.g., Carroll et al., 2000), rating scales provide a framework to evaluate EBT training methods, EBT implementation, and a platform for ongoing supervision. While rating scales have become more widespread in recent years, questions remain about the reliability and validity of such measures (Schoenwald & Garland, 2013), as well as how these measures can be used effectively in clinical settings. In research, highly trained raters, who often have no direct contact with project therapists, evaluate adherence. Research comparing independent ratings and therapist ratings has shown that therapists rate their adherence more favorably than observers (Chapman et al., 2013; Martino et al., 2009; Carroll, Nich, & Rounsaville, 1998).These findings suggest a favorability bias about one’s own therapy skills and EBT adherence. Beidas and Kendall (2010) note that this may, in part, explain the low rates of proficient treatment adherence by therapists; however they also found that therapist proficiency increased when supervision and organizational supports were provided. Indeed, a supervisor is one step removed from the therapeutic interaction, engendering more objectivity than the therapist. While increased supervision may be one way to improve therapist adherence, this assumes that the

supervisors are adhering to the treatment, and are accurately judging their supervisees. The latter assumption was examined by Martino et al. (2009) in a randomized trial testing Motivational Enhancement Therapy (MET; Ball et al., 2007). Comparisons of therapists', supervisors', and observers' treatment integrity ratings found that supervisors rated therapists more highly than observers, suggesting a favorability bias in judging treatment adherence among supervisor raters relative to trained observers (Martino et al., 2009).

If therapists rate their own skill and adherence preferentially, and supervisors also tend to provide higher ratings to supervisees, independent raters may provide assessments of therapist adherence that are less affected by favorability bias. Additionally, "independent" raters are also often associated with the treatment study and may have a stake in how well the treatment is administered, introducing another source of potential favorability bias (Perepletchikova & Kazdin, 2005).

The current study examined therapist adherence to an EBT by four sets of raters: therapists, supervisors, project-related experts (hereafter referred to as "TSF experts"), and non-project related or "external" raters. Specifically, we examined adherence ratings from the National Institutes Drug Abuse Clinical Trials Network trial, Stimulant Abuser Groups to Engage in 12-Step (STAGE-12; Donovan et al., 2013), a group and individual 12-step facilitation (TSF) treatment targeting stimulant abusers. During the STAGE-12 clinical trial, therapists, supervisors and TSF experts rated adherence to monitor treatment fidelity to the manualized intervention. Following the STAGE-12 trial, external raters also conducted adherence ratings as part of an independent study. The current study compared these four sets of adherence ratings. While group differences between rater may be due to the different perspective each rater category holds, we hypothesized that therapists would rate treatment adherence of their own sessions most favorably (the highest ratings), that ratings from other sources would be progressively lower as they became more distant from the therapist and the clinic, and less invested in therapists' successful adherence. Thus, we expected that therapists' ratings would be greater than supervisors' ratings, which would be greater than TSF expert ratings, and these in turn would be greater than the external ratings.

2. Method

2.1 Overview of STAGE-12 and Ancillary Fidelity Studies

The STAGE-12 study was a multisite randomized controlled trial comparing the STAGE-12 TSF intervention to treatment-as-usual, in a sample of adults receiving treatment for stimulant abuse or dependence in 10 outpatient programs (N=471; see Donovan et al., 2013). The STAGE-12 intervention was comprised of 8 sessions delivered over a 5-8 week period. The intervention included 5 group sessions based on the Project MATCH TSF manual (Baker, 1998; Carroll et al., 1998) adapted to a group format (Brown et al., 2002a; Brown et al., 2002b), and 3 individual sessions based on the first and last sessions from the Project MATCH TSF manual, with the addition of a 12-step intensive referral procedure (Timko et al., 2006). All STAGE-12 TSF sessions were digitally, audio-recorded for adherence monitoring by therapists, clinic supervisors and TSF experts. Treatment-as-usual sessions were not monitored for fidelity. The STAGE-12 study was approved by the University of Washington Institutional Review Board (IRB), as well as IRBs of all academic

institutions affiliated with participating sites. Participants in the STAGE-12 TSF condition, as compared to TAU, were more likely to be abstinent from stimulants during treatment but not at any subsequent assessment. TSF participants had significantly lower Addiction Severity Index (ASI; McLellan et al., 1992) Drug Composite scores at 3 months, and also had higher rates of 12-step attendance and involvement throughout the follow-up period (see Donovan et al., 2013).

Following the STAGE-12 trial, all TSF sessions were evaluated for adherence, competence and empathy by raters not affiliated with the STAGE-12 trial (i.e., external raters) using a ratings instrument based upon an expansion of the STAGE-12 TSF adherence scales (Campbell et al., 2013). The University of California, San Francisco and Oregon Health and Science University IRBs approved the procedures for this fidelity study. Greater competence and empathy in the delivery of the intervention were associated with better drug use and employment outcomes, while adherence was associated with better employment outcomes only (Guydish et al, 2014). In summary, there were four sets of STAGE-12 TSF adherence ratings; those conducted by therapists, supervisors, and TSF experts in the STAGE-12 parent study and those conducted by external raters in the fidelity study.

