HHS Public Access

Author manuscript

Psychother Res. Author manuscript; available in PMC 2015 August 26.

Published in final edited form as:

Psychother Res. 2012 ; 22(6): 720–730. doi:10.1080/10503307.2012.716528.

Do supervisors and independent judges agree on evaluations of therapists' adherence and competence?

Inga Dennhag,
Umeå University

Mary Beth Connolly Gibbons,
University of Pennsylvania

Jacques Barber,
University of Pennsylvania and Philadelphia VA medical center

Robert Gallop, and
West Chester University

Paul Crits-Christoph
University of Pennsylvania

Abstract

The current study examined the agreement between supervisors' and independent judges' evaluations of therapist adherence and competence in three treatments of cocaine dependence: supportive expressive therapy (SE), cognitive therapy CT), and individual drug counseling (IDC). We used data from The National Institute on Drug Abuse Collaborative Cocaine Treatment Study (*N*=295). Trained and experienced supervisors and independent judges rated treatment fidelity. At the therapist level of analysis, the agreement between supervisors' and independent judges' ratings was weak for SE competence, CT adherence, and CT competence. Moderate relations were found for IDC adherence and competence. Supervisors consistently rated adherence and competence more positively than judges in CT and IDC. The potential bias in supervisors' ratings is discussed.

Keywords

adherence; competence; cognitive therapy; supportive expressive therapy; drug counseling

To ensure that manual-based psychotherapies have been implemented adequately in research studies, it is necessary to measure treatment integrity. It is also important in an education context to evaluate if students are competent in a specific psychotherapy. Two aspects of treatment integrity have been investigated in the last three decades. *Adherence* is traditionally defined as the extent to which a therapist delivers the prescribed techniques or interventions, and also involves not delivering components that are proscribed by the therapy (Waltz, Addis, Koerner, & Jacobson, 1993). *Competence* refers to the quality or skill with which interventions are conducted. A skilled intervention is not only a proper and correctly delivered intervention, it is also a matter of the *appropriateness* of the intervention.

To be appropriate, an intervention should include relevant aspects of the context such as the patient's life situation and stages in therapy (Waltz et al., 1993). Therapist adherence and competence is believed to be maximized through training in the use of a treatment manual with regular supervision and assessments of adherence and competence (Barber, Krakauer, Calvo, Badgio, & Faude, 1997).

Four methods are used to measure treatment integrity within psychotherapy outcome research. These methods mirror different perspectives on adherence and competence of the patients, therapists, supervisors, and observers. The patient's perspective (Elliot & Williams, 2003; McCarthy & Barber, 2009) has rarely been used in clinical trials. The therapist's perspective has been implemented through the use of checklists that ask therapists to indicate if specific interventions were implemented (Brosan, Reynolds, & Moore, 2008; Carroll, Nich, & Rounsaville, 1998; Martino, Ball, Nich, Frankforter, & Carroll, 2009).

In clinical and training settings, however, it is not the patient or the therapist who evaluates therapist adherence/competence but most often it is the clinical supervisor who does so. As part of their role, supervisors provide formal or informal feedback on the therapy process with the purpose of improving therapists' skills and patients' outcomes.

Research studies often use independent judges to evaluate therapist adherence/competence from recorded sessions, in part because independent raters are considered more objective than supervisors. Independent judges are unaware of therapist and patient background and the context of the therapy process. However, even if raters are independent, researchers must still be aware of potential biases due to how raters are selected, trained, and how reliability is maintained (Hill & Lambert, 2004). Consequently, training for raters is often rigorous and careful. A number of studies (Barber, Crits-Christoph, & Luborsky, 1996; Shaw et al., 1999; Trepka, Rees, Shapiro, Hardy, & Barkham, 2004) have reported that independent judges' ratings of adherence or competence can be related positively to treatment outcome. However, the results are inconsistent and a recent meta- analysis on the topic concluded that, across studies, there is no correlation between treatment integrity and outcome (Webb, DeRubeis, & Barber, 2010).

It is important to know whether clinical supervisors' ratings of adherence/competence agree with independent judges' ratings. Because it is often recommended that clinical supervisors be actively involved in the process of disseminating evidence-based treatment to community settings by evaluating therapist adherence/competence (Amodeo, Storti, & Larson, 2010), the validity of such ratings needs to be examined. Comparison of supervisors' ratings to more objective independent judges' ratings is therefore an essential step in determining the validity of supervisors' ratings.

A second reason for investigating the agreement between supervisors' and independent judges' ratings is that in training and clinical practice settings supervisors' ratings are usually used, but we do not know if supervisors' ratings are comparable to independent measurement of adherence/competence. Supervisors might rate therapists overly high because of loyalties. Alternatively, supervisors might actually provide more accurate measurement of adherence/competence because of their contextual knowledge compared to

independent judges. Because supervisors are familiar with the full clinical context (i.e., diagnosis, severity of problem, background factors) for each patient, they maybe in a particularly advantageous position to evaluate the appropriateness of therapeutic interventions (Jacobson & Hollon, 1996; Kazantzis, 2003).

Only a few studies have directly compared supervisors' and independent judges' ratings of adherence/competence. Borders and Fong (1992) analyzed data from 43 trainees in a counseling program and found no agreement between supervisors' global ratings of trainees' counseling competence and independent judges' ratings of audiotapes from the same sessions. However, raters were trained differently (27 untrained supervisors vs. 2 trained independent judges) and they used different rating methods (supervisors rating method unknown vs. independent judges used systematic procedure). The variability in the ratings of the therapists' counseling competence was also much larger among the supervisors than for the independent judges. Chevron and Rounsaville (1983) examined nine therapists' skills in Interpersonal Psychotherapy (IPT) for depressed patients, through supervisors' ratings based on therapists' retrospective report of therapy sessions in supervision, and independent evaluators' ratings of videotaped sessions. Results showed low agreement between the rater types. Martino et al., (2009) investigated the agreement of therapist's adherence and competence between supervisors and independent judges in Motivational Enhancement Treatment (MET) for drug abusers. Martino et al. (2009) found that the agreement was poor regarding the extent to which specific interventions occurred in each session regardless the type of intervention. On average, supervisors provided higher ratings of adherence compared to independent judges. Furthermore, there was poor agreement between supervisors and independent judge about therapists' competence level, with the supervisors generally providing higher estimates than the judges for general MET strategies, but few differences for more advanced techniques. Although this small set of studies has shown a consistent finding of relatively low agreement between supervisors and independent judges, this relationship has only been examined in a few different psychotherapies, using a limited number of scales. It is not clear if this lack of agreement generalizes to other therapies and other adherence/competence scales.

