Published in final edited form as:
Psychother Res. 2011 May ; 21(3): 252–266. doi:10.1080/10503307.2010.551429.

A generalizability theory analysis of group process ratings in the treatment of cocaine dependence

PAUL CRITS-CHRISTOPH[1], **JENNIFER JOHNSON**[2], **ROBERT GALLOP**[3], **MARY BETH CONNOLLY GIBBONS**[1], **SARAH RING-KURTZ**[1], **JESSICA L. HAMILTON**[1], and **XIN TU**[4]

[1]University of Pennsylvania, Psychiatry, Philadelphia, USA

[2]Brown University, Psychiatry, Providence, USA

[3]West Chester University, Mathematics, West Chester, USA

[4]University of Rochester, 1Department of Biostatistics and Computational Biology, Rochester, USA

Abstract

Videotaped group drug counseling sessions were rated for alliance, self-disclosure, positive and negative feedback, group cohesion, and degree of participation of each group member. Interrater reliability was good to excellent for most measures. However, generalizability coefficients based on statistical models that included terms for patient, counselor, session, group, and rater revealed that some measures had inadequate dependability at the patient level if only two raters and two sessions were used to create patient-level scores. In contrast, good generalizability coefficients based on two raters and two sessions were obtained for alliance, non-positive learning statements received from counselor, participation variables, and self-disclosures about the past. The implications of the findings for the design of process-outcome studies are discussed.

Keywords

group psychotherapy; process research; alliance; substance abuse

Group therapy consists of a wide range of different approaches with the common element of a group format. Although groups differ in the extent to which they emphasize various processes of change, certain group process variables have been found to predict outcome or are assumed to be operative in most kinds of group treatments. These variables include the therapeutic alliance (see review by Johnson, Burlingame, Strauss, & Bormann, 2008), degree and quality of participation (i.e., degree of self-disclosure) in group therapy (Bloch, Crouch, & Reibstein, 1981; Coche, Dies, & Goettelmann, 1991; Coyne & Silver, 1980; Crouch, Bloch, & Wanlass, 1994; Lundgren & Miller, 1965; Tschuschke & Dies, 1994; Tschuschke, MacKenzie, Haaser, & Jaanke, 1996), amount of feedback provided to group members **(**Claiborn, Goodman, & Horner, 2001; Kivlighan, 1985; Morran, Stockton, & Bond, 1991), and group cohesion (see review by Burlingame, Fuhriman, & Johnson, 2002). Investigating the role of these types of group process variables in different kinds of group therapies can provide important information on the mechanisms of change for a specific variant of group therapy. The findings of such process studies can then be used to modify and enhance the therapy so that better treatment outcomes are obtained.

Correspondence concerning this article should be addressed to P. Crits-Christoph, University of Pennsylvania, Psychiatry, Philadelphia, USA. crits@mail.med.upenn.edu.

A potentially useful approach to the examination of process ratings of psychotherapy sessions is generalizability theory (Brennan, 2001; Cronbach, Nageswari, & Gleser, 1963; Hoyt, 2002; Shavelson & Webb, 1991; Wasserman, Levy, & Loken, 2009). Generalizability theory addresses the adequacy with which one can generalize from a sample of observations to a universe of observations from which the sample was randomly drawn. For example, generalizability theory can be used to examine interjudge reliability (Shrout & Fleiss, 1979). By incorporating multiple sources of error into reliability coefficients, reliability estimates calculated using generalizability theory are likely to be more accurate. This issue is particularly relevant to ratings of psychotherapy process because multiple sources of error are common, such as variation due to patient, session, the counselor, the rater, as well as other potential factors.

Assessing sources of variability in process ratings is critical to group psychotherapy research, which often poses questions about variation between group members within a group, or variation in group or individual-level variables over time. If the research goal is to obtain a stable estimate of patients' typical scores on a group process measure in order to relate the process measure to patient-level treatment outcome, it is important to know how much variability exists from session to session on the measure. Generalizability theory can be used to measure this variability and to determine how many sessions need to be sampled to yield an accurate overall level for each patient. If session to session variability is near zero, then a single session might likely be sufficient to measure a patient's typical level on a variable. Conversely, if session to session variability is very high, a large number of sessions may need to be evaluated to obtain a stable estimate of a patient's typical level. Similarly, the existence of significant variability from counselor to counselor might guide the decision to employ multilevel modeling (i.e., patients nested with counselor) when examining the relation of a process variable to treatment outcome. Thus, generalizability theory analysis of process ratings can inform the design of future studies through the specification of the number of raters, sessions, and/or counselors needed to investigate the role of process variables in both short and long-term outcomes.

Despite the usefulness of generalizability theory analysis of process ratings, we could locate no studies that have applied this approach to the study of group therapy to examine variability due to patients, raters, sessions, or counselors (and potential interaction effects between these factors). Although generalizability theory methods have not been used, there have been studies of process ratings over the course of group therapy that have documented average changes over sessions. For example, in a study of group cognitive behavioral therapy for social phobia, Woody and Adessky (2002) found that patients' individual alliances with the group leaders increased over sessions, while patients' individual sense of group cohesion did not. Similarly, Brossart, Patton, and Wood (1998) used growth curve modeling to document changes in the Group Climate Questionnaire (MacKenzie, 1983) over group therapy sessions conducted at a university counseling center, and Kipnes, Piper, and Joyce (2002) used both self-report and observer ratings of group cohesion to evaluate changes over sessions in psychodynamic groups for complicated grief. None of these studies, however, reported an index of the degree to which their assessments were stable (or could be generalized) across time, groups, and/or patients by providing a generalizability coefficient or reporting how many sessions would be needed to adequately assess a patient variable. Without assessing the dependability of a measure using generalizability theory analyses, the adequacy of such measures as predictors of other patient-level variables, such as outcome, is not known.

The goal of the current study was to conduct a generalizability theory analysis of group process variables that have been found to predict treatment outcome for group treatments across a range of patient populations. These variables included alliance, feedback, self-

disclosure, and cohesion. We used observer process rating data from group drug counseling for cocaine dependence to demonstrate the utility of generalizability analyses for group treatment studies and to plan future studies on the process of group drug counseling.

