Published in final edited form as:

Addict Disord Their Treat. 2016 June ; 15(2): 74–84. doi:10.1097/ADT.0000000000000083.

Assessing Treatment Integrity in Alcohol Behavioral Couple Therapy

Kevin A. Hallgren, Ph.D.,

Department of Psychiatry and Behavioral Sciences, University of Washington

Shirley M. Crotwell, M.S.,

Center on Alcoholism, Substance Abuse, and Addictions, University of New Mexico

Rosa E. Muñoz, M.S.,

Center on Alcoholism, Substance Abuse, and Addictions, University of New Mexico

Becky K. Gius, B.A.,

Center on Alcoholism, Substance Abuse, and Addictions, University of New Mexico

Barbara S. McCrady, Ph.D.,

Center on Alcoholism, Substance Abuse, and Addictions, University of New Mexico

Benjamin O. Ladd, Ph.D., and

Department of Psychology, Washington State University Vancouver

Elizabeth E. Epstein, Ph.D.

Department of Psychiatry, University of Massachusetts Medical School

Abstract

Objectives—Alcohol Behavioral Couple Therapy (ABCT) is an efficacious treatment for alcohol use disorders. Coding treatment integrity can shed light on the active ingredients of ABCT, but there are no published studies of treatment integrity instruments for ABCT. The present study describes the development and initial reliability of the Treatment Integrity Rating System – Couples Version (C-TIRS) for ABCT.

Methods—The C-TIRS was used to rate 284 first- and mid-treatment ABCT sessions of 188 couples in four randomized clinical trials.

Results—Average inter-rater reliability for distinguishing ratings between C-TIRS items was fair-to-good for quantity items (intraclass correlation [ICC] = 0.64) and poor-to-fair for quality items (ICC = 0.41). Five C-TIRS subscales were defined *a priori* to measure treatment components involving cognitive-behavioral therapy, spouse involvement, couple therapy, common therapeutic factors, and overall adherence to the treatment protocol and had adequate internal reliability ($\alpha =$

Corresponding author: Kevin A. Hallgren; Center for the Study of Health and Risk Behaviors; University of Washington; 1100 NE 45th St.; Seattle, WA 98105; Phone: 206-221-2380; Fax: 206-616-1705; khallgre@u.washington.edu.

All authors declare no conflicts of interest.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Alcohol Abuse and Alcoholism or the National Institutes of Health.

0.74–0.89). Inter-rater reliability was fair to good on seven of ten scales but poor on three scales (ICC range = 0.17–0.72).

Conclusions—The C-TIRS was designed to provide information about quantity and quality of the delivery of ABCT components; however, further refinement of the C-TIRS is warranted before it should be used in frontline practice. Clinical implications and recommendations for future research are discussed.

Keywords

alcohol behavioral couple therapy; behavioral coding; treatment fidelity; treatment integrity

Introduction

Background

Alcohol use disorders create many negative consequences to physical health and interpersonal relationships (1–2). Marriages where one partner has an alcohol use disorder are four times more likely to dissolve than marriages where neither partner has an alcohol use disorder (3). Alcohol use disorders are associated with poorer relationship satisfaction (4), greater marital conflict (5–6), and greater rates of intimate partner violence (7). These negative consequences, in turn, are associated with alcohol relapse (8–9). Spouses of individuals with alcohol use disorders also experience negative repercussions, including greater rates of mood and anxiety disorders and poorer physical health (3). Given the significant impact of alcohol use disorders on couples' relationships, individuals with alcohol use disorders and their spouses may benefit from treatments that simultaneously address both drinking and the intimate relationship.

Alcohol Behavioral Couple Therapy

Research has shown Alcohol Behavioral Couple Therapy (ABCT), a cognitive-behavioral treatment (CBT) that engages identified patients (IPs) with alcohol use disorders and their significant others (SOs), leads to more days of abstinence and higher relationship satisfaction than individual-based treatment (10–12). ABCT is manual-guided and is hypothesized to facilitate better treatment outcomes through enhanced relationship satisfaction and improved IP and SO skills for maintaining abstinence (13); however, the specific therapist behaviors that are most helpful in producing behavioral changes (i.e., active ingredients) are not well understood.

Using treatment integrity measures to better understand active ingredients could facilitate improved treatment delivery, for example, by focusing more squarely on active ingredients (increasing efficacy) and achieving improvement over fewer sessions (increasing efficiency). Treatment integrity measures also could help monitor the success of training efforts with clinicians, who often do not treat alcohol use disorders within a behavioral couple therapy framework despite the potential for increased client benefit (14). To this day, clinical trials continue to test the efficacy of ABCT (e.g., 15–17), and there is a need for adequate measures of treatment integrity to assess whether ABCT was delivered as intended (i.e., to assess internal validity).