Another limitation of existing studies is that the agreement between supervisors' and independent judges' ratings of adherence/competence was assessed only at the level of the individual session or patient. For clinical training purposes, however, the goal is typically to certify a therapist as achieving adequate adherence/competence in the delivery of a specific psychotherapy. In these situations it is therefore relevant to examine the agreement between supervisors and independent judges at the level of the therapist (i.e., averaging across multiple patients for each therapist). No previous study, however, has examined agreement between supervisors and independent judges at the therapist level.

Agreement between supervisors and judges is likely to be increased when data are examined at the therapist level. Previous research has shown that therapist performance can vary as a function of patient difficulty (Foley, O'Malley, Rounsaville, Prusoff & Weissman, 1987), but the impact of outliers would be reduced when data are analyzed at the therapist level with multiple patients per therapist.

The aim of the current study was to examine the agreement between the supervisors' and independent judges' ratings of treatment integrity in three different psychosocial treatments for cocaine dependence: supportive-expressive therapy (SE), cognitive therapy (CT), and individual drug counseling (IDC). A second goal was to go beyond existing studies by using multilevel modeling to examine agreement at the therapist level. A further aim was to compare the supervisors' and judges' mean ratings. Because Martino et al. (2009) found higher estimates for supervisors than for judges in mean adherence and competence, our hypotheses were that our results would show the same discrepancy in all three treatment conditions.

Method

Participants

Data were drawn from the National Institute on Drug Abuse Collaborative Cocaine Treatment Study (NIDA CCTS) (Crits-Christoph et al., 1999), a multicenter randomized clinical trial comparing treatments for cocaine dependence. The study was conducted at five clinical sites in the northeastern United States. The study was reviewed and approved by an Institutional Review Board (for more information about the NIDA CCTS see Crits-Christoph et al., 1999).

Patients—A total of 364 patients were randomly assigned to treatment but only those with data on variables of interest (n = 295) were included in this study. The patients met primary criteria for cocaine dependence diagnosis (100%), but also received additional diagnoses, including alcohol dependence (33%), axis I disorders other than substance dependence (40%), and a anti-social personality disorder diagnosis (44%). The mean age was 34 years (SD = 6.3, range = 22–57) and 79% of the patients were male. Fifty-nine percent of the subjects were Caucasian. Table 1 shows patient characteristics within the three treatment groups.

Therapists—The therapists were recruited through newspaper announcements or from the staff of the study sites. Based on therapists' earlier clinical experiences and therapeutic orientation, the supervisors for each treatment modality decided which therapists were suitable for being included in the study. In total, 39 therapists participated. The mean age was 40 years (SD = 5.9, range = 31–52), and 62% were male. Table 2 provides descriptive information about the therapists in each of the three treatment modalities.

Supervisors—There were three SE supervisors, five CT supervisors, and four IDC supervisors. The supervisors were coauthors of the treatment manuals and had extensive experience training therapists in their respective approach. In addition to direct clinical supervision of individual therapists, the supervisors coordinated and led training workshops together with the heads of the training units. As part of regular clinical supervision, the supervisors rated their supervisee therapists on adherence and competence in their own expert modality. The supervisors planned to evaluate sessions 2 and 4, and after that every fourth session (patients who dropped out early therefore had fewer sessions rated). The collection of the supervisors' ratings and the specification of particular sessions evaluated

was part of the protocol to maintain treatment fidelity in the NIDA CCTS and was unrelated to the current effort to compare supervisors' and independent judges' ratings.

The supervisors rated only their own supervisees. Supportive expressive therapy supervisors evaluated together 436 audiotapes, cognitive therapy supervisors evaluated 518 audiotapes, and individual drug counseling supervisors evaluated 396 audiotapes. The average number of sessions per patient rated by the supervisors were: SE: 4.6 (*SD* = 2.0, Range = 1–10), CT: 5.0 (*SD* = 2.3, Range = 1–13), IDC: 4.1 (*SD* = 2.1, Range = 1–9). The average number of patients per therapist rated by the supervisors were: SE: 7.8 (*SD* = 3.5, Range = 2–14), CT: 6.9 (*SD* = 2.3, Range = 4–13), IDC: 8.2 (*SD* = 3.9, Range = 1–13).

Independent Judges—Clinical experts in the respective treatment approaches were used as independent judges (for more information see Barber et al., 2004). These ratings were collected as part of an effort to independently evaluate treatment fidelity in the NIDA CCTS. Thus, the current study was a secondary analysis of these ratings that were collected for a separate purpose.

There were three SE judges, two CT judges, and two IDC judges. These judges were otherwise not associated with the study. The independent judges were blind to the therapy process and the sequence of the sessions. The judges rated only the tapes for the modality in which they were an expert. Approximately two sessions per patient were rated. One of the two tapes was randomly drawn from session 2 through 11, and the second tape was randomly drawn from session 12 to termination. Less than 2 sessions were obtained for some patients who dropped out of treatment early.

Supportive expressive independent judges evaluated 148 audiotapes, cognitive judges evaluated 192 audiotapes, and individual drug counseling judges evaluated 181 audiotapes. The average number of sessions per patient rated by the independent judge were: SE: *M* = 1.6 sessions, *SD* = .6; CT: *M* = 1.9 sessions, *SD* = .4; IDC: *M* = 1.8 sessions, *SD* = .5.