Methods

Overview

Videotaped sessions of group drug counseling drawn from the National Institute on Drug Abuse Cocaine Collaborative Treatment Study (NIDA CCTS; Crits-Christoph et al., 1999) were used to evaluate the generalizability of process ratings. The NIDA CCTS was a randomized multicenter clinical trial that compared four manual-guided treatments for cocaine dependence: individual drug counseling (IDC) plus group drug counseling (GDC), cognitive therapy (CT) plus GDC, supportive expressive (SE) psychodynamic therapy plus GDC, and GDC alone. The primary finding was that IDC+GDC achieved superior results compared to the other three treatments on a composite drug use outcome measure (Crits-Christoph et al., 1999). It is noteworthy that all four of the treatment groups in the NIDA CCTS achieved very positive outcomes, and all groups included manual-based group drug counseling. Furthermore, the combination of individual plus group drug counseling was not superior to group alone on reduction in use of cocaine from baseline to the month 12 assessment despite the fact that the patients who received IDC+GDC had substantially more treatment sessions compared to those who receive GDC alone. Thus, it appears that the group counseling may have been crucial to the success of these treatments.

Participants

Patients—A total of 487 patients were randomized to the four treatment modalities in the NIDA CCTS. For the current report, patients who were randomized to these treatment modalities but did not attend any group sessions were excluded. In addition, patients who could not be identified on videotapes (primarily because they were heard but were out of the camera view) were not included. Otherwise, all members present in a selected group session were rated for the process variables.

For the NIDA CCTS as a whole, the patient sample consisted of individuals aged 18–60, with a diagnosis of cocaine dependence, who had used cocaine at least once in the past 30 days and reported a stable living situation. Patients were excluded if they were diagnosed with current opioid dependence or opioid dependence in early partial remission, dementia or other irreversible organic brain syndrome, evidenced psychotic symptoms, were a current imminent suicide or homicide risk, or had a life threatening or unstable medical illness. Patients were also excluded if they were unwilling to discontinue a current psychotherapeutic treatment, had an impending incarceration, were hospitalized > 10 days in the past 30 days for index episode of cocaine use, were currently mandated for treatment by legal or Children & Youth Services, or resided in a halfway house at time of screening. All patients who participated in the NIDA CCTS provided written informed consent.

Counselors—Ten group drug counselors participated in the original study. Of these 10, eight were men, eight were Caucasian, and three had a master's degree (the rest had Bachelor's or Associates degrees). The counselors had an average of 6.9 years of clinical experience and were 42.6 (range: 30-62) years of age.

Group Drug Counseling Treatment

The GDC treatment (Dailey, Mercer, & Carpenter, 1999) is designed to educate clients about the important concepts in addiction recovery and to provide a supportive group atmosphere in which members can express feelings, discuss problems and learn to draw

strength from one another. GDC relies heavily on group support, but has a psychoeducational format, with 12 standard rotating sessions covering symptoms of cocaine addiction, the process of recovery (two sessions), managing craving, relationships in recovery, self-help groups, establishing a support system, managing feelings in recovery, coping with guilt and shame, warning signs of relapse, coping with high-risk situations, and maintaining recovery. Although group drug counseling is a specific form of group counseling tailored to problems with substance abuse/dependence, such counseling groups are the primary mode of addiction treatment in the U.S. (SAMHSA, 2010) and would also be expected to share a number of non-specific elements with general (non-substance abuse) group therapy approaches. The GDC model also encourages participation in 12-Step self-help recovery programs such as Cocaine Anonymous (CA), Narcotics Anonymous (NA), and Alcoholics Anonymous (AA).

Treatment was 6 months in duration. Group Drug Counseling sessions (1.5 hours) were held once a week for the treatment period. Treatment was free of charge. Group membership was rolling—that is, members were added to the group as they were recruited into the study. Assignment to a specific group was based on the fit between group and patient schedules. Each of the participating five sites of the NIDA CCTS generally had two or three groups running simultaneously, for 14 total groups during the course of the study. Group attendance on any given day averaged 4.43 members (range = 1 to 11). If the counselor was not available, the group session was not held. For the NIDA CCTS sample as a whole (n = 487), the average number of group sessions attended was 8.6 (SD = 7.2) for IDC+GDC, 9.5 (SD = 7.2) for CT+GDC, 8.8 (SD = 6.8) for SE+GDC, and 8.6 (SD = 7.2) for GDC alone.

Treatment Process Measures

Therapeutic alliance—The alliance was assessed using the observer version of the Working Alliance Inventory (WAI; Horvath & Greenberg, 1986). We employed the Raue, Goldfried, and Barkham (1997) version of the observer WAI, since it includes a formal rater manual and has recently been shown to have good convergent validity with patient and therapist perspectives on the alliance (Stiles et al., 2002). The WAI is a 36-item instrument, with each item rated on a 7-point Likert-type scale. Based upon Bordin's (1979) general definition of the alliance, this scale assesses the patient's affective bond with the therapist and the agreement between patient and therapist on the goals and tasks of treatment. We evaluated each member's alliance with the group counselor (not the alliance of the group as a whole).

Research has found strong support for the reliability and validity of the WAI (Horvath, 1994). Using the Raue et al. (1997) version, Stiles et al. (2002) report high internal consistency reliability for the three subscales (bond: .95; tasks: .94; goals: .94). Interrater reliability for the WAI observer scale has been reported to be .75 (for two judges pooled) (Tang & DeRubeis, 1999). Of standard measures of the alliance, the WAI – observer version was the most highly correlated (r = .48) with treatment outcome in a sample of cocaine/alcohol dependence patients receiving 12-step facilitation treatment (Fenton, Cecero, Nich, Frankforter, & Carroll, 2001).

Quality of participation in group—The quality of participation was assessed by the frequency of self-disclosure statements by the patient. To code self-disclosure of patient statements, we employed the response mode coding system developed by Gibbons et al. (2002) adapted to patient statements rather than therapist statements. Self-disclosures were defined as statements that reveal something personal about the patient's experiences or feelings and were divided into two types: (1) “here-and- now” self-disclosures that were statements reflective of a current emotional reaction to the group or group members, and (2)

statements that described emotionally significant events in the past. Simplistic statements about drug use, such as “I used last week,” were not coded as self-disclosures. Gibbons et al. (2002) report a reliability of .71 for self-disclosure at the level of the individual statement for three judges pooled.