Integrity in Alcohol Treatments

Treatment integrity is conceptualized as the degree to which a treatment is implemented adequately and as intended (18). Thus, measuring treatment integrity requires an assessment of both the degree to which specific treatment components were present (i.e., *quantity* of treatment) and the appropriateness or skillfulness of the delivery (i.e., *quality* of treatment). Although there is a lack of treatment integrity measures developed specifically for ABCT, multiple treatment integrity measures have been developed for individual alcohol treatment (e.g., 19–22). Because ABCT is largely based on a CBT model of alcohol treatment, the treatment integrity measure described in the present study was conceptualized as an extension of existing treatment integrity measures for individual-based CBT.

Treatment Integrity in CBT for Alcohol Use Disorders

Carroll and colleagues (20) developed the MATCH Tape Rating Scale to assess treatment integrity and discriminability for CBT compared to Motivational Enhancement Therapy and Twelve-Step Facilitation in a multisite randomized clinical trial of individual-based alcohol treatments (23). Treatment manuals were used to guide the development of Likert-type items to assess the quantity (but not quality) of the delivery of hypothesized active ingredients for each treatment. Coding manuals were developed through an iterative process, and confirmatory factor analysis indicated good fit for a three-factor model that corresponded to the three treatments. The CBT factor, which formed the basis for many items in the measure used in the present study, included items assessing therapists' review of homework, skill training (e.g., coping with cravings, addressing interpersonal difficulties), reviewing high-risk situations, and relapse prevention. Inter-rater reliability for the CBT factor was modest (intraclass correlation [ICC] = .46) and internal reliability was adequate (Cronbach's alpha = .70). Carroll and colleagues (24) modified the instrument into the Yale Adherence and Competence Scale, which yielded separate quantity and quality ratings of therapist behaviors thought to be common across addiction treatment modalities. The modified measure demonstrated excellent inter-rater reliability for quantity and quality CBT-component scales (ICCs = 0.88) in an alcohol- and cocaine-use disorder treatment trial consisting of CBT, twelve-step facilitation, and medication management.

Morgenstern and colleagues (21) later used the original MATCH Tape Rating Scale and a modified version of the instrument, the Improving Addiction Counseling through Technology Transfer Tape Rating Scale, to evaluate CBT adherence among community-based substance use treatment providers who received one week of CBT training. Each instrument was modified to assess both the quantity and quality of treatment components. Inter-rater reliability was high for quantity and quality items, with median ICCs around 0.85. In summary, a small number of instruments have assessed individual-based CBT treatment integrity for almost two decades with "fair" to "excellent" inter-rater reliability (25).

We are only aware of one treatment integrity rating system for general (non-alcohol focused) behavioral couple therapy, which was used to distinguish traditional behavioral couple therapy from integrative behavioral couple therapy in two clinical trials (26–27). The measure consisted of 17 Likert-type scale items assessing the use of change-focused interventions (characterizing traditional behavioral couple therapy) or acceptance-focused

interventions (characterizing integrative behavioral couple therapy). The items had high inter-rater reliability (ICCs = 0.74 to 0.95) and were able to distinguish the two treatments; however, detailed psychometric analyses of the instrument have not been reported. To our knowledge, there has been little or no additional work published on behavioral couple therapy treatment integrity rating systems more generally. Further, there are no published coding systems that assess treatment integrity for ABCT, which includes SO-specific and couple-specific treatment components not assessed in existing measures. This gap may be due to the complexity of couples' treatment, the relative recency of research on active ingredients, and the time- and cost-intensive nature of developing and testing treatment integrity measures. It is for precisely these reasons that the development and testing of treatment integrity measures is warranted.

The Present Study

The aim of the current study was to develop and test a treatment integrity rating system for use with ABCT. It was hoped that developing such a rating system could inform subsequent research, including research on active ingredients of ABCT, and could provide a model for developing rating systems for non-alcohol focused behavioral couple therapies. The present report describes the development of the Treatment Integrity Rating System – Couples Version (C-TIRS), provides an overview of the scale content, reports descriptive statistics and psychometric properties, and discusses recommendations for further improving treatment integrities in couple therapies.

Materials and Methods

Participants

IP and SO participants were enrolled in one of four randomized clinical trials of ABCT, referred to here as the Program for Alcoholic Couples Treatment (28), Men's (29), Women's I (11), and Women's II (30) studies. In total, 218 heterosexual dyads were included in the original studies, and 188 had at least one first-treatment session ($n = 169$) or mid-treatment session ($n = 115$; session 9 for Women's II, session 8 for all other studies) used in the present analysis. Reasons for missing recordings included dropping out of treatment (65 sessions), sessions not being recorded or missing from the archive (60 sessions), inaudible or very poor audio quality (22 sessions), and other reasons (5 sessions). Sessions were more commonly unavailable for the Men's study compared to others, $\chi^2(1) = 5.96$, $p = .01$, largely due to a greater rate of missing and inaudible tape recordings.