Because the independent judges' ratings were collected for a separate study (not with the intent to compare to supervisors' ratings), the sessions selected for study were not necessarily the same sessions rated by the supervisors. Within IDC, of the 693 sessions assessed by either the supervisor or judge, there were 68 (9.8%) sessions rated by both supervisor and judge. Similarly for CT, of the 833 sessions assessed by either the supervisor or judge, there were 71 (8.5%) sessions rated by both. For SE, of the 537 sessions assessed by either the supervisor or judge, there were 50 (9.3%) sessions rated by both.

Treatments

The treatments lasted up to six months; patients were seen twice a week during the first three months and once a week during the last three months. All of the treatments have been described more thoroughly previously (Crits-Christoph et al., 1997). The mean (*SD*) number of treatment sessions actually attended by patients in the three treatment conditions were: IDC = 11.9 (10.5), CT = 15.5 (10.6), SE = 15.7 (11.3).

Brief supportive expressive therapy provided in this study followed the general SE psychodynamic therapy treatment manual (Luborsky, 1984) and a more specific variant of it developed for cocaine abusers (Mark & Luborsky, 1992). According to this model, the cocaine problems are viewed in the context of an understanding of the person's intrapsychic and interpersonal functioning. The therapist helps the patients via support and interpretations of main core conflictual relationship themes, to clarify and solve the most interfering problems.

Cognitive therapy as conducted in the present study followed the manual of Beck, Wright and Newman (1993) for substance users. The therapy is based on the assumption that the substance abuse is related to individuals' maladaptive beliefs and related thought processes. The therapist uses Socratic questioning, advantages-disadvantages analysis, monitoring of drug-related beliefs, activity scheduling, behavioral experiments, and role playing to help the patient to change maladaptive emotion-thought-behavior patterns related to drug abuse.

Individual Drug Counseling followed a specific manual (Mercer & Woody, 1992) and is based on the 12-step philosophy. The treatment focuses primarily on helping the patient recover in specific stages. These stages contented how to achieve and maintain abstinence from drugs by encouraging behavioral changes, such as avoiding drug triggers, structuring one's life, and engaging in healthy behaviors. Participation in self-help groups was strongly recommended.

Therapist training

Prior to participating in the main clinical trial, therapists received intensive training in their respective treatment modalities. Therapist training began by reading the treatment manual, and continued with workshops and training cases. Four 2-day workshops were provided in total, and consisted of didactic presentations, role-plays, and discussions of case examples to implement the treatment modality. Four training cases were given to every therapist. The training cases had to last for at least a month to be counted as a training case. All therapy sessions were audiotaped and sent to the supervisor for adherence and competence rating within three days. Those therapists that achieved adequate levels of competence were invited to participate in the main trial for each treatment (for more information see Barber & Crits-Christoph, 1996).

Measures

Two scales were used for the SE condition (one for supervisors and one for independent judges), and one scale in the CT and IDC conditions. The reason for using two scales in the SE condition was that the original supervisors' scale had a limited number of items and therefore the subsequent effort (Barber et al., 1997) to obtain independent ratings of treatment fidelity included the development of a longer scale.

Adherence/Competence Scale for SE Cocaine Dependence—(ACS-SEC, Barber et al., 1997). This scale measures therapist adherence, appropriateness, and the quality of SE interventions. Items are rated separately on adherence, appropriateness, and quality on a scale from 1 (low) to 7 (high), but if adherence is rated as 1 (not at all) the quality rating is

assigned the same number as the appropriateness rating. Because appropriateness and quality scales were found to be highly correlated (r = .94), they were averaged to form an overall competence scale. The scale originally contained three subscales: supportive (13 items), expressive (31 items), and cocaine abuse (11 items), but only the first two scales were used in the current study because the cocaine abuse items were not included on the rating scales used by the supervisors. The supportive items on the scale address how the therapist builds the alliance with the patient. The expressive items address how the therapist uses the techniques in SE such as clarifications and interpretations of core conflicts. Barber et al. (2008) found the interjudge reliability for the ACS-SEC total adherence scale was good (ICC [2,2] = .84), but was weaker for the ACS-SEC total competence score (ICC [2,2] = .68). Barber, Foltz, Crits-Christoph, and Chittams (2004) found that ACS-SEC adherence sale discriminated SE from CT and IDC, but the same was not obtained for the competence scale. In the current study, the internal consistency was excellent for both adherence and competence scales when patient was the unit for the analysis (adherence α = .93, competence α = .98) and when therapist was the unit for the analysis (adherence α = .95, competence α = .99). The scale was rated by independent judges on SE therapy sessions.

Supervisor's adherence/competence scale for TCACS—This scale was developed specifically for use in the NIDA CCTS (Crits-Christoph et al., 1997). The scale measures SE adherence (the frequency of SE interventions) with 6 items, and competence (if the interventions were delivered with quality) with 7 items. The items are rated from 1 (low) to 7 (high) on a Likert scale. The adherence and the competence scale include items related to supportive techniques, the development of relationship episodes, interpretations of Core Conflictual Relationship Themes (CCRT), relating drug use to CCRT, maintaining treatment integrity, and addressing the patient's experience of the therapist or the transference. The competence scale also consisted of one more item; the therapist's general relatedness and supportiveness. In the current study, the internal consistency was fair to good for the adherence and competence scales when the patient was the unit for the analysis (adherence α = .62, competence α = .86) and good when the therapist was the unit for the analysis (adherence α = .84, competence α = .85).