Feedback—Patient and counselor statements toward each targeted group member were coded using the Gibbons et al. (2002) system, with feedback indexed as the frequency of “learning statements” in the session towards each relevant patient. Learning statements were defined as any statement that helped the patient become aware of a thought, feeling, or behavior. Such learning statements might simply point out a patient's thoughts, feelings, or behaviors, describe a causal link between a thought, feeling, or behavior, or describe a pattern of behaviors or link past behavior to present behavior. In the Gibbons et al. (2002) study, the interjudge reliability for coding individual statements as learning statements was .77. At the level of the session, the intraclass correlation (three judges pooled) was also .77. In terms of validity, significant differences between therapists and between treatment modalities were found for the frequency of “learning statements” in the Gibbons et al. study, suggesting that the measure was sensitive to variations in the therapeutic process.

The extent to which feedback is positive was coded using the “approval” category from the Hill Counselor Verbal Response Modes Category System (Hill, 1986). In this system for coding statements in psychotherapy sessions, “approval” statements are defined as those that provide emotional support, approval, reassurance, or reinforcement. Approval may imply sympathy or tend to alleviate anxiety by minimizing client's problems. Reliability for coding therapist statements using this system was previously reported to be .67 (per judge reliability as assessed by Kappa) (Hill et al., 1988).

Level of participation in group—Measures of quantity of participation included total time talking (excluding counselor talk time) and number of turns-at-talk in each group session by each patient. Turns-at-talk that were less than 3 seconds in duration (e.g., brief utterances such as “uh-huh”) were excluded from analyses.

Group cohesion—Because groups were rolling, with membership changing constantly (new members added and existing members dropping out), it was not anticipated that a consistent sense of group cohesion would emerge and continue over time. However, we viewed this as an interesting empirical question and therefore evaluated group cohesion in an exploratory way on a subset of tapes. We used the Harvard Health Plan Group Cohesiveness Scale (Budman et al., 1982, 1989) to evaluate overall cohesion for each rated group session. For the purposes of this scale, cohesion is defined as group connectedness as evidenced by working together towards a common therapeutic goal, constructive engagement around common themes, and openness to sharing personal material. Separate ratings are made by trained judges on the following dimensions: focus, interest/involvement, trust, facilitative behavior, and bonding, as well as a global cohesiveness rating. Each of these dimensions is rated on a 1 (very slight) to 9 (very strong) scale. Three additional variables that bear on cohesiveness are also rated: affective intensity, conflict, and global quality (capturing therapist interventions and unusual events).

Raters and Procedures

Alliance, feedback, and self-disclosure—Five judges who were trained, experienced clinicians with a Master's or Ph.D. degree were hired to provide expert judgment and ratings of the observed sessions on the alliance, feedback, self-disclosure, and cohesion instruments. All raters had previously worked as judges in psychotherapy studies.

Judges worked independently as they were trained and as they rated the main study tapes. The 2-month training consisted of review and discussion of instruments, rating of five training tapes, and discussion of discrepancies between judges. While rating actual study sessions, judges participated in monthly (total of 12) recalibration sessions to maximize reliability and prevent rater drift. Non-study tapes were used for these recalibration sessions. The judges completed ratings of 387 separate GDC videotapes containing a total of 417 patients. Each patient was present in an average of 3.86 separate group sessions. Each tape was rated by two judges using a balanced incomplete block design to assign judges to tapes.

Participation—Because the participation variables were based on the number of turns at talk and time of speaking, it was assumed that non-clinically trained judges could accomplish this task. A total of 27 undergraduate students were hired and trained to code participation in the GDC sessions. Following training, these judges participated in monthly meetings to address any problems that may have arisen. These judges independently coded participation in 1030 group sessions that included a total of 440 patients, with each patient present in an average of 7.5 ($SD = 5.6$) group sessions. One judge coded each session, with the exception that a second judge coded 57 of the 1030 tapes, which included a total of 119 patients.

Cohesion—Cohesion ratings were made by two Ph.D. clinicians (included in the five who rated alliance, feedback, and self-disclosure) who had clinical and research experience with group therapy approaches. Training and re-calibration for the cohesion ratings was done using the same steps and procedures as was utilized for the alliance, feedback, and self-disclosure ratings. A total of 76 GDC session videotapes were randomly chosen to be rated for cohesion from among the 387 tapes rated on the other process variables.

Statistical Analysis

Preliminary analyses and graphical displays examined the distributions of the process variables. The extent to which the distribution for each variable deviated from a normal distribution was tested using the Shapiro-Wilk Statistic. The Shapiro-Wilk (1965) test statistic ranges from 0 to 1, with larger values indicating a more near normal measure. The *p*-value assesses whether the data deviate significantly from normality but is extremely sensitive to sample size (Metz, Haccou, & Meelis, 1994). Therefore, we used the following thresholds to indicate substantial deviations in normality: .950 for large sample sizes ($n > 500$), .900 for moderate sample sizes (between 100 and 500), and .850 for small sample sizes ($n < 100$). When the Shapiro-Wilk statistic is below the respective threshold based on the effective sample size, it indicates substantial deviations in normality. For variables that were non-normally distributed, appropriate transformations were then conducted based on a Box-Cox transformation analysis (Box & Cox, 1964).

The primary statistical analysis was guided by generalizability theory. Generalizability theory provides a framework within which multiple sources ("facets") of variability in a given set of measurements can be simultaneously estimated by including these multiple sources in an analysis of variance design and generating variance components for each random effect. In this study, we partitioned variation in alliance (WAI total score), feedback, self-disclosure, and participation ratings into variation due to rater, counselor, patient (within counselor), session (within patient; this effect coded for each of the appropriately four rated sessions attended by each patient), group (this effect coded for the specific group on a given day that a patient attended and indexed the extent to which patients within a given group, on a given day, were more similar than patients in a different group session; using this definition, the number of specific groups was the number of tapes, e.g., 387 for alliance ratings), and the following interactions: patient by session, patient by rater, group by

rater, and group by patient. All of these sources of variance were specified as random effects. Other models were also attempted that specified more interactions, including three-way and four-way interactions, but a statistical solution with these models could not converge due to the high level of nesting and overlap among the sources of variation. These other sources of variation (interactions) are therefore contained within the residual variation.