2.2 Ratings Participants: Therapists, Supervisors, TSF Expert, and External Raters

Therapist eligibility included being credentialed, completing a therapist assessment, and willingness to be randomly assigned to provide either the STAGE-12 intervention or usual care (the control condition). Across all 10 sites, 106 therapists met eligibility criteria; 20 were randomly selected within site (2 per site) to provide the STAGE-12 intervention. Additional therapists (n=4) were recruited as needed through the course of the trial (Donovan et al., 2013). Therapists self-identified as 58% White, 17% African American, and 17% multi-racial. In terms of education, 37% reported having a Masters degree, Bachelor (25%), Associates (25%) and high school graduation (< 1%). Mean age was 52 (SD = 7.1), most (83%) were licensed or certified in their field, and they had on average 11 years (SD = 6.7) of substance abuse counseling experience.

One supervisor at each of the 10 sites supervised the STAGE-12 intervention, and 4 supervisors were added as replacements when needed. Most supervisors self-identified as White (71%), and one supervisor identified in each category of African American, multi-racial, “other”, and unreported. Twenty-eight percent of supervisors reported having a Doctorate, 57% a Masters degree, and one supervisor each reported high school education or missing data. Mean age was 48 (SD = 12.2), most (93%) were licensed or certified in their field, and they had an average of 14 years (SD = 6.0) of substance abuse counseling experience.

TSF expert raters were 4 therapists experienced in substance abuse treatment and trained in the TSF intervention. One had a master’s degree, two were doctoral candidates, and one was a licensed clinical psychologist. External raters were recruited from local university graduate programs to conduct independent fidelity ratings of STAGE-12 sessions for the fidelity study (Campbell et al., 2013). Nine raters were employed over the course of the study; seven held masters degrees and two had doctoral degrees. Years of experience in substance abuse or mental health treatment ranged from 0 to 11 (mean = 4.7). An expert rater, not affiliated

with the STAGE-12 trial, co-rated approximately 6% of the sessions to monitor reliability of the external raters' scores. The expert rater was a doctoral level clinical psychologist with extensive experience training raters and conducting fidelity monitoring. Training procedures for therapists, supervisors and TSF expert raters are reported in Donovan et al., (2013); training and reliability procedures for external raters are described in Campbell et al. (2013).

2.3 Measures

During the STAGE-12 trial, therapists, supervisors and TSF expert raters used a TSF adherence measure developed for the STAGE-12 study. Adherence items assessed the degree to which the therapist delivered content prescribed for a specific session (e.g., "To what extent did the individual counselor encourage the participant to identify and agree to attend specific 12-step meetings prior to the next individual session"). There were four content rating forms, one used for all group sessions and three corresponding to STAGE-12 individual sessions 1, 2 and 3. Adherence items were rated using a 7-point Likert scale (1 = not at all to 7 = extensively). The adherence scale designated for group sessions contained 12 items. The number of adherence items differed for each individual session, ranging from 3 to 9 items, due to differences in session content. In addition to adherence items corresponding to the content of each session, ratings for all sessions included a global session item: "Overall, how well did the counselor conduct this specific session?" This was the sole item designed to tap therapist competence. Coding on this item encompassed therapist adherence insofar as therapists could not obtain a high score on the global session item if they did not demonstrate adherence to the intervention. Additionally, raters were asked to consider the "overall" performance, and were provided scoring instruction for guidance.

External raters used the Twelve Step Facilitation Adherence Competence Empathy Scale (TSF ACES: Campbell et al., 2013). The TSF ACES used the same adherence items and global session rating as those used in the STAGE-12 clinical trial, but included additional ratings of competence to accompany each adherence item ("How well did the counselor handle this item?"), proscribed behaviors (behaviors inconsistent with the counseling protocol), and a Global Empathy item adopted from the Motivational Interviewing Treatment Integrity (MITI) scale (Moyers et al., 2005). TSF ACES items were 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). Inter-rater reliabilities for the summary measures derived for each session (i.e., mean adherence, mean competence, mean proscribed behaviors, global empathy, global session rating) assessed with intraclass correlations were as follows: .91 for mean adherence, .90 for mean competence, .83 for mean proscribed behaviors, .80 for the global session rating and .69 for global empathy. Internal consistencies computed with Cronbach's alpha for summary measures that are based on multiple items were acceptable for mean adherence (.69) and mean competence (.71) and low for proscribed behaviors (.47). A version of individual session #2 conducted with participants 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). We eliminated ratings of this session from our analysis due to the poor internal consistency of adherence ratings for the session. Relationships between the

summary measures with each other were all in the expected directions (e.g., proscribed behaviors were negatively correlated with all other measures). The TSF ACES Rating Manual and forms can be found at http://ctndisseminationlibrary.org/PDF/795_TSFACES.pdf, and psychometric properties are reported in further detail in Campbell et al. (2013).

There were two summary measures in common for the STAGE-12 study ratings and the external ratings: the Mean Adherence rating for each session and the Global Session Performance rating ("Overall, how well did the counselor conduct this specific session. These two measures were used in the current analysis across the four sets of raters.