Cognitive Therapy Adherence and Competence Scale—(CTACS, Barber, Liese, & Abrams, 2003). The CTACS scale was also developed specifically for the NIDA CCTS (Crits-Christoph et al., 1997) and derived from the Cognitive Therapy Scale (Young & Beck, 1980). This scale was used by both the supervisors and independent judges for rating CT sessions. The 21-item scale contains five areas of CT therapy integrity: cognitive therapy structure (i.e., agenda, mood check, homework), development of a collaborative relationship (i.e., socialization to CT model, warmness, collaboration), development and application of case conceptualization (i.e., eliciting automatic thoughts, core beliefs and schemas), cognitive and behavioral techniques (i.e., guided discovery, asking for evidence, use of alternative and behavioral techniques), and a final global rating of overall performance as a cognitive therapist. Each item was rated separately on adherence to the manual, appropriateness of the intervention, and the quality of the intervention on a scale ranging from 0 (low) to 6 (high). In general, if an expected intervention did not occur, as for example agenda setting, adherence was rated as zero as were appropriateness and quality.

However, if the rater thought it was appropriate for the therapist to not engage in a certain intervention and the therapist did not engage in that behavior, then the appropriateness rating was scored from 1 to 6. Because of the high correlation (r > .90, Barber, Liese, & Abrams, 2003) between appropriateness and quality scores, an overall competence score was computed as the mean of those scales. Barber et al. (2003) investigated the interrater reliability in the present study's CT sample (n = 92) and found that the ICC total was .67 for adherence, and .73 for competence, indicating acceptable reliability. The scale has also discriminated CT from SE therapy and IDC (Barber et al., 2004). Internal consistency for the current study was good to excellent for the 21-item scale (Patient as the unit of analysis: adherence rated by supervisors α = .88 ; competence rated by supervisors α = .97; adherence rated by judges α = .89; competence rated by judges α = .96. Therapist as the unit of analysis: adherence rated by supervisors α = .89; competence rated by supervisors α = .97; adherence rated by judges α = .96; competence rated by judges α = .98.).

Adherence/Competence Scale for IDC for Cocaine Dependence—(ACS-IDCCD, Barber, Mercer, Krakauer, & Calvo, 1996). The ACS-IDCCD measured adherence (the frequency of IDC interventions) and competence (the quality of those interventions). This scale was used by both supervisors and independent judges for rating IDC sessions. The scale contains 38 items, with 34 of these items covering the main techniques for IDC (Mercer & Woody, 1992). For the current study, the mean of the 34 items was used to create an overall score (see Barber et al., 2006). Each item is rated separately on adherence and competence on Likert scales ranging from 1 (low) to 7 (high). Barber, Mercer et al. (1996) found the interjudge reliability of two judges pooled in the ACS-IDCCD to be acceptable for both adherence and competence, as measured by intraclass correlations. The scale has also discriminated IDC from cognitive and dynamic therapy (Barber et al., 2004; Barber, Mercer et al., 1996). For the current sample the internal consistency was good to excellent for the 34 item scale (Patient as the unit of analysis: adherence rated by supervisors α = .96 ; competence rated by supervisors α = .94; adherence rated by judges α = .84; competence rated by judges α = .92. Therapist as the unit of analysis: adherence rated by supervisors α = .97 ; competence rated by supervisors α = ,98; adherence rated by judges α = .97; competence rated by judges α = .97.).

Statistical procedures

Three analyses were conducted: (1) a preliminary analysis to examine the dependability of patient and therapist level scores for the supervisors' ratings, (2) assessment of the relation between supervisors' and independent judges' ratings, and (3) comparison of the mean levels of supervisors' and independent judges' ratings. All three analyses were conducted using multilevel modeling techniques.

The dependability of supervisors' adherence/competence scores was evaluated using a three level model (sessions within patients within therapists, with all factors specified as random). Generalizability coefficients were calculated from these models to index the dependability of measurement at the patient level (i.e., averaging over sessions) and also at the therapist level (i.e., averaging over patients) using derivations of the formulas given in Webb, Shavelson, and Haertel (2006) as applied to our three level model. This was an important

preliminary analysis because low dependability would attenuate the relationship between supervisors' and independent judges' ratings. Moreover, low dependability at the patient level would call into question comparison of supervisors and independent judges at the patient level (i.e., there is variability over sessions and therefore sessions are not interchangeable). In fact, we (Dennhag, Gibbons, Barber, Gallop & Crits-Christoph, in press) found that the dependability of patient level scores for the independent judges' adherence/competence ratings used in the current report was inadequate (i.e., generalizability coefficients < .70) for all scales. However, at the therapist level, acceptable generalizability coefficients for the judges' ratings were found for CT and IDC (CT adherence = .77, competence = .75; IDC adherence = .71, competence = .77), with lower values for SE therapy (adherence = .58, competence = .65) (Dennhag et al., in press). The independent judges' data used in Dennhag et al. (in press) study were also used in the present study.

For the analyses examining therapist differences and also the relation between supervisors' and independent judges' ratings, a three level model was implemented (sessions nested within patients nested within therapists; all 3 factors random), with the addition of rater type (supervisor vs. judge) as a fixed factor. [1]

The three level equation is:

$$Y_{ijk}=\mu\ldots+\alpha\times Rater_{ijk}+\varepsilon_{j(k)}+\varepsilon_{k}+\varepsilon_{i(j(k))}$$

Where Y_{ijk} is the outcome for the j^{th} session for the j^{th} patient nested within the k^{th} therapist. $\mu\ldots$ is the overall mean for the reference group (supervisor was set as the reference group), α is the estimated on-average difference in the overall mean between the supervisor and judge $\varepsilon_{i(j(k))}$ accounts for the random variance within patients (due to sessions) plus interactions, $\varepsilon_{j(k)}$ accounts for the random patient variance nested within therapists, ε_{k} accounts for the random therapist variance.