Variance components were calculated with the SAS Proc Mixed procedure (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006) using the restricted maximum likelihood estimation method (REML) to estimate the variance components. As discussed by Swallow and Monohan (1984) there are two general classes of variance component estimators: likelihood based approaches and the non-likelihood classical ANOVA-type estimation methods, i.e., minimum variance quadratic unbiased estimation (MIVQUE). SAS Proc Mixed will implement variance component estimation through the likelihood based approaches (maximum likelihood and REML) and MIVQUE approach. We proceeded with the REML based approach due to its useful properties such as yielding unbiased estimates, consistency, asymptotic normality, and efficiency (Verbeke & Molenberghs, 2000). In addition, Swallow and Monahan (1984) provide simulation evidence favoring REML and maximum likelihood over MIVQUE (Littell et al., 2006).

Variance components estimate the population covariation between random factors (e.g., sessions) and the dependent variable (e.g., alliance). Unlike a correlation coefficient (which is standardized on a scale from −1 to +1), variance components are expressed in units of the dependent variable. The test of the statistical significance of a variance component examines whether it is different from zero (i.e., no meaningful variation due to the specified effect) and is based on a mixture of Chi-square distributions (Verbeke & Molenberghs, 2000). The SAS code for the Proc Mixed procedure was:

```
proc mixed data = group_dataset covtest method = reml;

class patient group counselor session rater; model alliance = ;

random patient group counselor session rater patient * rater patient * session group
* rater group * patient;
```

An interclass correlation coefficient (ICC) was calculated that estimated the single judge reliability for rating individual sessions. The ICC was calculated as given in this formula:

$$\frac{\widehat{\sigma}^2_{TOT} - \left(\widehat{\sigma}^2_{Rater} + \widehat{\sigma}^2_{Patient\times Rater} + \widehat{\sigma}^2_{Group\times Rater} + \widehat{\sigma}^2_{Residual}\right)}{\widehat{\sigma}^2_{TOT}}$$

where $\widehat{\sigma}^2_{TOT}$ is the sum of the following estimated variance components: $\widehat{\sigma}^2_{Patient}$, variance attributable to patients; $\widehat{\sigma}^2_{Rater}$, variance attributable to rater; $\widehat{\sigma}^2_{Patient\times Rater}$, variance attributable to the patient by rater interaction; $\widehat{\sigma}^2_{Group\times Rater}$, variance attributable to the group by rater interaction; $\widehat{\sigma}^2_{Residual}$, variance attributable to the error variance (i.e., not attributed to other factors in the design). This ICC formula above corresponds to the McGraw and Wong (1996) ICC (A, 1) model (one judge) that assesses reliability based on the absolute level, rather than the relative ranking, of scores.

The reliability of two judges averaged was also calculated by dividing the sum of the Rater terms plus residual by 2 in the above formula. This corresponds to McGraw and Wong's (1996) ICC (A, 2) model.

Generalizability coefficients were calculated to examine the expected between-patient dependability for study designs in which various numbers of sessions for each patient are rated by each of two judges, using a relative decision rule (e.g., relative rankings of scores; Wasserman, Levy, & Loken, 2009). The term dependability is used to describe the accuracy of generalizing from a person's observed score on a given construct to the ideal mean score a person would have received across all relevant observation contexts. In the same way an interjudge reliability coefficient examines how well scores generalize across judges, a generalizability coefficient examines how well scores generalize across relevant observation contexts (e.g., sessions, counselors). With regard to sessions, for example, the generalizability coefficient provides an index of whether or not the scores from one session can be assumed to be representative of other sessions and therefore can be used to examine individual differences between patients on a specific measure. Generalizability coefficients can also be calculated that indicate the extent to which averaging two (or any number) of sessions creates a more dependable score. The generalizability coefficient for a design with one rated session was calculated as the ratio of the patient variance component to the sum of all variance components involving patient, as given in the following formula:

$$\frac{\widehat{\sigma}^2_{Patient}}{\widehat{\sigma}^2_{Patient}+\frac{\widehat{\sigma}^2_{Patient\times session}}{n_{session}}+\frac{\widehat{\sigma}^2_{Patient\times Group}}{n_{Group}}+\frac{\widehat{\sigma}^2_{Patient\times rater}}{n_{rater}}+\frac{\widehat{\sigma}^2_{Residual}}{n_{group}\times n_{rater}}}$$

where $n_{session}$ is the average number of sessions per patient, n_{Group} is the number of specific groups (i.e., tapes), and n_{rater} is the number of raters. A minimum generalizability coefficient of .70 is generally thought to be acceptable for observational ratings (Allen & Yen, 1979), though a higher level (e.g., above .80) is preferred as is true for reliability coefficients (Cardinet, Johnson, & Pini, 2010).

For the participation variables, generalizability coefficients were calculated using the larger sample of tapes on which only a single judge coded participation, and therefore the rater term was not included in the statistical model. For the cohesion variables, generalizability coefficients without a separate patient by group term were not included in the model because group was the unit of analysis (i.e., the residual).

Results

Baseline Demographic and Clinical Characteristics of Sample

Table I provides baseline demographic and clinical data for the samples of patients rated on these process variables, as well as the characteristics of the full sample of 487 randomized patients from the NIDA CCTS. Comparisons of each of the respective subsamples to those patients who were excluded in the subsamples revealed that the Alliance/Feedback sample (n = 417) were slightly older than those excluded (n = 70) from this sample ($t(485) = 1.97$, p = .049). No other significant differences were apparent for the samples of patients included, versus excluded, from the total randomized sample.

Mean Levels and Distributions of Process Variables

Means and standard deviations of the variables (untransformed), Shapiro-Wilk tests for deviation from normality, and transformations applied are provided in Table II. The variables that were counts of learning statements or self-disclosures all had non-normal distributions that were best transformed into binary variables (presence or absence of that type of statement in a session) due to a large stack of zero responses. The participation

variables and most of the cohesion ratings did not deviate significantly from a normal distribution, though two did (Trust and Conflict) and needed log transformations to correct.

Sources of Variability in Process Ratings

Table III presents variance components and percent of total variance for the process ratings; the statistical significance of the variance components is presented at both the .05 and .01 levels (the latter are more meaningful given the large number of tests conducted in a generalizability theory design). In general, variance components for patient differences were statistically different from zero, while differences between sessions were typically small and non-significant. For example, patient differences represented 13% of the total variance for the WAI total score, 23% of the variance for past self-disclosures, and 19% of the variance in percent time speaking. Across the nine group cohesion rating scales, patient variance was 16% to 34% of the total variance. Counselor effects ranged from 1% (number of self-disclosures in the here-and-now) to 14% (WAI total score) and none were significant at the .01 level. Rater differences were generally about 5% of the total variance, although for some scales a higher percent (10–14%) of total variance was evident (number of positive learning statements received from other patients; group cohesion scales: focus, interest/involvement, conflict). Group effects were highly variable, with some scales showing large group effects (percent time speaking: 40%) and other scales showing minimal group effects (number of non-positive learning statements received from other patients: 1%).