2.4 Ratings Training

Therapists and supervisors did not receive formal training in the ratings procedure. The ratings instruments were introduced during the therapist training, which was required of supervisors as well. Therapists and supervisors were provided with a coding manual that described how scores were assigned to each item. The coding manual was developed jointly with the author of STAGE-12, individuals from the research team, and some of the TSF expert raters. Both therapists and supervisors received feedback about their performance either during the counselor certification process, or when a session was deemed unpassable by one of the TSF expert raters. Feedback was provided in written format with suggestions about how to improve future sessions. The feedback would often reference the coding manual; neither therapists nor supervisors co-rated sessions with TSF expert raters.

The TSF expert raters also attended the STAGE-12 TSF therapist training, or watched the training video. During the first weeks of the study, there was only one designated rater, supervised by a clinical psychologist not charged with rating, but with extensive experience in clinical trials and fidelity monitoring. Together with the research team and developers of STAGE-12, the coding manual was developed. As other raters were added, the first rater and supervisor had an initial meeting with new raters to discuss ratings procedures and the coding manual. Subsequently, the expert raters attended monthly conference calls to discuss ratings issues and review the co-rated sessions. Two randomly selected sessions (1 group, 1 individual) were rated by all raters prior to the calls and discussed during the calls. This procedure lasted throughout data collection period, and resulted in a total of 28 sessions being co-rated.

External raters viewed the STAGE-12 TSF therapist training video and completed a one-day ratings training. Raters achieved a criterion level of inter-rater reliability with a ratings expert on audio-recorded, practice sessions conducted by STAGE-12 therapists to become certified. Audio recordings of TSF sessions were then randomly assigned to certified raters in sets of 20; one session per set was randomly assigned to the study expert for co-rating to monitor ratings consistency. See Campbell et al. (2013) for more detail.

2.5. Procedures

2.5.1 Ratings—During the STAGE-12 trial, STAGE-12 therapists self-rated each session after its completion; supervisors would conduct occasional TSF sessions when clinic

therapists were absent (e.g., vacation, illness). To obtain mutually exclusive rater groupings, we removed 86 ratings of supervisor-conducted sessions because they had self-rated the sessions as therapists, leaving a total of 919 therapist self-ratings. STAGE -12 clinic supervisors and TSF expert raters rated a random selection of approximately 35% of the audio-recorded TSF sessions. However, supervisors did not complete all ratings assigned to them. As a result, there were 362 TSF expert ratings and 329 supervisor ratings. Table 1 shows the number of raters for every group, along with the number of sessions each group rated. Supervisors did not co-rate sessions with expert raters, nor did supervisors or therapists receive any feedback from experts regarding their ratings. As described above, therapists and supervisors did receive written feedback on expert rated sessions regarding therapist performance that had been scored as unpassable.

The external study team rated 966 TSF session audio recordings that were both audible and complete (Campbell et al., 2013). Ratings from individual session #2 which had produced anomalous alpha coefficients (N=19) were excluded from analysis, leaving 947 external session ratings.

2.5.2 Matching ratings across sources and aggregating by participant—To enable a direct comparison between ratings from different sources for the same therapy sessions, we matched ratings across their four sources, resulting in a total of 277 audio recordings rated by a therapist, a supervisor, a TSF expert rater, and an external rater.

These audio recordings included both individual and group sessions. A single rating for a single group session can apply to more than one participant. Consequently, the 277 matched audio recordings translate to ratings for 324 participant sessions, received by 160 participants, and with an average of 2.03 (sd = 1.09) ratings per participant.

2.6 Analysis Plan

Two rating values were calculated or available for each session. The first was the mean of the adherence items for a given session (Mean Adherence), and the second was the single item, Global Session Performance rating for each session. Means were calculated using all session ratings available for each participant across the four sets of raters. Therapist, supervisor and TSF expert raters rated these items using a scale from 1 (Not at all) to 7 (Extensively), while external raters in the ancillary fidelity study, used a scale ranging from 1 (not at all) to 6 (extensively; Campbell et al., 2013). To ensure comparability across rating groups, adherence ratings conducted by therapists, supervisors and TSF expert raters were recoded by collapsing ratings of 6 or 7 to 6 (extensively) to give the same 6-point rating scale across all rating groups.

There are three potential levels of nesting in these data: nesting of sessions within a single participant, nesting of participants within a single therapist, and nesting of therapists within a single clinic or site. At the level of sessions within participant, we took the mean of each measure across all sessions received by a single participant. This gives a single Mean Adherence rating and a single Global Session Performance rating for each participant and each rating group. Aggregating all sessions per participant into mean values also allowed us to reduce the levels of nesting necessary in the analytic model by one level.

Participants were assigned to a primary therapist, with the intention that the same therapist would provide the three individual STAGE-12 sessions. The therapist leading the group was not the primary therapist for all participants in a given group. For the 160 participants in the analysis, 22.5% received sessions delivered by more than one counselor, so that nesting participants within counselor was not possible. However, for those participants who had only one therapist, we calculated intraclass correlation coefficients (ICCs) to assess interdependence of session scores for sessions provided by the same therapist on different participants (ICC = .0014). This suggests that the level of overall rating shown by a therapist does not vary by participant or, stated another way: therapists show the same level of competence regardless of the participant. As the ICC was essentially 0, we did not attempt to control for nesting of participant within therapist.