From the multi-level model specified above, the Variance-Covariance matrix for the clustered data within the unit of analysis (therapist), is estimated as a function of the random effects above as well as the residual variance. Within the multilevel/mixed model literature, the derivation of the variance-covariance matrix depends on two components: the G matrix and the R matrix (Verbeke and Molenberghs, 2000). The G-matrix consists of the variance estimates per the random effects specified above. The R-matrix models the residual error. The variance-covariance matrix is estimated as a linear combination of the G and R matrices, dependent on the random effects of the model. Correlation estimates (r_T) were derived to quantify the association between judge and supervisor scores of adherence and competence at the therapist level based on the respective variance-covariance term divided by the total variance. In addition, intraclass correlation coefficients based on variance components (ICC_{VarT}) were derived to index the size of therapist differences on adherence / competence scales. The ICC_{VarT} for assessing therapist differences was calculated as

[1]Additional information regarding the equations and models can be obtained directly from the first author.

follows: $ICC_{VaraT} = \frac{\sigma^2_{TOT} - \sigma^2_e - \sigma^2_{P(T)}}{\sigma^2_{TOT}}$, where $\sigma^2_{TOT} = \sigma^2_e + \sigma^2_{P(T)} + \sigma^2_T$ (σ^2_{TOT} refer to the total variance, $\sigma^2_{P(T)}$ is the patient within therapist variance and σ^2_e is the residual variance).

To test for mean differences between supervisors and judges, the α-term in the above three level model equation, corresponding to the cross-classified factor identifying the rating type (Judge versus Supervisor), was examined along with its statistical significance. Effect sizes for the contrast between the mean scores for judges and supervisors were based on the standardized difference for linear mixed models described by Verbeke and Molenberghs (2000). Estimation of multilevel structures described above used SAS 9.2 (Littell, Milliken, Stroup, Wolfinger, & Schabeinberger, 2006) with restricted maximum likelihood estimation (REML).

Results

Descriptive Characteristics of Study Variables

Table 3 presents descriptive characteristics of the study variables, including means and intraclass correlation coefficients reflecting the size of any therapist effect for each of the scales. The ICC_{VarT} indicates that small to moderate size therapist differences existed on all scales, with the exception that large therapist differences were evident on the IDC supervisors' adherence and competence scales.

Dependability of Supervisors' Ratings

At the patient level, generalizability coefficients were marginally acceptable to excellent for the IDC supervisor scores (adherence = .96, competence = .76). The other scales all had inadequate generalizability coefficients at the patient level (SE adherence = .43, SE competence = .51; CT adherence = .46, CT competence = .66). However, therapist level dependability of measurement was acceptable to excellent for CT and IDC condition (CT adherence = .76, CT competence = .79; IDC adherence = .97, IDC competence = .82). For SE therapy, dependability was acceptable for competence (.71), but low for adherence (.16) at the therapist level. Based on the inadequate patient level dependability of the judges' ratings for all 3 treatments (Dennhag et al., in press), and the inadequate patient level dependability of the supervisors' ratings for SE and CT, we focused further analyses on the therapist level only. In addition, the SE adherence scale was dropped from further analyses because of inadequate dependability at the therapist level.

Correlations between Supervisor and Independent Judge Ratings

Correlations at the therapist level computed from the multilevel model showed that agreement between supervisors'and independent judges' ratings of treatment fidelity was weak for SE and CT (r_T = .137 to .257), but moderate between IDC's supervisors and independent judges (r_T = .507 for adherence, r_T = .544 for competence) (Table 4). The weakest correlation was between CT supervisors and independent judge on ratings of competence (r_T = .137). We also conducted these same analyses but restricted the samples of sessions to only those for which both the supervisors and independent judges had rated

the same sessions. Results were very similar, with 4 of the five correlations differing for those in Table 4 by less than .04, and the fifth (SE competence) differing by .07 (i.e., a correlation of .303, compared to .233).

Mean Levels for Supervisors' Ratings Compared to Judges' Ratings

Using multilevel modeling, comparisons at the therapist level on mean levels of adherence for supervisors and judges revealed that, as hypothesized, supervisors rated therapists higher on adherence than did independent judges: CT: t (14) = 11.81, $p < 0.001$, $d = 1.16$, 95% CI = 0.59 – 1.74; IDC: t (11) = 37.61, $p < 0.001$, $d = 3.87$, 95% CI = 3.21 – 4.54. The effect sizes were moderate to very large according to Cohen's (1988) definitions. With regard to competence, supervisors rated therapists more positively than judges in CT (t (14) = 8.15, p <0.001, $d = .80$, 95% CI = 0.23 – 1.38) and in IDC (t (11) = 37.41, $p < 0.001$, $d = 3.83$, 95% CI = 3.17 – 4.49), but not in SE (t (11) = −.14, $p = 0.89$, $d = -0.01$, 95% CI = −0.68 – 0.65). The effect sizes for CT and IDC were large to very-large, respectively.

Discussion

The current study investigated the agreement between supervisors' and independent judges' ratings of treatment integrity. There were two major findings. First, at the therapist level of analysis, the agreement on adherence and competence ratings was statistically significant but relatively low for SE (only adherence) and CT, but moderate for IDC. In cognitive therapy, independent raters and supervisors essentially did not agree at all in regard to competence. Second, supervisors, compared to judges, rated therapists substantially more positively on adherence in CT and IDC conditions. Moreover, CT and IDC supervisors rated therapists higher than judges on competence, which was not found in SE.

Earlier investigations found low agreement between supervisors and independent judges when the patient is the unit of analysis on adherence (Martino et al., 2009) and competence (Borders & Fong, 1992; Chevron & Rounsaville, 1983; Martino et al., 2009). This study is the first to confirm this relative lack of agreement at the therapist level.