The Women's I and Women's II studies recruited female IPs, the Men's study recruited male IPs, and the Program for Alcoholic Couples Treatment study recruited both male and female IPs. Across all studies, IPs met criteria for DSM-III or DSM-IV (31–32) alcohol abuse or alcohol dependence and had consumed alcohol within 30–60 days before the initial screening. Exclusionary criteria included current drug dependence with physiological dependence, current psychotic disorder, or significant cognitive impairment. Couples in which the SO had an alcohol use disorder were excluded in the Program for Alcoholic Couples Treatment and Men's studies. All of the couples were in committed heterosexual relationships and both partners were willing to come to treatment. Couples were recruited

from the community primarily through newspaper advertisements and other treatment agencies. All participants could receive up to 12–20 ABCT sessions (depending on the study) that were 90 minutes long. Inclusion criteria, recruitment procedures, structure of treatment, and therapist training are further detailed in Supplemental Digital Content (Table S.1).

Of the 188 dyads, 37.8% came from Men's ($n = 71$), 27.1% from Women's II ($n = 51$), 25% from Women's I ($n = 47$), and 10.1% from the Program for Alcoholic Couples Treatment ($n = 19$). Mean IP age was 43.5 ($SD = 10.3$) years and 54.3% of IPs were female. Mean SO age was 44.1 ($SD = 11.4$) years. Most dyads were married (85.1%); the remainder were cohabitating but unmarried (7.4%), in a committed relationship but not cohabitating (3.7%), separated (2.7%), or unknown (1.1%). Most IPs and SOs were White (91.5% and 79.8% respectively), and the second most common race was Black/African American (4.3% and 3.7% for IPs and SOs, respectively).

Treatments

The four original clinical trials included the same overarching ABCT components: (a) individual CBT to improve IP coping skills and abstinence self-efficacy, (b) unilateral family therapy to help SOs implement behavioral strategies to reduce IP drinking and appropriately respond to IP alcohol use if it occurred, and (c) behavioral couple therapy to attenuate relationship distress and improve relationship satisfaction and communication (13). The treatment protocols in the present study also emphasized the importance of common factors such as building rapport, continuity between sessions, empathy, and support for client efforts. Additional details about the treatment protocols are available in Supplemental Digital Content (Tables S.1 and S.2; also see 33).

The C-TIRS

The C-TIRS used in the current study evolved from the rating system developed for individual-based CBT by Carroll et al. (20, 24) and refined by Morgenstern et al. (21). Like previous versions of the measure, the C-TIRS assesses CBT treatment elements; unlike other measures, it also includes items that assess treatment integrity for SO- and couple-specific treatment components, common therapeutic factors, and overall adherence to the ABCT protocol. The C-TIRS was developed initially for Women's I and preliminary data were presented by McCrady et al. (11), then modified for Women's II, and further refined for use in the current study.

The C-TIRS has 37 items with two ratings per item to reflect the quantity and quality of the delivery of each treatment element. Although "quantity" and "adherence" are often used interchangeably, for the present study, "quantity" was defined as the degree to which each treatment element was delivered, which is conceptually different from "adherence" – for example, a therapist might spend considerable time delivering a treatment component (high quantity), but this might be contrary to what is prescribed by the treatment manual for that session (low adherence). Thus, low quantity ratings did not necessarily reflect poor adherence; however, non-adherence was incorporated into quality ratings in conjunction with the overall skillfulness (i.e., competence) of the delivery. Ratings were made on five-

point Likert-type scales (1 = “not at all”/“poor” quantity/quality, 5 = “extensively”/“excellent” quantity/quality).

C-TIRS items were divided into five *a priori*-defined therapeutic domains, including CBT components (17 items), SO components (2 items), couple-specific components (7 items), common therapeutic factors (10 items), and general adherence/skillfulness (1 item; see examples in Table 1, see Supplemental Digital Content Table S.2 for full list). The general adherence item assessed the overall coverage of prescribed treatment components and overall skillfulness of delivering the session. The C-TIRS is freely available for download from <http://casaa.unm.edu>.

The present study rated first- and mid-treatment sessions, which were selected due to the similarity in the content covered in these sessions across the four parent studies and because recordings of these sessions were consistently available across studies. In all studies, first-session material included rapport building, presentation of rationale, skill training, assigning homework, rationale for partner involvement in treatment, and partner skill training (e.g., partner self-recording); some studies also included agenda setting, motivational enhancement, and anticipating high-risk situations. In all studies, mid-treatment session material included assessment of drinking and relationship satisfaction, review of homework, presentation of rationale, skill training, assigning homework, partner skill training, reciprocity enhancement, and communication enhancement; some studies also included agenda setting, functional analysis, developing alternative behaviors to drinking, relapse prevention, and anticipating high-risk situations. More detailed descriptions of session content by study are provided in the Supplemental Digital Content (Table S.2). All studies specified that therapists should address the couple’s relationship, involve both partners, and adhere to common factors and to the general structure of the treatment manual in every session.

Coders

Seven coders were chosen from a psychology doctoral program based on their interest in and familiarity with alcohol use disorder treatment and couples therapy. Six coders were in a clinical psychology program and one was in an evolutionary psychology program. Five coders were female and four had master’s degrees at the time of coding.