By taking the average fidelity score (either global or adherence) across all sessions received by a participant, we eliminated nesting within participant, leaving 4 rater types and 160 participants. As clients were always seen in the same clinic (site), we set the model so that participants are nested within site, and we included site as a factor in the model to control for site differences. The analytic model, then, included fixed effects for rater type (therapist, supervisor, TSF expert, external) and random effects for subject.

ICCs were calculated to examine the inter-rater reliabilities (Shrout & Fleiss, 1979) for Mean Adherence and Global Session Performance. The four rating groups (therapist, supervisor, TSF expert, external) were compared to the expert rater who coded 59 randomly selected sessions for the fidelity study. The number of sessions coded by the expert overlapped to varying degrees with the four groups (N = 52 therapist, N = 23 supervisor, N = 24 TSF expert, and N = 59 external). In order to account for clustering within the four rating groups, the ICC was calculated using additional random effects in both numerator and denominator. For the Global Session Performance, and because global fidelity is based on a single item, a model with random session and rater within session effects was used to estimate the ICC. For Mean Adherence, and because Mean Adherence is based on a different number of items depending on the type of session, ICCs were calculated slightly differently. A random item (reflecting both session and item within session) was added to both the numerator and the denominator as is the case in ICC calculations for random effects. Both models used a fixed effects model of intercept only.

3. Results

3.1 Rating Stability, Practice Effects, and Interdependence of Group Ratings

Before comparing ratings across the four rater groups, we assessed rating stability, practice effects and interdependence of group ratings. First, because raters provided ratings on multiple sessions over time, on multiple participants, we assessed whether ratings of a single individual on a single adherence indicator appear stable. We examined the stability of the Global Session Performance ratings using a standard linear model for repeated measures, and testing for significant change over time within person, using ordering of sessions. No significant effects for time were found (p = .91) indicating that the ratings were stable over time (meaning, across all sessions received by that single participant).

Second, we assessed practice effects on the scoring of sessions, as the number of sessions rated by therapists and external raters is similar (916 and 947 respectively), and almost three times the number rated by supervisors and TSF expert raters (329 and 362 respectively). Using a standard linear model for repeated measures, we tested for differences in change over time (ordered sessions received by the same participant) as a function of type of rater (therapist, supervisor, TSF expert, external). We found no interaction between rater type and time ($p = .12$) indicating that the level of the stability of ratings did not differ by type of rater.

Third, as group sessions included more than one participant, one group rating would apply to all participants in the group (ratings are perfectly correlated for participants in that group). However, participant fidelity ratings used in the analysis were based on the mean of all group and individual sessions attended by each participant. To assess how interdependence of group ratings may affect participant means, we identified the pattern of sessions rated for each participant. As there are 3 individual and 5 group sessions, the mean fidelity rating for a single participant may be based on any combination of these 8 sessions. For 160 participants in the analysis we identified 71 patterns of sessions included in the participant mean. This suggests that participant mean fidelity scores were substantially independent, despite some instances where participants attended the same group.

3.2 Reliability Ratings

For Mean Adherence there was a larger range of ICCs (.39 for therapist, .55 for supervisor, .62 for TSF expert, and .79 for external). For Global Session Performance, agreement was fair across groups (.57 for therapist, .61 for supervisor, .61 for TSF expert and .61 for external).

3.3 Adherence and Competency Comparisons

Least square means and standard errors (SE) derived from the mixed effects model are depicted in Figure 1. For the Mean Adherence outcome, the therapist rating (Mean = 5.26, SE= .045), was lower than the supervisor (Mean = 5.52, SE=.045) and TSF expert ratings (Mean = 5.57, SE=.045), and higher than the external ratings (Mean = 5.14, SE=.045). For the Global Session Performance rating, the therapist rating (Mean = 5.70, SE= .046) was higher than supervisor (Mean = 5.59, SE=.047), TSF expert (Mean = 5.61, SE=.046), and external ratings (Mean = 4.93, SE=.046).

Findings from individual contrasts between each set of mean values are shown in Table 2, with Mean Adherence contrasts listed on the left half of the table and Global Session Performance contrasts listed on the right. The first three contrasts comparing therapist ratings to other ratings of Mean Adherence are statistically significant, but the directionality differs by contrast. Specifically, therapist ratings were significantly lower than either supervisor (estimate = −.264, SE = .057) or TSF expert (estimate = −.317, SE = .057). This is counter to our expectation that therapists would rate their own adherence higher than either supervisors or TSF expert raters. At the same time, therapists did rate their own adherence higher than did external raters (estimate = .117, SE= .057). Finally, Mean Adherence ratings by external raters were significantly lower than either supervisor

(estimate = .381, SE= .057) or TSF expert (estimate = .434, SE= .057) ratings. Global Session Performance ratings (also presented in Table 2) were significantly higher for therapists than external raters (estimate = .7782, SE = .060, p <.001). Both supervisor (estimate = .665, SE= .060, p < .001) and TSF expert ratings of Global Session performance (estimate = .681, SE= .060, p < .001) were significantly higher than external ratings.