In addition to the main effects described above, several interaction terms were of note. For some of the variables, the Patient × Group interaction term was statistically significant and nearly as large, if not larger, in magnitude than the Patient variance component. This Group by Patient interaction indicates that some patients scored similarly to other group members in some sessions but not others, while other patients tended to consistently be similar to or different from their group-mates. Patient by Session interactions, indicative of scores varying over sessions for some patients but not for others, were generally small, an exception being non-positive learning statements from other patients, for which 17% of the total variance was attributed to this interaction.

The Patient by Rater interaction examines the extent to which the raters had better agreement for some patients than for others. For the most part, these effects were small. However, the Focus and Trust group cohesion ratings had larger Patient by Rater interaction effects (Table III).

Group by Rater effects indexed the extent to which the raters varied in their ratings for some specific groups, but not for others. Significant Group by Rater interactions were evident for all variables with the exception of non-positive learning statements received from counselors (note: Group by Rater effects could not be examined for the cohesion variables). Relatively large Group by Rater effects were found for the WAI total score and number of positive learning statements received from counselor (20% and 19%, respectively, of total variance explained by the Group by Rater effects).

Interjudge Reliability of Process Ratings

Interjudge reliabilities, calculated from the variance component model, were generally good to excellent at the level of the individual observations (tapes) (Table IV). As expected, the interjudge reliability of the participation variables was very good (per judge reliability of .92 for percent time speaking and .88 for number of turns at talk), justifying our decision to use only one judge for the bulk of the tapes coded. Similarly, the interjudge reliability values of the group cohesion rating scales were all high, with eight or nine scales achieving a reliability (two judges combined) of .90 or higher. However, interjudge reliability of coding

learning statements and self-disclosures was variable. While the reliability (two judges pooled) for coding non-positive learning statement was adequate, the reliability of coding positive learning statements, particularly from other patients, was not (.59). Judges agreed relatively easily on the coding of self-disclosures in the past (reliability of two judges pooled: .85), but less so on self-disclosures in the here-and-now (reliability of two judges pooled: .64).

Generalizability Coefficients

Generalizability coefficients that index the ability of the rating scales to dependably discriminate between patients are given in Table V for a variety of potential study designs across a range of number of raters and number of sessions. Based on using two raters, each rating two sessions, these coefficients were good (> .80) for the alliance, number of non-positive learning statements received from the counselor, number of self-disclosures concerning the past, and the participation variables. However, generalizability coefficients based on two raters and two sessions were unacceptably low (below .70) for the positive learning statement scores, non-positive learning statements from other patients, self-disclosure in the here-and-now, and all of the group cohesion scales.

For most of the variables that had inadequate generalizability coefficients based on two raters and two sessions, increasing the number of raters and/or sessions would boost generalizability coefficients to acceptable levels (Table V). For example, six of the nine group cohesion scales achieved adequate (>.70) generalizability coefficients if four raters and two sessions were to be used, and the remaining three scales became adequate if four sessions and four raters were to be used. To achieve an adequate generalizability coefficient for the number of non-positive learning statements received from other patients, four raters and eight sessions would be required. An adequate generalizability coefficient would be achieved for self-disclosures in the here-and-now with either eight judges rating four sessions or four judges rating eight sessions. However, even with eight raters and 12 sessions coded, the numbers of positive learning statements from counselors and from other patients do not reach acceptable levels.

Item Analysis of WAI

Because the WAI was a multi-item scale, there was the possibility that the large Group by Rater interaction found for the WAI total score may have been driven by variability among items, with raters finding it easier to rate certain items, compared to other items, when rating certain groups. We attempted to add an item facet to the generalizability theory analysis described above in order to evaluate any potential Group by Rater by Item interaction. However, convergence could not be obtained with this model due to the added complexity of the model involving two-way and three-way interaction terms with the item facet, as well as reduced variability for the item scores compared to the total scale, and therefore variance components could not be estimated. Consequently, to evaluate items on the WAI, we calculated interjudge reliabilities and generalizability coefficients for each of the WAI items using the same methods as were implemented for the WAI total score. Calculation of interjudge reliability for the 36 WAI items revealed that reliability was highly consistent across items, with 35 of the 36 items displaying *ICC* (A, 2) values between .70 and .80, with the remaining item higher than .80. Generalizability coefficients were also acceptable (≥ .70) for the majority of items (29 of 36). One WAI item (the client is aware that the therapist is genuinely concerned for his/her welfare) had a generalizability coefficient of .53, and the remaining six items had generalizability coefficients between .64 and .70.

Discussion

The results of this generalizability study indicate that most group process variables can be rated with a high degree of interjudge reliability from videotapes of group drug counseling sessions. In addition, across all measures, patient variability (patient main effects plus patient by session and patient by group interactions) was generally larger than session, counselor, or rater variability, leading to adequate to good generalizability coefficients for many variables based on simple study designs (two to four raters; two sessions). The good interjudge reliability and ability to measure meaningful patient-to-patient variability using these scales indicates that it is possible to study process constructs (e.g., alliance, feedback, self-disclosure, group cohesion) in the context of group drug counseling. Thus, while group drug counseling is often perceived as something different from standard group therapy, the current data suggest that it is possible to investigate this treatment approach in terms of group therapy constructs. These findings take on particular importance because group drug counseling is such a widely used clinical modality but attracts very little research. Some version of group drug counseling is offered by 93% of the drug treatment programs in the United States (SAMHSA, 2010).

Despite the good to excellent overall interjudge reliability for most of the scales, Group by Rater interaction effects were apparent for several scales, suggesting that agreement between raters was better for some groups than for others. Similarly, in a generalizability study Hoyt (2002) found large Therapist by Rater effects when the therapist was the target of ratings. There may be characteristics of certain group sessions that make rater agreement especially difficult. For example, if the group leader dominates discussion in a given group session, raters may find it more difficult to evaluate the alliance between individual patients and the group leader compared to a group session in which all patients participated. An agenda for future research might be to examine the relation between group characteristics and degree of rater agreement within each group to further understand how specific group factors influence rater agreement. Such result could be used to improve rater manuals and rater training in ways that decrease the amount of disagreement that can occur with specific types of groups.