Training included intensive one-week training, a 1.5-day workshop with an expert coder (last author), review of the C-TIRS and parent study manuals, and group and individual practice with discussion. After five months of training, coders were determined to have achieved an adequate level of proficiency based on the requirement that inter-rater reliability ICCs for the quantity items of the full C-TIRS exceed 0.60. During the training period, it became apparent that it was difficult to obtain adequate inter-rater reliability for quality ratings. We worked to improve the precision of the quality rating definitions, but ICCs for quality items remained low and therefore were not considered when certifying coder proficiency. Reliabilities for subscales were not used for determining coder proficiency. Coders continued to attend weekly meetings throughout the study to reduce coder drift.

Coding Procedures

First and mid-treatment session audiotapes were digitized, transcribed, and randomly allocated to single coders. A random subset of 33 sessions (11.6% of total) was blindly assigned to be rated by all coders to assess inter-rater reliability. This subset size was similar to those used in similar work (e.g., 19–20). Relative to the full sample, this subsample had a proportionate number of first-session and mid-treatment sessions and a proportionate number of sessions from each of the four parent studies.

Coders reviewed relevant sections of treatment manuals before and during the rating process to be reminded of the manual-specified treatment components, which in turn informed item ratings. Coders listened to sessions in their entirety while taking notes; final ratings were made only after listening to the entire session.

Analytic Plan

The analyses used here were similar to those of previous treatment integrity studies (e.g., 19–21). Inter-rater reliability was computed using two-way, single-measures, consistency ICCs, which estimate the consistency in coder ratings and provide conservative estimates of agreement (e.g., correcting for chance agreement; 34). ICCs were interpreted using the rules of thumb put forward by Cicchetti (25), with ICCs indicating poor (<0.40), fair (0.41–0.60), good (0.61–0.75), or excellent (>0.75) inter-rater reliability. Internal consistency was examined using Cronbach's alpha for the four scales with multiple items. Pearson correlations among the C-TIRS scales were used to assess convergent and discriminant validity.

Results

Reliability analyses focused on coders' ability to distinctively rate different C-TIRS items within each session produced mean ICC values of 0.64 (range across sessions = 0.44 to 0.78) for quantity items and 0.41 (range across sessions = 0.06 to 0.64) for quality items. This indicates that, on average, coders had fair-to-good agreement in the patterns of quantity ratings and poor-to-fair agreement for quality items.

Descriptive statistics for C-TIRS scales are provided in Table 2; inter-rater reliability estimates are presented in Table 3. Inter-rater reliabilities for scale ratings were variable (Table 3), with ICCs for quantity scales ranging from 0.28 to 0.72, indicating a mix of poor (common factors), fair (SO components, couple components, overall adherence), and good reliability (CBT components). ICCs for the quality scales were lower, ranging from 0.17 to 0.46, indicating poor (SO components, couple components) and fair reliability (CBT components, common factors, overall adherence); however, many reliability estimates were near the upper or lower bounds of the interpretive categories offered by Cicchetti (25), and these interpretations should be interpreted as rules of thumb. Overall, these results suggest a variable degree of measurement error in coder ratings.

Cronbach's alphas indicated that internal reliability was acceptable for quantity scales: 0.74 (CBT components), 0.74 (SO components), 0.74 (couple components), and 0.81 (common

factors). Similarly, Cronbach's alphas were acceptable for quality scales: 0.85 (CBT components), 0.80 (SO components), 0.82 (couple components), and 0.89 (common factors).

Quality and quantity ratings within similar ABCT domains were highly correlated with each other: $r_s = 0.81$ (CBT components), 0.80 (SO components), 0.54 (couple components), 0.86 (common factors), and 0.75 (overall adherence), all $p_s < .001$, indicating that high quantity ratings usually co-occurred with high quality ratings in similar domains. Correlations between different scales are presented in Table 4. All correlations in Table 4 were positive and significant ($p < .001$) except the correlation between couple component quantity and overall adherence quantity ($p = 0.93$).

Discussion

The present study developed and tested the C-TIRS, a behavioral coding measure of treatment integrity for ABCT. The results indicated several areas of strength in the C-TIRS as well as several weaknesses of the instrument.

Inter-rater reliability focused on distinguishability between items was fair-to-good for quantity ratings and poor-to-fair for quality ratings. These findings suggest that C-TIRS items likely were distinct enough to reliably capture some of the differences in the degree to which therapists delivered the elements of ABCT but not the quality with which they provided elements of treatment. At the scale level, measurement error was highly variable and often quite prominent. Using Cicchetti's classification system for ICCs, reliability was good for the CBT-components quantity scale, fair for the SO-components, couple-components, and overall adherence quantity scales, and poor for the common factors quantity scale. Reliability was fair on three of the quality subscales (CBT components, common factors, overall adherence) but poor on two others (SO components and couple components), despite internal consistency being adequate for all quality and quantity scales.