4. Discussion

This study compared treatment adherence and global performance ratings across four groups of raters in a community based, multi-site trial of 12-step facilitation: counselor self-ratings, supervisors, TSF expert, and external raters. This is the first study, to our knowledge, that compared across these different sets of raters. The addition of independent, external raters in our study afforded us the opportunity to "observe the observers" and compare ratings across groups with varying degrees of independence from the administered treatment. Reliability calculations indicated that therapists exhibited poor reliability, and they were less reliable than other groups. In this study, external ratings were significantly lower than therapist, supervisor and TSF expert ratings for Mean Adherence and Global Session Performance. Ratings by supervisors and TSF expert raters did not differ significantly. Therapist self-ratings were significantly lower than supervisor ratings for Mean Adherence and were marginally higher than supervisor ratings for Global Session Performance. Therapist ratings were also less reliable than ratings by the other groups.

The reliability results help us understand and better put into context the comparisons between rater groups. Global Session Performance ratings were similarly reliable across all groups of raters when compared to a "standard" of an expert rater not affiliated with the main trial. However, Mean Adherence ratings showed a wide range of reliability, and rater reliability of therapists was particularly low. Such a result is not surprising given that the therapists did not have exposure to ratings procedures, and they had less practice at coding overall. While supervisors were also untrained in ratings procedures, they listened to a variety of therapists' sessions, giving them more practice at coding, as well as a broader comparison about the range in quality of treatment adherence. Adequate training in ratings procedures might ameliorate inconsistencies between self-ratings and expert raters, and training could be considered as an option for improving self-ratings of treatment adherence. The poor rater reliability of therapists affects interpretation of this group's ratings results, and limits our ability to comment upon the correspondence of their ratings with the other groups.

The finding that external raters were more critical than therapists and supervisors is consistent with prior research (Martino et al., 2009; Carroll, Nich, & Rounsaville, 1998). External raters also provided ratings more critical than TSF expert raters. While this kind of comparison has not been tested in the past (study-designated fidelity ratings vs. fidelity ratings from a non-study team), such an outcome has several possible explanations. External raters may have less favorability bias because they had no ongoing relationship with those being rated, or due to more rigorous rater training and supervision than that used by counselors, supervisors and TSF expert raters (see Campbell et al., 2013). Finally, external raters were more comprehensive in their ratings, examining competence, empathy and

proscribed behavior, in addition to adherence and global performance. Perhaps this led raters to focus on additional target behaviors, resulting in more thorough and more critical ratings. It may be helpful to note that, while differences between means were statistically significant, they were modest in absolute terms (5.57 TSF expert vs. 5.14 external for Mean Adherence; 5.61 TSF expert vs. 4.93 external for Global Session).

Previous literature indicates therapists and supervisors tend to rate treatment adherence more favorably than expert raters (Martino et al., 2009). Our results were consistent with this finding for only one comparison: therapists rated themselves more favorably than did supervisors on Global Performance. However, supervisor and TSF expert ratings did not differ from each other and both were more favorable than counselor ratings for Mean Adherence. It is useful to consider the distinction between Mean Adherence and Global Session Performance, given the pattern of results. The Global Performance measure includes assessment of overall competence along with adherence, involving a higher level of inference, thus introducing an increased likelihood of rater bias (Hoyt, 2000). Relatively higher therapist ratings on Global Performance may reflect seeing oneself as doing well overall, or as possessing good counseling skills generally. However, interpretation of therapist ratings should be tempered with the finding that therapists showed poor reliability with the expert rater.

Therapists' lower self-ratings on Mean Adherence, relative to supervisors and TSF expert raters, may reflect counselors' self-perceived difficulties in adhering to the manualized content. They may have felt lower self-efficacy in content adherence than they did in overall performance. Alternatively, supervisors and TSF expert experts may have had a leniency bias in their judgment of Mean Adherence. Supervisors tend to form leniency bias in ratings of supervisees due to the working relationships that form between supervisor and supervisee (Gonsalvez & Freestone, 2007). In addition, both supervisors and TSF expert raters may have inadvertently rated sessions more generously (i.e., "give the benefit of the doubt") in order to provide feedback that kept counselors motivated to continue participating in the study. Furthermore, supervisors and TSF expert raters may have been motivated to leniently rate sessions in order to demonstrate adequate treatment adherence during the trial and avoid losing counselors due to loss of proficiency.

The lack of differences in ratings by supervisors and TSF expert raters may be explained by their non-independence, since these groups had ongoing contact including shared information about ratings. For example, TSF expert raters provided written summaries for most sessions that fell below expected performance levels, including recommendations on how to improve treatment adherence. These directives were vetted through the supervisor, exposing both the supervisor and therapist to the feedback. While therapists may have had access to the feedback, the supervisor was in essence acting as, and be being trained as an *observer* by the TSF expert, and therefore one might expect ratings between these groups to be similar.