There may be several reasons for the low agreement in SE and CT, but moderate agreement in IDC. First, it may be that techniques in IDC are more straightforward to identify in therapy sessions. Items in the IDC adherence/competence scale such as "monitored attendance at 12-step groups" may leave less room for disagreements compared to items in the CT adherence/competence scale such as "elicited core beliefs/schemas; effectively related these to patient's problems." A second reason for the low agreement in regard to CT was that the CTACS competence scale used here, which was derived from the Cognitive Therapy Scale (Young & Beck, 1980), may be problematic for assessing competence. This concern is particularly relevant because the Cognitive Therapy Scale has been widely used to evaluate CT competence in large scale efforts to disseminate CT (e.g., Stirman et al., 2010). Our data are consistent with the results of a previous study (Jacobson et al., 1996 (with more information in Jacobson and Gortner's article (2000)), in which there was no agreement between experts who rated CT sessions with the Cognitive Therapy Scale. The authors concluded that either the Cognitive Therapy Scale was an unreliable instrument or that experts cannot agree on "good" cognitive therapy. However, Barber et al. (2003)

investigated the independent judge interrater reliability of CTACS in the present study's CT sample and found acceptable interjudge reliability for both CT adherence and CT competence. Nevertheless, further scale development with regard to adherence and competence assessment in CT may be indicated to increase the validity of such ratings.

Relatively low agreement between supervisors and judges in rating SE competence may also have been a function of the scales used. In fact, the interjudge reliability of the independent judges' SE competence scale was marginal (.68). Perhaps more importantly, the comparison between SE supervisors' and judges' ratings was utilized with different scales, which was not the case in the CT or IDC condition for which the same measures were compared. It is therefore possible that somewhat different scale content was compared in SE. The use of different scales might also explain why supervisors rated adherence and competence more positively than judges did in all three conditions, except for SE competence. However, the SE judges' scale was largely an expansion of the number of items, not a revision of the constructs to be measured, and therefore the problem may lie with the definitions of the underlying constructs.

The mean differences between supervisors and independent judges found here for CT and SE parallel findings from Martino et al.'s (2009) study that found higher ratings for supervisors than from judges in adherence and in fundamental competence strategies. The present replication of comparisons in two different treatment modalities strengthens the indication that supervisors and judges actually represent different rater perspectives, at least in terms of average levels. Martino et al. (2009) discuss the possibility that their result could be due to training differences of the raters. In the current study, supervisors and judges were all highly trained and the result is less likely to depend on training.

Supervisors are more familiar with the context and may exaggerate their judgment because of loyalty. Another possible explanation for the supervisors' more positive ratings is that their context knowledge gives them an advantage to rate therapists more sensitively and carefully than independent judges (Jacobson & Hollon, 1996; Kazantzis, 2003). Supervisors, for example, might be better than the independent judges at judging when it is appropriate to refrain from an intervention. However, it is also possible that supervisors' ratings were biased upwards based on their relationships with their supervisees. The independent judges were unaware of the patient's condition, the therapist's background, or the therapy process. They had no loyalties to the therapists and were therefore potentially less biased than the supervisors.

Regardless of whether the higher levels of ratings provided by supervisors reflect bias or a more informed position, clinical trainers and researchers should be aware that such ratings may be elevated compared to independent judges' ratings. This is particularly important to consider if the goal in the training or research setting is to certify a therapist as competent based on the therapist achieving a certain absolute score level on a competence scale. Furthermore, if trainees are consistently given feedback by supervisors that their adherence and competence is very positive, the trainees may have little incentive to alter their approach. An interesting approach would be to allow independent judges to provide feedback during therapists' training. That would give another perspective of feedback to

both the therapist and the supervisor that may result in behavior change, without risking a rupture in the alliance between therapist and supervisor.

Limitations of this study included the facts that supervisors' and judges' ratings were often obtained from different sessions for each patient and that judges' ratings were obtained from a small sample of sessions. Another shortcoming of the study was that the supervisors and the independent judges were not randomized to their different conditions, and therefore not free from selection bias. They were selected because of their suitability and competence in each respective area. Raters' characteristics, perceptions, and attitudes might interfere with their ability to be objective (Fiske, 1977). The current study limited this impact through careful training of the raters, the use of protocol for rater procedures, the use of treatment manuals, and ongoing meetings to calibrate the raters. Another limitation is that the generalizability of the current findings to other types of treatments, and other patient populations, is not known. It may be that adherence/competence is particularly difficult to evaluate in a substance dependent population because of a high frequency of disruptive life events that interfere with implementation of standard therapy techniques. A final limitation is that certain other variables not measured here (e.g., patient difficulty, quality of the supervisory relationship) might moderate the relationship between supervisor and independent judge ratings. Further research is needed to investigate the role of such potential moderator variables.

In summary, this study supports the view that supervisors and independent judges have somewhat different perspectives when they rate treatment integrity. The current study findings also add substantially to our understanding of those different perspectives by investigating data at the therapist level. Ratings from supervisors and independent judges have low to moderate agreement, and that should be taken into consideration when evaluating therapists in a training context. Consistent with prior research, we also found that supervisors had higher mean levels than independent judges on adherence and competence ratings. This finding was evident in three different treatment modalities, suggesting that a systematic differences in supervisors' versus independent judges' perspectives on treatment fidelity.

References

Amodeo M, Storti SA, Larson MJ. Moving empirically supported practices to addiction treatment programs: recruiting supervisors to help in technology transfer. Substance Use and Misuse. 2010; 45:968–982. [PubMed: 20397880]

Barber JP, Crits-Christoph P. Development of a therapist adherence / competence rating scale for supportive-expressive dynamic psychotherapy: A preliminary report. Psychotherapy Research. 1996; 6:81–94. [PubMed: 22242608]

Barber JP, Foltz C, Crits-Christoph P, Chittams J. Therapist's adherence and competence and treatment discrimination in the NIDA Collaborative Cocaine Treatment Study. Journal of Clinical Psychology. 2004; 60:29–41. [PubMed: 14692007]

Barber JP, Crits-Christoph P, Luborsky L. Effects of therapist adherence and competence on patient outcome in brief dynamic therapy. Journal of Consulting and Clinical Psychology. 1996; 64:619–622. [PubMed: 8698958]

Barber JP, Gallop R, Crits-Christoph P, Barrett MS, Klostermann S, McCarthy KS, Sharpless BA. The role of the alliance and techniques in predicting outcome of suportive-expressive dynamic therapy for cocaine dependence. Psychoanalytic Psychology. 2008; 25:461–482.