Correlations among the C-TIRS scales provide some evidence of convergent and discriminant validity of the C-TIRS; however, these should be interpreted with caution for scales with poor reliability. Quantity scales were highly correlated with quality scales that measured the same construct, and likewise, correlations between different scales were typically positive and medium or large (e.g., up to $r = 0.75$). Although these correlations were often high, they are consistent with other treatment integrity studies (e.g., 19). Several interpretations are possible based on these high correlations. For example, it is possible that items may have measured similar content, that coders had difficulty distinguishing quantity from quality or had difficulty distinguishing between different scales, or that the degree of quality and quantity of delivery of ABCT components tends to be related. The non-significant correlation between quantity of couple-specific treatment components and quantity of overall adherence was unexpected given that coders were instructed to include couple therapy elements in their ratings of overall adherence. It is possible that coders instead gave greater weight to therapists' delivery of other components and had difficulty incorporating couple therapy elements into their ratings of overall adherence.

Mean quantity ratings were typically within the low-to-middle range of the full possible range of scale scores. It should be noted that this does not necessarily reflect poor adherence, but rather, reflects the fact that many of the specific items within a particular scale were often absent during sessions (i.e., low quantity) because not every ABCT component was always prescribed or delivered in each session. Quantity ratings in this range are also consistent with other treatment integrity studies of well-controlled clinical trials (20, 35). For quality ratings, mean values tended to be near the mid-range of the scales, which is consistent with an average treatment quality that would be described as acceptable for a clinician who was well-trained in delivering ABCT.

Sources of Unreliability and Measurement Error

Overall, reliability data on the C-TIRS were similar to data reported by Carroll et al. (20) in the first iteration of their CBT treatment integrity rating system. We would expect that later refinements of the C-TIRS also will show improved reliability. Nonetheless, our experiences with the C-TIRS suggested several potential sources of unreliability that are briefly discussed here.

First, many C-TIRS items measured multidimensional or non-specific treatment elements. For example, SO quality ratings consisted of only two items assessing therapists' incorporation of SOs into the content and process of the session. These items may be easier to measure if they were more specific – for example, assessing SO involvement in each of the specific content and process domains in the treatment. This complication was evidenced anecdotally during ongoing coder meetings, and could be tested empirically in future work by modifying the scale content.

The C-TIRS also did not initially include behaviorally-specific anchors for each rating, (e.g., a list of the required clinician behaviors for being rated at each scale level for each item). Over the three-year coding period of the present study, we maintained a growing list of behavioral anchors in attempt to improve inter-rater reliability for future work.

Another measurement difficulty arose due to the present sample being restricted to clinical-trial therapists who were well trained and supervised in delivering ABCT. This likely restricted the overall range of therapist behaviors (i.e., few sessions with very poor ratings), which limited the amount of “signal” that could be detected. This restricted range may have attenuated reliability and also prevented further validity testing (e.g., comparing ratings for ABCT vs. non-ABCT sessions).

In addition, coders were graduate students with minimal couple therapy experience. Although they received extensive training in ABCT concepts and the C-TIRS, it is possible that more experience in ABCT, or couple therapy more generally, would have led to greater agreement among coders.

Clinical Implications

Although the present study yielded mixed findings about the reliability of the C-TIRS, it offers potential insights for clinical practice. First, we view this study as a stepping stone for developing a clinical tool that is reliable, valid, and able to be used on a broad scale for

assessing treatment integrity in ABCT. For example, an improved C-TIRS could provide useful information for supervisors in clinical practice settings about how well clinicians are delivering ABCT, which could facilitate clinical training and improve supervision of the treatment of couples. It should be noted, however, that the current version of the C-TIRS does not include cut-offs for categorizing individual therapists as proficient versus non-proficient in their delivery of ABCT, suggesting that C-TIRS scores should not be used to decide whether a therapist is “competent” in delivering ABCT.

In its present form the C-TIRS scales and items could serve as a content guide for clinicians who deliver or supervise ABCT. The C-TIRS domains and items were developed by experts in ABCT and are designed to capture the ingredients in ABCT that are thought to facilitate successful treatment outcomes. The scale could serve as a checklist for the components of ABCT that should be considered when planning, conducting, or reviewing ABCT treatment sessions. The C-TIRS also could be used as a guide for training clinicians in ABCT by providing a comprehensive listing of the elements of the treatment and helping clinicians to better understand the multi-faceted nature of ABCT.

Treatment integrity ratings could supplement other forms of information about therapists’ delivery of ABCT, such as clinician self-reports. However, treatment integrity ratings should be viewed in context with couples’ unique presentations (e.g., highly volatile couples or IPs who continue to drink throughout the course of treatment may require deviations from typical treatment protocols). In general, the information obtained through ratings of treatment integrity could provide relatively objective measures of treatment process that could be used in conjunction with other contextual and clinically relevant data.