4.1 Limitations and future directions

A primary strength of this study lies in the rare opportunity for external ratings to be included in the ratings comparisons. The adherence ratings scale was one that had

demonstrated initial reliability and validity (see Campbell et al., 2013). This study was conducted in community-based treatment programs and utilized counselors already employed in these settings potentially limiting the generalizability of findings given some distinct features of programs participating in the CTN (Arfken et al., 2005). The current study employed a relatively large sample size of ratings; almost 100% of sessions had self-ratings by counselors and ratings by the external raters. However, the set of ratings available from all four categories of raters was limited to about 35% of all sessions. Calculating reliability with the expert rater included an even smaller subset, such that the overlap between expert rater and all four of the rater groups was only 16 sessions, constituting a limitation in interpreting the reliability results. Interpretation of study findings is also limited due to several differences in ratings procedures across the ratings (e.g., self-ratings were based on recollection of performance after a session ended, while other raters listened to digitally recorded sessions). There were differences in ratings training, including a lack of assessment of ratings competence for therapists and supervisors during the STAGE-12 study. Because training can provide more acuity in terms of understanding the intervention, as well as the ratings process, this difference is important and may contribute substantially to group differences. Additionally, there was no measure of how the rater groups compared in terms of basic knowledge or understanding about the treatment protocol. Adherence scales were slightly different (i.e., same Likert descriptive anchors but different scale lengths) between the STAGE-12 raters and the external raters; differences that could affect how raters used the scales. All of these differences may have influenced differences in ratings outcomes. Another limitation should be noted regarding the rating scale itself. Specifically, therapist competence was only captured by one item (Global Session Performance), and therefore this construct may not be measured as accurately and comprehensively as adherence.

Procedural differences and differences in ratings across rater groups underscore the challenges of developing practical and effective methods of fidelity monitoring of EBTs in clinical practice. While funding and licensing agencies increasingly call for use of EBTs, community-based organizations implementing them will seek the simplest, most reliable and cost-effective methods of monitoring EBT delivery. For research purposes, results suggest that raters unaffiliated with the treatment being tested may provide the most objective ratings, or ratings that are most reliable with an expert rater and least impacted by favorability bias. For clinical practice, results suggest that there may be a role for on-site therapists or supervisors rating adherence. While there were statistically significant differences among raters on the Mean Adherence and Global Session ratings, these differences may not be regarded as clinically meaningful among frontline clinical providers (Miller & Manuel, 2008). For instance, the largest mean difference on the Mean Adherence rating was .43 and between the expert rater and the external raters (Means = 5.57 and 5.14 respectively) and .77 on the Global Session Performance rating between therapists and external ratings (Means = 5.70 and 4.93 respectively). Future research that employs more rigorous training for therapists on adherence monitoring may better reveal whether therapist self-ratings are suitable for monitoring adherence to EBTs. Based on our experience, the recommendations for employing therapists in the self-rating and treatment adherence process are as follows: 1) rating scales used in clinical settings should measure specific

behaviors; 2) therapists should undergo adequate training on the ratings procedures, the more inferential the measure is, the more training should be required; and 3) this should be supported by ongoing clinical supervision that focuses on both adherence and competence in the EBT.

Future research should examine the impact of training therapists on self-rating to determine whether this group can achieve acceptable reliability and objectivity in ratings. Researchers might also examine the process of supervision, and determine its impact on treatment adherence. To do so, appropriate supervision measures would need to be developed. Treatment adherence and its relation to participant outcomes was not a focus of the current paper (see Guydish et al, 2014); however, a related research question is whether specific rater groups differ in terms of predicting treatment outcomes.

Acknowledgments

This work was supported by the National Institute on Drug Abuse (R01 DA025600), by the NIDA Clinical Trials Network – Pacific Northwest Node (U10 DA013714) and Western States Node (U10 DA015815), and by the NIDA San Francisco Drug Abuse Treatment Research Center (P50 DA009253). The authors would like to thank all the individuals who conducted ratings of the STAGE-12 intervention audio files, as well as the STAGE-12 lead node, in particular Dennis Donovan and Suzanne Doyle, who assisted in answering numerous questions regarding the protocol and accompanying ratings data. We are grateful to Alan Bostrom and Mable Chan, who assisted in early phases of data analysis for this study.

Referecnes

Arfken CL, Agius E, Dickson MW, Anderson HL, Hegedus AM. Clinicians' beliefs and awareness of substance abuse treatments in research- and nonresearch affiliated programs. Journal of Drug Issues. 2005; 35(3):547–558.

Baker, S. Twelve step facilitation for drug dependence. Psychotherapy Development Center, Department of Psychiatry; Yale University, New Haven, CT: 1998.

Ball SA, Martino S, Nich C, Frankforter TL, Van Horn D, Crits-Christoph P, et al. Site matters: Motivational enhancement therapy in community drug abuse clinics. Journal of Consulting and Clinical Psychology. 2007; 75:556–567. [PubMed: 17663610]

Beidas RS, Kendall PC. Training Therapists in Evidence-Based Practice: A Critical Review of Studies From a Systems-Contextual Perspective. Clinical Psychology-Science and Practice. 2010; 17(1):1–30. [PubMed: 20877441]

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. 2002a; 97:677–689. [PubMed: 12084137]

Brown TG, Seraganian P, Tremblay J, Annis H. Matching substance abuse aftercare treatments to client characteristics. Addictive Behavior. 2002b; 27:585–604.