In the current study, interjudge reliability was assessed based on the absolute level rather than the relative ranking of judges' scores. It should be noted that modeling the reliability of a relative vs. absolute decision rule is not unique to generalizability theory analysis. Standard intraclass correlation coefficient analyses also can be conducted using a relative or absolute decision rule (McGraw & Wong, 1996; Shrout & Fleiss, 1979). Decisions about whether reliability should be based on the absolute level or the relative ranking of judges' scores depends on the specific research question and the potential uses of the measurement scale.

Although meaningful patient variability was evident for many of the scales, Patient by Rater, Patient by Group, and Patient by Session interaction effects were also apparent for some variables, meaning that some patients' scores varied more across raters, groups, and sessions than did other patients' scores. The presence of these interaction effects served to lower generalizability coefficients for several of the variables, including feedback and self-disclosure ratings. The high interjudge reliability but low generalizability coefficient (for two raters and two sessions) for these scales (e.g., group cohesion scales) indicates that judges can agree on their ratings within a given group session, but such scores then vary considerably across groups or sessions, especially for some patients.

For the variables in the current study that had inadequate generalizability coefficients with two sessions and two raters, analyses of various potential study designs revealed that

minimally adequate levels for generalizability coefficients could be achieved by increasing the numbers of sessions and/or raters. For example, with four raters and eight sessions, feedback ratings of the number of non-positive learning statements received from other patients can be made dependably to differentiate patients. We can envision, however, two practical issues with such designs. The first is that the amount of resources to employ a relatively large number of raters, each rating large number of sessions, becomes a burden that effectively would discourage the pursuit of such a study. Second, only a portion of patients typically stay in psychotherapy long enough to generate enough sessions for such a design. In the NIDA CCTS, 47.4% of patients had fewer than eight GDC sessions. Thus, for part of a sample there would be less than the needed number of sessions to differentiate patients if eight sessions are needed.

A further methodological problem with rating large numbers of sessions in order to create a single patient variable is that patients improve over the course of treatment. Thus, process ratings obtained from later-in-treatment sessions might often be influenced by the improvements made by patients. This issue renders using such an average score over the course of many therapy sessions problematic for predicting treatment outcome—a common purpose of process rating studies.

To the extent that the problematic generalizability coefficients (for two raters and two sessions) found here for some variables is typical of rating of group therapy in general, there are important implications for the use of such scales in understanding the relation of process to outcome. As with low interjudge reliability, a low patient-level generalizability coefficient will attenuate the relation of a process variable to patient outcome if the process variable is scored using the same number of sessions and groups as is used to calculate the low generalizability coefficient. The relation of the process variable to outcome will therefore be underestimated at best and possibly not detected at all. For example, based on the current data, ratings of self-disclosure in the here-and-now based on two to four raters and two sessions (generalizability coefficients of .50 to .57) will probably yield scores at the patient level that provide little chance of detecting a relation of this variable to treatment outcome unless a very large sample size is used to detect a small (attenuated) effect.

In the context of planning a process-outcome study, what then can be done with variables that have inadequate generalizability coefficients even when multiple sessions are combined? One solution is for the investigator to go back to the drawing board and revise the rating scales. However, to the extent that generalizability coefficients examining patient variability are compromised by Patient by Session interactions, another solution is to focus on short-term session outcomes rather than long-term (termination) outcomes. If interjudge reliabilities are high, but process variables change from session-to-session for some patients, overall patient-based scores are compromised but session-based process scores can be used to predict session (or weekly) outcomes. Another possibility is to use brief self-report measures of process that can be filled out at each session and provide lower-cost, higher-frequency assessments of process measures.

It is also important to keep in mind that the generalizability coefficients calculated here examine the patient-level dependability of the mean score across levels of factors of interest (e.g., sessions). Many studies of group therapy examine change in process variables over time (e.g., Brossart, Patton, & Wood, 1998; Kipnes, Piper, & Joyce, 2002). Although generalizability coefficients assessing the dependability of an average over sessions were good to excellent for many of the process variables examined in the current study, for some process variables the assumption that the mean reflects a stable “true score” may be highly unlikely. In rolling groups in particular, the addition or subtraction of a key member may have a profound effect on process and therefore the mean across sessions with and without

such key individuals may be highly misleading. In cases where the mean is not an appropriate way to aggregate across sessions, the dependability of other functions (e.g., linear, quadratic trends over sessions) can be examined in generalizability theory models. Because a limited number of sessions were used, the current study did not assess whether there were stable individual differences in particular patterns (e.g., linear, quadratic) of change over sessions. The existence of such dependable patient differences in patterns of change over time would also be a prerequisite to using such an individual difference variable as a predictor of treatment outcome. If no dependable individual differences in mean session scores, or patterns of scores over sessions, is evident, then aggregating over sessions would not be indicated.

This study has several limitations. Most notably, it is not known whether the findings generalize beyond the context of manual-based group drug counseling for cocaine dependence. It is possible that greater variability in the process variables would be found if naturalistic groups (not manual-based) were examined, if other forms of group therapy were examined, or if group therapy for other types of patients besides those with cocaine dependence was examined. Greater variability on the process ratings might enhance both interjudge reliability and generalizability coefficients. Another limitation was that this study employed rolling groups. The manner in which process variables develop over time, and the extent of variability between patients, sessions, and groups, may be quite different in closed groups versus rolling groups. Furthermore, the stage of therapy was not addressed in the current study. Some studies have found that group process constructs like certain aspects of cohesion or group climate vary over stage of therapy (Brossart et al., 1998). This variability could be taken into account in a generalizability study by incorporating stage of therapy as another design feature.

Despite these limitations, the current study illustrates the value of using generalizability theory to examine the dependability of measurements and the implications of generalizability coefficients for the planning of process-outcome studies. Moreover, the findings suggest that group drug counseling, a very common form of treatment, can be studied using standard process variables. Process research on group drug counseling can help clarify the mechanism of action of this modality and point to ways to enhance the effectiveness of this and similar treatment approaches.

References

Allen, MJ.; Yen, WM. Introduction to measurement theory. Waveland Press; Prospect Heights, IL: 1979.