Based on our experience with the present study, we caution clinicians and clinical supervisors to anticipate the challenges of properly and reliably assessing treatment integrity. This suggests that in the future, frontline providers might benefit from alternative methods to assess treatment integrity more pragmatically, accurately, and efficiently (36), which is consistent with the strategic plans of the Institute of Medicine (37) and the National Institute of Mental Health (38). Future measure development may attempt to “slim down” some of the scales and items that performed well in the present study, while reformulating others that performed less well. While the coders in the present study were highly trained and had an extensive period of time to code the study tapes, resources to train and code tapes are limited in many clinical settings, which may make reliable and valid coding even more difficult and reduce reliability further in those contexts. Alternatively, recent advances in computer-automated coding suggest that that efficient and accurate treatment integrity coding could eventually be achievable on a wide scale through machine learning tools (e.g., 39–40) that could be applied to the C-TIRS. This work in frontline settings has not yet been tested, and a computer-automated coding system for ABCT would still require a psychometrically sound coding system that can be used reliably by trained coders (39).

Limitations and Strengths

The present study has several limitations worth noting. During the training period, we were unable to achieve criterion reliability for quality ratings and instead only required a criterion level of reliability for quantity ratings in order to begin coding study tapes. The difficulty

with obtaining adequate quality ratings has not been reported in previous research (21) but was found both in the present study and with previous informal use of the C-TIRS. Given that the treatment integrity ratings were being used in the context of a larger study of mechanisms of behavior change in ABCT, the necessity to complete the goals of the larger study led to the decision to use C-TIRS quantity ratings alone for the main study.

The therapists rated in the present study received extensive training in ABCT and were instructed to follow specific ABCT protocols, potentially restricting the range of observed behaviors and reducing the generalizability of the results to community treatment providers. Generalizability is further reduced by the use of only first- and mid-treatment sessions, which were the only sessions consistently recorded and covering similar content across the parent studies. Thus, results should be interpreted with the caveat that therapist behaviors may have been different during sessions not captured in the present study.

The sample was predominantly Caucasian, which may seriously limit the validity of the results for other groups. For example, research on communication in different racial and ethnic groups has shown some differences between non-Hispanic whites and other diverse groups. While expressed emotion, or expressions of over-involvement (“smothering”), criticism or hostility vary from individual to individual; these expressions also vary cross-culturally (41). Moreover, the phenomenon of familial expressed emotion predicts different outcomes depending on the ethnicity of individuals; higher expressed emotion has been linked with poorer outcomes in non-Hispanic whites but not in Hispanic individuals in schizophrenia treatment (42–43). At present, there is a dearth of literature on minority couples in treatment. Future study should explore these potential differences. Additionally, the use of audiotapes precluded the ability to account for non-verbal behaviors, which have been associated with perceived treatment credibility and therapeutic alliance (43). Lastly, the lack of a comparison group prevented us from validating that the measure could distinctively identify ABCT compared to other forms of treatment.

The present study also has many strengths. Most significantly, the present study examined specific therapist behaviors during ABCT using a coding method that has not been examined in detail in previous literature. The coding instrument was developed from an individual-based treatment integrity instrument used in a well-known clinical trial of alcohol treatments, and to our awareness, is the first treatment integrity rating system for couple-based alcohol treatment. Coders received rigorous training in the conceptual model of ABCT and in the use of the C-TIRS. Participants were combined from four treatment studies, allowing for a larger sample compared to many other ABCT studies. The sample also included many female IPs, who are often under-studied in the alcohol treatment literature. The findings of fair to good reliability for three of the C-TIRS quantity scales suggest the potential for the measure to be improved and used in future research and training.

Summary and Future Directions

This study is the first to report on the development and psychometric properties of a treatment integrity measure designed specifically for ABCT, a well-validated and empirically-supported treatment for alcohol use disorder (12). As a first study of a new

treatment integrity rating system, results were similar to other such studies (e.g., 20). Further refinements to the C-TIRS will, hopefully, improve the reliability of the measure. We hope that researchers and clinicians can learn from both our successes and challenges and continue working toward developing improved treatment integrity measures that will enhance both clinical training and empirical research on ABCT.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was funded by NIAAA grant numbers R01AA03984, R01AA018376, R01AA007070, R37AA007070, T32AA007569, and T32AA018108.