Campbell BK, Manuel JK, Turcotte Manser S, Peavy KM, Stelmokas J, McCarty D, Guydish J. Assessing fidelity of treatment delivery in group and individual 12-step facilitation. Journal of Substance Abuse Treatment. 2013; 44:169–176. doi: 10.1016/j.jsat.2012.07.003. PMID: 22944595. [PubMed: 22944595]

Carroll KM, Nich C, Rounsaville BJ. Utility of therapist session checklists to monitor the delivery of coping skills treatment for cocaine abusers. Psychotherapy Research. 1998; 8:307–320.

Carroll KM, Connors GJ, Cooney NL, et al. Internal validity of Poject MATCH Treatments: Dicriminability and integrity. Journal of Consulting and Clinical Psychology. 1998; 66:290–303. [PubMed: 9583332]

Carroll KM, Nich C, Sifry RL, Nuro KF, Frankforter TL, Ball SA, Fenton L, 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]

Chapman JE, McCart MR, Letourneau EL, Sheidow AJ. Comparison of youth, caregiver, therapist, trained, and treatment expert raters of therapist adherence to a substance abuse treatment protocol. Journal of Consulting and Clinical Psychology. 2013; 81(4):674–680. [PubMed: 23668668]

Donovan DM, Daley DC, Brigham GS, Hodgkins CC, Perl HI, Garrett S, Doyle S, Floyd AS, Knox PC, Botero C, Kelly T, Killeen T, Hayes C, Baumhofer NK, Seamans C, Zamarelli L. Stimulant abuser groups to Engage in 12-Step (STAGE-12): A multisite trial in the NIDA Clinical Trials Network. Journal of Substance Abuse Treatment. 2013; 44(1):103–114. doi: 10.1016/j.jsat.2012.04.004. [PubMed: 22657748]

Garner BR. Research on the diffusion of evidence-based treatments within substance abuse treatment: A systematic review. Journal of Substance Abuse Treatment. 2009; 36(4):376–399. [PubMed: 19008068]

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

Glasner-Edwards S, Rawson R. Evidence-based practices in addiction treatment: Review and recommendation for public policy. Health Policy. 2010; 97(2-3):93–103. [PubMed: 20557970]

Gonsalvez CJ, Freestone 1,J. Field supervisors' assessments of trainee performance: Are they reliable and valid? Australian Psychologist. 2007; 42(1):23–32.

Guydish J, Campbell BK, Manuel JK, Delucchi K, Le T, Peavy M, McCarty D. Does Treatment Integrity Predict Client Outcomes in 12-Step Facilitation for Stimulant Abuse? Drug and Alcohol Dependence. 2014; 134(1):330–336. [PubMed: 24286966]

Henggeler SW, Melton GB, Brondino MJ, Scherer DG, Hanley JH. Multisystemic therapy with violent and chronic juvenile offenders and their families: The role of treatment fidelity in successful dissemination. Journal of Consulting and Clinical Psychology. 1997; 65(5):821–833. [PubMed: 9337501]

Hoyt WT. Rater bias in psychological research: When is it a problem and what can we do about it? Psychological Methdos. 2000; 5(1):64–86.

McLellan AT, Kushner H, Metzger D, Peters R, Smith I, Grissom G, Argeriou M. The fifth edition of the Addiction Severity Index. Journal of Substance Abuse Treatment. 1992; 9:199–213. [PubMed: 1334156]

Manuel JK, Hagedorn HJ, Finney JW. Imlementing Evidence-Based Psychosocial Treatment in Specialty Substance Use Disorder Care. Psychology of Addictive Behaviors. 2011; 25(2):225–237. [PubMed: 21668085]

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

Miller WR, Manuel JK. How large must a treatment effect be before it matters to practitioners? An estimation method and demonstration. Drug and Alcohol Review. 2008; 27(5):524–528. [PubMed: 18608445]

Miller WR, Mount KA. A small study of training in motivational interviewing: Does one workshop change clinician and client behavior? Behavioural and Cognitive Psychotherapy. 2001; 29(04):457–471.

Miller WR, Zweben J, Johnson WR. Evidence-based treatment: Why, what where, when and how? Journal of Substance Abuse Treatment. 2005; 29(4):267–276. [PubMed: 16311179]

Moyers TB, Martin T, Manuel JK, Hendrickson SM, Miller WR. Assessing competence in the use of motivational interviewing. Journal of substance abuse treatment. 2005; 28(1):19–26. [PubMed: 15723728]

Perepletchikova F, Kazdin AE. Treatment integrity and therapeutic change: Issues and research recommendations. Clinical Psychology: Science and Practice. 2005; 12(4):365–383.

Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological bulletin. 1979; 86(2):420. [PubMed: 18839484]

Schoenwald SK, Garland AF. A review of treatment adherence measurement models. Psychological Assessment. 2013; 25(1):146–156. [PubMed: 22888981]

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

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(4):620–630. [PubMed: 8370857]

Highlights

- We examine treatment adherence and competency of a 12-step facilitation treatment.
- We compare ratings between therapists, supervisors, expert, and external raters.
- The external raters rated most critically on all measures (adherence and competency).
- Therapists exhibited poor reliability with the designated expert.
- Raters more removed from the research process appear to provide less biased ratings.

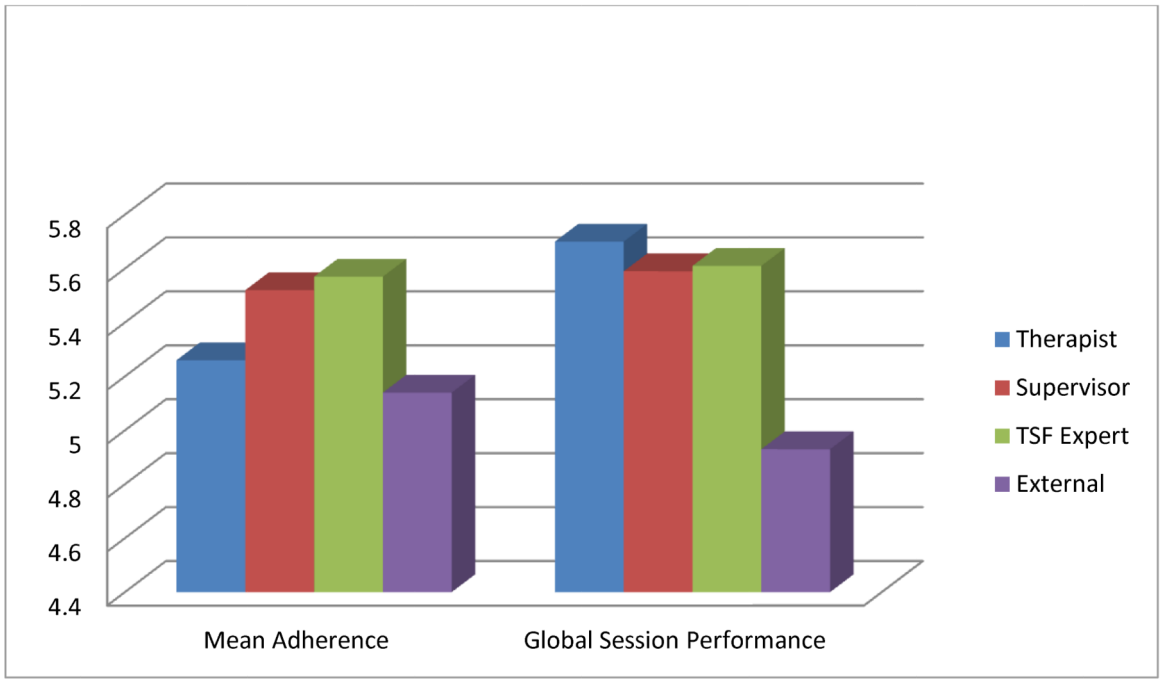

Figure 1. Least Square Means Comparison of Rater Groups

*Mean Adherence is the mean of 3-9 adherence items, depending on the content of each session. Global Session Performance is the rating of a single item: "Overall, how well did the counselor conduct this specific session?" The outcome measure is the mean of ratings (Mean Adherence, Global Session Performance) for all sessions delivered to a single patient, and the mean of those patient-level ratings aggregated by Rating Group.

Table 1

Summary of number of raters and number of sessions rated in each of four rating groups.

	Participants (N)	Sessions Rated (N)*
Supervisors	14	329
TSF Expert raters	4	362
Therapists	24	916
External raters	9	947

* Number of sessions after removal of corrupted files, duplicate ratings, and ratings by the same person as Therapist and Supervisor.

Table 2

Contrasts for Least Square Means across Rating Groups for measures of Mean Adherence and Global Session Performance*

	Mean Adherence				Global Session Performance			
Contrast	**Estimate**	**SE**	**T**	**P**	**Estimate**	**SE**	**t**	**p**
Therapist-Supervisor	−0.26	0.06	−4.63	<.0001	0.11	0.06	1.86	0.064
Therapist-TSF Expert	−0.32	0.06	−5.54	<.0001	0.10	0.06	1.61	0.108
Therapist-External	0.12	0.06	2.05	.041	0.78	0.06	12.96	<.0001
Supervisor-TSF Expert	−0.05	0.06	−0.92	0.356	−0.02	0.06	−0.25	0.802
Supervisor-External	0.38	0.06	7.67	<.0001	0.67	0.06	11.07	<.0001
TSF Expert-External	0.4	0.06	7.59	<.0001	0.68	0.06	11.34	<.0001

*Mean Adherence is the mean of 3-9 adherence items, depending on the content of each session. Global Session Performance is the rating of a single item: “Overall, how well did the counselor conduct this specific session?” The outcome measure is the mean of ratings (Mean Adherence, Global Session Performance) for all sessions delivered to a single patient, and the mean of those patient-level ratings aggregated by Rating Group.