Barber JP, Gallop R, Crits-Christoph P, Frank A, Thase ME, Weiss RD, Gibbons C. The role of therapist adherence, therapist competence, and alliance in predicting outcome of individual drug counseling: Results from the National Institute Drug Abuse Collaborative Cocaine Treatment Study. Psychotherapy Research. 2006; 16:229–240.

Barber JP, Krakauer I, Calvo N, Badgio PC, Faude J. Measuring adherence and competence of dynamic therapists in the treatment of cocaine dependence. Journal of Psychotherapy, Practice and Research. 1997; 6:12–14. Retrieved from: http://jppr.psychiatryonline.org/cgi/reprint/6/1/12. [PubMed: 9058557]

Barber JP, Liese B, Abrams M. Development of the cognitive therapy adherence and competence Scale. Psychotherapy Research. 2003; 13:205–221.

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]

Beck, AT.; Wright, FD.; Newman, CF.; Liese, BS. Cognitive therapy of substance abuse. New York: Guilford press; 1993.

Borders LD, Fong ML. Evaluations of supervisees: brief commentary and research report. The Clinical supervisor. 1992; 9:43–51.

Brosan L, Reynolds S, Moore RG. Self-evaluation of cognitive performance: do therapists know how competent they are? Behavioural and Cognitive Psychotherapy. 2008; 36:581–587.

Bryk, A.; Raudenbush, S. Hierarchical Linear Modeling: applications and data analysis methods. Newbury Park, CA: Sage Publishing; 1996.

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

Chevron ES, Rounsaville BJ. Evaluating the clinical skills of psychotherapists. A comparison of techniques. Archives of general psychiatry. 1983; 40:1129–1132. Retrieved from http://archpsyc.ama-assn.org/cgi/reprint/40/10/1129. [PubMed: 6625860]

Cohen, J. Statistical power analysis for the behavioral sciences. 2nd ed.. Hillsdale: NJ: Erlbaum; 1988.

Crits-Christoph P, Mintz J. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. Journal of Consulting and Clinical Psychology. 1991; 59:20–26. [PubMed: 2002139]

Crits-Christoph P, Siqueland L, Blaine J, Frank A, Luborsky L, Onken LS, Moras K. The NIDA Collaborative Cocaine Treatment Study: rationale and methods. Archives of General Psychiatry. 1997; 54:721–726. Retrieved from http://archpsyc.ama-assn.org/content/vol54/issue8/index.dtl. [PubMed: 9283507]

Crits-Christoph P, Siqueland L, Blaine J, Frank A, Luborsky L, Onken LS, Beck AT. Psychosocial treatments for cocaine dependence. Archives of General Psychiatry. 1999; 56:493–502. Retrieved from http://archpsyc.ama-assn.org/cgi/content/full/56/6/493. [PubMed: 10359461]

Crits-Christoph P, Siqueland L, Chittams J, Barber JP, Beck AT, Frank A, Woody G. Training in cognitive, supportive-expressive, and drug counseling therapies for cocaine dependence. Journal of consulting and clinical psychology. 1998; 66:484–492. [PubMed: 9642886]

Dennhag ID, Gibbons M-B, Barber JP, Gallop R, Crits-Christoph P. How many treatment sessions and patients are needed to create a stable score of adherence and competence in the treatment of cocaine dependence? Psychotherapy Research. (in press).

Elliot M, Williams D. The client experience of counseling and psychotherapy. Counselling Psychology Review. 2003; 18:34–38.

Everitt BS. The Analysis of Repeated Measures: A Practical Review with Examples. The Statistician. 1995; 44:113–135. Retrieved from http://www.jstor.org/pss/2348622.

Fiske, DW. Methodological issues in research on the psychotherapist. In: Gurman, AS.; Razin, AM., editors. Effective psychotherapy. New York: Pergamon Press; 1977. p. 23-43.

Foley SH, O'Malley S, Rounsaville BJ, Prusoff BA, Weissman MM. The relationship of patient difficulty to therapist performance in interpersonal psychotherapy of depression. Journal of Affective Disorders. 1987; 12:207–217. [PubMed: 2956305]

Goldstein, H. Models in Educational and Social Research. New York, NY: Oxford University Press; 1987.

Hill, CE.; Lambert, MJ. Methodological issues in studying psychotherapy processes and outcomes. In: Lambert Michael, J., editor. Bergin and Garfield's Handbook of Psychotherapy and Behavior change. 5th ed. New York: Wiley; 2004. p. 84-135.

Jacobson NS, Dobson KS, Truax PA, Addis ME, Koerner K, Gollan JK, Prince SE. A component analysis of cognitive-behavioral treatment for depression. Psychology. 1996; 64:295–304.

Jacobson NS, Gortner ET. Can depression be de-medicalized in the 21st century: scientific revolutions, counter-revolutions and the magnetic field of normal science. Behaviour research and therapy. 2000; 38:103–117. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10660997. [PubMed: 10660997]

Jacobson NS, Hollon SD. Prospectus for future comparisons between drugs and psychotherapy: lessons from the CBT-versus-pharmacotherapy exchange. Journal of Consulting and Clinical Psychology. 1996; 64:104–108. [PubMed: 8907089]

Kazantzis N. Therapist competence in review of the contemporary empirical evidence. Behaviour Change. 2003; 20:1–12.

Littell, RC.; Milliken, GA.; Stroup, WW.; Wolfinger, RD.; Schabenberger, O. SAS System for Mixed Models. 2nd ed. Cary NC: SAS Institute Inc; 2006.

Luborsky, L. Principles of psychoanalytic psychotherapy: A manual for Supportive-Expressive treatment. New York: Basic Books; 1984.

Mark, D.; Luborsky, L. A manual for the use of supportive-expressive psychotherapy in the treatment of cocaine abuse. Philadelphia: Department of Psychiatry, University of Pennsylvania; 1992.