References

- Centers for Disease Control and Prevention (CDC). Alcohol-Related Disease Impact (ARDI). Atlanta, GA: CDC; 2012.
- Epstein EE, Fischer-Elber K, Al-Otaiba Z. Women, aging, and alcohol use disorders. *J Women Aging*. 2007; 19(1–2):31–48. [PubMed: 17588878]
- McCrary BS. Treating alcohol problems with couple therapy. *J Clin Psychol*. 2012; 68(5):514–525. [PubMed: 22504611]
- Epstein EE, McCrary BS. Behavioral couples treatment of alcohol and drug use disorders: Current status and innovations. *Clin Psychol Rev*. 1998; 18(6):689–711. [PubMed: 9779329]
- Marshall MP. For better or for worse? The effects of alcohol use on marital functioning. *Clin Psychol Rev*. 2003; 23(7):959–997. [PubMed: 14624823]
- Whisman MA. Marital distress and DSM-IV psychiatric disorders in a population-based national survey. *J Abnormal Psychol*. 2007; 116(3):638–643.
- Rotunda RJ, O'Farrell TJ. Marital and family therapy of alcohol use disorders: Bridging the gap between research and practice. *Prof Psychol Res Practice*. 1997; 28(3):246–252.
- Maisto SA, McKay JR, O'Farrell TJ. Relapse precipitants and behavioral marital therapy. *Addictive Behaviors*. 1995; 20(3):383–393. [PubMed: 7653319]
- Tracy SW, Kelly JF, Moos RH. The influence of partner status, relationship quality and relationship stability on outcomes following intensive substance-use disorder treatment. *J Stud Alcohol*. 2005; 66(4):497–505. [PubMed: 16240557]
- McCrary BS, Stout R, Noel N, et al. Effectiveness of three types of spouse-involved behavioral alcoholism treatment. *British J Addict*. 1991; 86(11):1415–1424.
- McCrary BS, Epstein EE, Cook S, et al. A randomized trial of individual and couple behavioral alcohol treatment for women. *J Consult Clin Psychol*. 2009; 77(2):243–256. [PubMed: 19309184]
- Powers MB, Vedel E, Emmelkamp PG. Behavioral couples therapy (BCT) for alcohol and drug use disorders: A meta-analysis. *Clin Psychol Rev*. 2008; 28(6):952–962. [PubMed: 18374464]
- McCrary, BS.; Epstein, EE. Couple therapy in the treatment of alcohol problems. In: Gurman, AS.; Lebow, JL.; Snyder, DK., editors. *Clinical Handbook of Couple Therapy*. 5. New York: Guilford Press; 2015. p. 555–584.
- Fals-Stewart W, Birchler GR. A national survey of the use of couples therapy in substance abuse treatment. *J Subst Abus Treat*. 2001; 20:277–283.
- Epstein, EE.; McCrary, BS. *Overcoming Alcohol Use Problems: A Cognitive-Behavioral Treatment Program Therapist Guide*. 1. New York: Oxford University Press; 2009.
- Schumm JA, O'Farrell TJ, Kahler CW, et al. A randomized clinical trial of behavioral couples therapy versus individually based treatment for women with alcohol dependence. *J Consult Clin Psychol*. 2014; 92(6):993–1004. [PubMed: 25045910]

17. Walitzer K, Dermen K, Shyhalla K, et al. Couple communication among problem drinking males and their spouses: A randomized controlled trial. *J Fam Ther.* 2013; 35:229–251.
18. Perepletchikova F, Kazdin AE. Treatment integrity and therapeutic change: Issues and research recommendations. *Clin Psychol: Sci and Practice.* 2005; 12(4):365–383.
19. Campbell BK, Manuel JK, Manser ST, et al. Assessing fidelity of treatment delivery in group and individual 12-step facilitation. *J Subst Abus Treat.* 2013; 44:169–176.
20. Carroll KM, Connors GJ, Cooney NL, et al. Internal validity of Project MATCH treatments discriminability and integrity. *J Consult Clin Psychol.* 1998; 66(2):290–303. [PubMed: 9583332]
21. Morgenstern J, Morgan TJ, McCrady BS, et al. Manual-guided cognitive-behavioral therapy training: A promising method for disseminating empirically supported substance abuse treatments to the practice community. *Psychol Addict Beh.* 2001; 15(2):83–88.
22. Moyers TB, Martin T, Manuel JK, et al. Assessing competence in the use of Motivational Interviewing. *J Subst Abus Treat.* 2005; 28(1):19–26.
23. Project MATCH Research Group. Project MATCH: Rationale and methods for a multisite clinical trial matching patients to alcoholism treatment. *Alcohol Clin Exper Res.* 1993; 17:1130–1145. [PubMed: 8116822]
24. Carroll KM, Nich C, Sifry RL, et al. A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug Alcohol Depen.* 2000; 57(30):225–238.
25. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess.* 1994; 6(4):284–290.
26. Christensen A, Atkins DC, Berns S, et al. Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. *J Consult Clin Psychol.* 2004; 72(2): 176–191. [PubMed: 15065953]
27. Jacobson NS, Christensen A, Prince SE, et al. Integrative behavioral couple therapy: An acceptance-based, promising new treatment for couple discord. *J Consult Clin Psychol.* 2000; 68(2):351–355. [PubMed: 10780137]
28. McCrady BS, Noel NE, Stout RL, et al. Comparative effectiveness of three types of spouse involvement in outpatient behavioral alcoholism treatment. *J Stud Alcohol.* 1986; 47:459–467. [PubMed: 3795960]
29. McCrady BS, Epstein EE, Hirsch LS. Maintaining change after conjoint behavioral alcohol treatment for men: Outcomes at six months. *Addiction.* 1999; 94:1381–1396. [PubMed: 10615723]
30. McCrady, BS.; Epstein, EE.; Hallgren, KA.; Cook, S.; Jensen, NK. Women with alcohol dependence: A randomized trial of couple versus individual plus couple therapy. Manuscript under review
31. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 3. Washington, DC: Author; 1980.
32. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4. Washington, DC: Author; 1980.
33. McCrady, BS.; Epstein, EE. Overcoming alcoholism: A couples-focused program. New York: Oxford; 2009.
34. Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology.* 2012; 8(1):23–34. [PubMed: 22833776]
35. Tober G, Clyne W, Finnegan O, et al. Validation of a scale for rating the delivery of psychosocial treatments for alcohol dependence and misuse: The UKATT Process Rating Scale (PRS). *Alcohol Alcohol.* 2008; 43(6):675–682. [PubMed: 18818190]
36. Schoenwald SK, Garland AF, Chapman JE, et al. Toward the effective and efficient measurement of implementation fidelity. *Admin Pol Mental Health Mental Health Serv Res.* 2011; 38(1):32–43.
37. Institute of Medicine. [Accessed July 24, 2015] Developing evidence-based standards for psychosocial interventions for mental disorders. 2015. Available at: <http://iom.nationalacademies.org/Activities/MentalHealth/PsychosocialInterventions.aspx>