Martino S, Ball S, Nich C, Frankforter TL, Carroll KM. Correspondence of motivational enhancement treatment integrity ratings among therapists, supervisors, and observers. Psychotherapy research. 2009; 19:181–93. [PubMed: 19396649]

McCarthy KS, Barber JP. The Multitheoretical List of Therapeutic Interventions (MULTI): initial report. Psychotherapy research. 2009; 19:96–113. [PubMed: 19065285]

Mercer, D.; Woody, G. Addiction counseling. Philadelphia, USA: Unpublished manuscript, Veterans Affairs Medical Center, University of Pennsylvania; 1992.

Shaw BF, Elkin I, Yamaguchi J, Olmsted M, Vallis TM, Lowery A, Imber SD. Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression. Journal of Consulting and Clinical Psychology. 1999; 67(6):837–846. Retrieved from http://spider.apa.org/ftdocs/ccp/1999/december/ccp676837.html. [PubMed: 10596506]

Stirman SW, Bhar SS, Spokas M, Brown GK, Creed TA, Perivoliotis D, Beck AT. Training and consultation in evidence-based psychosocial treatments in public mental health settings: The access model. Professional Psychology: Research and Practice. 2010; 41:48–56.

Trepka C, Rees A, Shapiro DA, Hardy GE, Barkham M. Therapist competence and outcome of cognitive therapy for depression. Cognitive Therapy and Research. 2004; 28:143–157.

Verbeke, G.; Molenberghs, G. Linear Mixed models for Longitudinal Data. New York: Spinger-Verlag; 2000.

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 JP. Therapist adherence/competence and treatment outcome: A meta-analytic review. Journal of consulting and clinical psychology. 2010; 78:200–11. [PubMed: 20350031]

Webb, NM.; Shavelson, RJ.; Haertel, EH. Reliability coefficients and generalizability theory. In: Rao, CR.; Sinharay, S., editors. Handbook of Statistics. Vol. 26. Amsterdam, Netherlands: Elsevier; 2006. p. 1-124.

Young, J.; Beck, AT. Cognitive therapy scale rating manual. Philadelphia, PA: Unpublished manuscript, University of Pennsylvania; 1980.

Table 1

Demographic and Clinical Characteristics of Patients

Patient	**SE (*n*=94)**	**CT (*n*=103)**	**IDC (*n*=98)**
Age - Mean (Range), years	33.7 (22–48)	35.0 (22–57)	33.3 (22–55)
Men (%)	75 (80)	86 (83)	71 (72)
Race/Ethnicity			
White (%)	56 (60)	59 (57)	58 (59)
African-American (%)	36 (38)	41 (40)	36 (37)
Asian (%)	2 (2)	0 (0)	1 (1)
Hispanic (%)	0 (0)	1 (1)	3 (3)
Other Race (%)	0 (0)	2 (2)	0 (0)
Employed (%)	59 (63)	63 (61)	61 (62)
Living alone (%)	33 (35)	44 (43)	43 (44)
Cocaine use past 30 days, Mean (*SD*)	9.7 (6.9)	9.8 (8.2)	10.4 (7.6)

Note. SE = Supportive Expressive Therapy, CT = Cognitive Therapy and IDC = Individual Drug Counseling.

Table 2

Characteristics of Therapists

Therapist	**SE (*n*=12)**	**CT (*n*=15)**	**IDC (*n*=12)**
Age - Mean (Range), years	38.9 (31–48)	40.0 (31–52)	40.1 (31–49)
Men (%)	8 (67)	12 (80)	4 (33)
Etnicity			
White (%)	11(92)	14 (93)	9 (75)
African-American (%)	0	1 (7)	3 (25)
Other (%)	1 (8)	0	0
Degree			
MD (%)	1 (8)	0	0
PhD/PsyD/EdD (%)	8 (67)	12 (80)	0
MSW (%)	0	3 (20)	0
MA (%)	3 (25)	0	5 (42)
BA, AA, RN (%)	0	0	7 (58)
Years of Clinical Experience, Mean (*SD*)	11.7 (6.1)	13.4 (8.4)	13.8 (8.5)
Years of Substance Use	8.7 (7.4)	6.6 (6.2)	10.4 (6.5)
Clinical Experience Mean (*SD*)			

Note. SE = Supportive Expressive Therapy, CT = Cognitive Therapy and IDC = Individual Drug Counseling.

Table 3

Descriptive Characteristics of Study Variables

Variable	*M (SD)*	ICC_{VarT}
SE Judge Scale		
Adherence	2.5 (.3)	.09
Competence	3.8 (.3)	.20
SE Supervisor Scale		
Adherence	3.1 (.5)	.23
Competence	3.8 (.4)	.25
CT Judge Scale		
Adherence	3.5 (.6)	.42
Competence	4.0 (.5)	.41
CT Supervisor Scale		
Adherence	4.1 (.4)	.21
Competence	4.4 (.4)	.24
IDC Judge Scale		
Adherence	2.3 (.3)	.19
Competence	4.0 (.4)	.17
IDC Supervisor Scale		
Adherence	3.6 (.8)	.74
Competence	5.2 (1.1)	.74

Note. Means calculated using therapist as unit of analysis. Mean of scale items is shown. SE and IDC items rated on 1 to 7 scale; CT items rated on 0 to 6 scale. ICC_{VarT} = intraclass correlation coefficient indicating the size of the effect for differences between therapists for each scale.

Table 4

Supervisors' and Independent Judges' Agreement on Treatment Fidelity at the Therapist Level

Judge	Supervisor	
	Supportive Expressive Therapy	
		Competence
Competence		.233*
	Cognitive Therapy	
	Adherence	Competence
Adherence	.257**	
Competence		.137*
	Individual Drug Counseling	
	Adherence	Competence
Adherence	.507*	
Competence		.544*

Note. The estimated correlation estimates (r_T) between the scales at the therapist level are shown.

* $p < .05$,

** $p < .01$