38. National Institute of Mental Health. [Accessed July 24, 2015] Strategic objective 4: Strengthen the public health impact of NIMH-supported research. 2015. Available at: <http://www.nimh.nih.gov/about/strategic-planning-reports/strategic-objective-4.shtml#s44>
39. Atkins DC, Steyvers M, Imel ZE, et al. Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Impl Sci*. 2014; 9(49):1–11.
40. Xiao, B.; Bone, D.; Van Segbroeck, M. [Accessed July 2014, 2015] Modeling therapist empathy through prosody in drug addiction counseling; *Proceedings of Interspeech-2014*. 2014. p. 213-217. Available from http://www.isca-speech.org/archive/archive_papers/interspeech_2014/i14_0213.pdf
41. López SR, Ramírez García JI, Ullman JB, et al. Cultural variability in the manifestation of expressed emotion. *Fam Proc*. 2009; 48(2):179–194.
42. Breitborde NJK, López SR, Kopelowicz A. Expressed emotion and health outcomes among Mexican-Americans with schizophrenia and their caregiving relatives. *J Nerv Mental Dis*. 2010; 198(2):105–109.
43. Dowell NM, Berman JS. Therapist nonverbal behavior and perceptions of empathy, alliance, and treatment credibility. *J Psychother Integrat*. 2013; 23(2):158–165.

Table 1

Example C-TIRS Items

Scale	Item	Description
CBT Components		
	1	Assessment of client's drinking, cravings, and relationship satisfaction
	5	Cognitive-behavioral skill training
	14	Enhancing motivation to change
SO Components		
	18	Partner content (e.g., partner-related triggers, interventions, decision-making)
	19	Partner process (e.g., encouraging partner to take active role in supporting abstinence)
Couple-Specific Components		
	20	Reciprocity enhancement
	21	Communication skills
	26	Addressing relational issues that arise in session
Common Therapeutic Factors		
	27	Consistency of problem focus
	35	Empathy
General Adherence		
	37	General adherence to the prescribed content and structure

Table 2

Descriptive Statistics for C-TIRS Scales

	Possible Range	Quantity			Quality		
		M(SD)	Min	Max	M(SD)	Min	Max
First Session							
CBT Components	5–85	31.59(6.41)	15	47	44.78(5.70)	18	60
SO Components	0–10	5.45(1.73)	2	9	5.93(1.59)	2	10
Couple-Specific Components	7–35	15.24(3.65)	7	26	21.43(2.76)	13	31
Common Therapeutic Factors	7–50	31.66(4.82)	19	44	32.37(5.41)	20	47
General Adherence	1–5	3.70(0.76)	2	5	3.46(0.82)	1	5
Mid-Treatment Session							
CBT Components	5–85	36.55(8.42)	18	57	48.54(8.05)	20	69
SO Components	0–10	5.40(1.91)	2	10	5.69(1.78)	2	10
Couple-Specific Components	7–35	20.39(4.57)	9	30	22.06(4.10)	11	32
Common Therapeutic Factors	7–50	32.59(5.24)	21	46	32.91(6.07)	13	47
General Adherence	1–5	3.34(0.98)	1	5	3.10(1.06)	1	5

Table 3

Inter-Rater Reliability for C-TIRS Scale Ratings

	Quantity Ratings	Quality Ratings
CBT Components	.72	.45
SO Components	.49	.37
Couple Components	.59	.17
Common Factors	.28	.46
General Adherence	.50	.43

Note. Values are two-way, single-measures, consistency intra-class correlations.

Table 4

Correlations Among C-TIRS Scales

C-TIRS Scale	1.	2.	3.	4.	5.
1. CBT Interventions		.61	.63	.62	.60
2. SO Interventions	.52		.58	.55	.61
3. Couple Interventions	.42	.33		.74	.63
4. Common Factors	.51	.40	.40		.75
5. Overall Adherence	.46	.34	.01	.49	

Note. Values below the diagonal indicate correlations among quantity ratings, values above the diagonal indicate correlations among quality ratings. All correlations significant at $p < .001$ except the correlation between quantity of couple-interventions and overall adherence.