Bloch S, Crouch E, Reibstein. Therapeutic factors in group psychotherapy. Archives of General Psychiatry. 1981; 38:519–526. Retrieved from http://archpsyc.ama-assn.org/. [PubMed: 7235852]

Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. Psychotherapy: Theory, Research, and Practice. 1979; 16:252–260. doi:10.1037/h0085885.

Box GEP, Cox DR. An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological). 1964; 26:211–252.

Brennan, RL. Generalizability theory. Springer; New York, NY: 2001.

Brossart DF, Patton MJ, Wood PK. Assessing group process: An illustration using tuckerized growth curves. Group Dynamics: Theory, Research, and Practice. 1998; 2:3–17. doi:10.1037/1089–2699.2.1.3.

Budman, SH.; Demby, A.; Koppenaal, G.; Sabin-Daley, B.; Scherz, B.; Hunter, M.; Feldstein, M. Group Cohesiveness Scale (GCS-I): Rater manual. Harvard Community Health Plan, Mental Health Research Program; Boston, MA: 1982.

Budman SH, Soldz S, Demby A, Feldstein M, Springer T, Davis MS. Cohesion, alliance, and outcome in group psychotherapy. Psychiatry. 1989; 52:339–350. [PubMed: 2772092]

Burlingame, GM.; Fuhriman, A.; Johnson, J. Cohesion in group psychotherapy.. In: Norcross, J., editor. A guide to psychotherapy relationships that work. Oxford University Press; Oxford: 2002. p. 71-88.

Cardinet, J.; Johnson, J.; Pini, G. Quantitative methods series. Routledge/Taylor & Francis; New York, NY: 2010. Applying generalizability theory using EduG..

Claiborn CD, Goodyear RK, Horner PA. Feedback. Psychotherapy: Theory, Research, Practice, Training. 2001; 38:401–405. doi:10.1037/0033-3204.38.4.401.

Coche E, Dies RR, Goettelmann K. Process variables mediating change in intensive group therapy training. International Journal of Group Psychotherapy. 1991; 41:379–397. [PubMed: 1885254]

Coyne R, Silver R. Direct, vicarious, and vicarious-process experiences. Small Group Research. 1980; 11:419–429. doi:10.1177/104649648001100407.

Crits-Christoph P, Siqueland L, Blaine J, Frank A, Luborsky L, Onken LS, Beck A. Psychosocial treatments for cocaine dependence: Results of the NIDA Cocaine Collaborative Study. Archives General Psychiatry. 1999; 56:493–502. doi:10.1001/archpsyc.56.6.493.

Cronbach LJ, Nageswari R, Gleser GC. Theory of generalizability: A liberation of reliability theory. The British Journal of Statistical Psychology. 1963; 16:137–163.

Crouch, EC.; Bloch, S.; Wanlass, J. Therapeutic factors: Interpersonal and intrapersonal mechanisms.. In: Fuhriman, A.; Burlingame, GM., editors. Handbook of group psychotherapy: An empirical and clinical synthesis. Wiley & Sons; New York, NY: 1994. p. 269-318.

Dailey, DC.; Mercer, DE.; Carpenter, G. Counseling for cocaine addiction: the collaborative cocaine treatment study model. 1999. Available at http://archives.drugabuse.gov/TXManuals/DCCA/DCCA1.html

Fenton LR, Cecero JJ, Nich C, Frankforter TL, Carroll KM. Perspective is everything: The predictive validity working alliance instruments. Journal of Psychotherapy Practice & Research. 2001; 10:262–268. Retrieved from http://jppr.psychiatryonline.org/content/vol10/issue4/. [PubMed: 11696653]

Gibbons MB, Crits-Christoph P, Levinson J, Gladis M, Siqueland L, Barber JP, Elkin I. Therapist interventions in the interpersonal and cognitive therapy sessions of the Treatment of Depression Collaborative Research Program. American Journal of Psychotherapy. 2002; 56:3–26. Retrieved from http://www.ajp.org/archives/56-01-2002.html. [PubMed: 11977782]

Hill, CE. An overview of the Hill Counselor and Client Verbal Response Modes Category Systems.. In: Greenberg, LS.; Pinsof, WM., editors. The psychotherapuetic process: A research handbook. Guilford; New York, NY: 1986. p. 131-160.

Hill CE, Helms JE, Tichenor V, Spiegel SB, O'Grady KE, Perry ES. Effects of therapist response modes in brief psychotherapy. Journal of Counseling Psychology. 1988; 35(3):222–233. doi: 10.1037/0022-0167.35.3.222.

Horvath, AO. Empirical validation of Border's pantheoretical model of the alliance: The Working Alliance Inventory perspective.. In: Horvath, AO.; Greenberg, LS., editors. The working alliance: Theory, research, and practice. Wiley; New York, NY: 1994. p. 109-128.

Horvath, AO.; Greenberg, LS. The development of the Working Alliance Inventory.. In: Greenberg, LS.; Pinsof, W., editors. The psychotherapeutic process: A research handbook. Guilford; New York, NY: 1986.

Hoyt WT. Bias in participant ratings of psychotherapy process: An initial generalizability study. Journal of Counseling Psychology. 2002; 49:35–46. doi:10.1037/0022-0167.49.1.35.

Johnson JE, Burlingame GM, Strauss B, Bormann B. The therapeutic relationship in group psychotherapy. Gruppenpsychotherapie und Gruppendynamik. 2008; 44:52–89.

Kipnes DR, Piper WE, Joyce AS. Cohesion and outcome in short-term psychodynamic groups for complicated grief. International Journal of Group Psychotherapy. 2002; 52:483–509. doi:10.1521/ijgp.52.4.483.45525. [PubMed: 12375484]

Kivlighan DM. Feedback in group psychotherapy: Review and implications. Small Group Behavior. 1985; 16:373–385. doi:10.1177/0090552685163007.

Littell, RC.; Milliken, GA.; Stroup, WW.; Wolfinger, RD.; Schabenberger, O. SAS system for mixed models. 2nd ed.. SAS Institute; Cary, NC: 2006.

Lundgren D, Miller D. Identity and behavioral changes in training groups. Human Relations Training News. 1965; 9 Spring, 1965.

MacKenzie, KR. The clinical application of a group climate measure.. In: Dies, RR.; MacKenzie, KR., editors. Advances in group psychotherapy: Integrating research and practice. International Universities Press; New York, NY: 1983. p. 159-170.

McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996; 1:30–46. doi:10.1037/1082-989X.1.1.30.

Metz JAJ, Haccou P, Meelis E. On the Shapiro-Wilk test and Darling's test for exponentiality. Biometrics. 1994; 50:527–530.

Morran DK, Stockton R, Bond L. Delivery of positive and corrective feedback in counseling groups. Journal of Counseling Psychology. 1991; 38:410–414. doi:10.1037/0022-0167.38.4.410.

National Institute on Alcohol Abuse and Alcoholism & National Institute of Drug Abuse. Request for applications for group therapy for individuals in drug abuse and alcoholism treatment (Publication No. RFA-DA-04-008). Department of Health and Human Services; Washington DC: 2003.

Raue PJ, Goldfried MR, Barkham M. The therapeutic alliance in psychodynamic-interpersonal and cognitive-behavioral therapy. Journal of Consulting and Clinical Psychology. 1997; 65:582–587. doi:10.1037/0022-006X.65.4.582. [PubMed: 9256559]

Shavelson, RJ.; Webb, NM. Generalizability theory: A primer. Sage; London: 1991.

Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin. 1979; 86:420–428. doi:10.1037/0033-2909.86.2.420. [PubMed: 18839484]

Stiles WB, Agnew-Davies R, Barkham M, Culverwell A, Goldfried MR, Halstead J, Shapiro DA. Convergent validity of the Agnew Relationship Measure and the Working Alliance Inventory. Psychological Assessment. 2002; 14:209–220. doi:10.1037/1040-3590.14.2.209. [PubMed: 12056083]

Substance Abuse and Mental Health Services Administration, Office of Applied Studies. National Survey of Substance Abuse Treatment Services (N-SSATS): 2009. Data on Substance Abuse Treatment Facilities. HHS Publication No. (SMA); Rockville, MD: 2010. p. 10-4579.DASIS Series: S-54

Swallow WH, Monahan JF. Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. Technometrics. 1984; 28:47–55.

Tang TZ, DeRubeis RJ. Sudden gains and critical sessions in cognitive behavioral therapy for depression. Journal of Consulting and Clinical Psychology. 1999; 67:894–904. doi: 10.1037/0022-006X.67.6.894. [PubMed: 10596511]

Tschuschke V, Dies RR. Intensive analysis of therapeutic factors and outcome in long-term inpatient groups. International Journal of Group Psychotherapy. 1994; 44:185–208. [PubMed: 8005718]

Tschuschke V, MacKenzie KR, Haaser B, Janke G. Self-disclosure, feedback, and outcome in long-term inpatient psychotherapy groups. Journal of Psychotherapy Practice and Research. 1996; 5:35–44. Retrieved from http://jppr.psychiatryonline.org/content/vol5/issue1/.

Verbeke, G.; Molenberghs, G. Linear mixed models for longitudinal data. Springer; New York, NY: 2000. p. 64-74.

Wasserman R, Levy K, Loken E. Generalizability theory in psychotherapy research: The impact of multiple sources of variance on the dependability of psychotherapy process ratings. Psychotherapy Research. 2009; 19:397–408. doi:10.1080/10503300802579156. [PubMed: 19235094]

West, BT.; Welch, KB.; Galecki, AT. Linear mixed models: a practical guide using statistical software. CRC; Boca Raton, FL: 2007. p. 27-29.

Woody SR, Adessky RS. Therapeutic alliance, group cohesion, and homework compliance during cognitive-behavioral group treatment of social phobia. Behavior Therapy. 2002; 33:5–27. doi: 10.1016/S0005-7894(02)80003-X.

Table I

Demographic and Clinical Characteristics of Patient Samples Rated on Process Measures

Characteristic	Alliance, feedback, self-disclosure ratings (*n* = 417)	Participation coding (*n* = 440)	Original full study sample (*n* = 487)
% Non-minority	58.3	58.6	57.9
% Employed	59.9	60.6	60.3
% Living alone	69.1	69.6	69.6
% Crack and injectors	81.7	81.3	81.1
% Male	76.5	77.1	76.8
Age, mean (*SD*)	34.1±6.39	33.9±6.29	33.9±6.30
Years education, mean (*SD*)	13.0±2.06	13.0±2.02	13.0±2.00
Days cocaine past 30, mean (*SD*)	10.5±7.71	10.5±7.74	10.4±7.76
Years cocaine use, mean (*SD*)	6.8±4.69	6.9±4.73	6.9±4.75
Days alcohol past 30, mean (*SD*)	7.4±7.91	7.4±7.91	7.4±7.89

Table II

Mean, SD, Normality Test, and Transformation Applied for Process Variables

	Test of deviationfrom normal distribution					
Variable	**Number of patients**	**Mean**	***SD***	**Shapiro-Wilk statistic**	***p*-value**	**Transformation applied**
Alliance – WAI total score	416	187.34	29.03	0.993	0.034	None
Number of positive learning statements						
Received from counselor	417	.297	.667	.502	.001	Binary
Received from other patients	417	.122	.497	.261	.001	Binary
Number of non-positive learning statements						
Received from counselor	417	1.30	1.93	.694	.001	Binary
Received from other patients	417	.693	1.345	.577	.001	Binary
Number of self-disclosures						
Here-and-now	417	.478	1.046	.511	.001	Binary
Past	417	7.571	5.315	.958	.012	None
Participation variables						
Total time speaking per session	440	1.246	.882	.909	.001	None
Number of turns at talk per session	440	19.819	13.034	.906	.001	None
Group cohesion	Number of counselors					
Focus	10	5.63	1.66	.952	.66	None
Interest/involvement	10	6.08	1.84	.954	.69	None
Trust	10	5.75	1.23	.766	.004	Log
Facilitative behavior	10	5.82	2.19	.948	.61	None
Bonding	10	5.72	1.69	.973	.94	None
Global cohesiveness	10	5.56	1.70	.915	.252	None
Global quality	10	4.90	1.93	.861	.049	None
Affective intensity	10	5.50	2.12	.872	.072	None
Conflict	10	2.21	1.50	.727	.002	Log

Note. n = 417 patients; 10 counselors; 14 groups (rolling membership); 387 specific group sessions (tapes); 3.86 sessions were attended, on average, by each patient; five raters (two per session) for alliance, feedback, and self-disclosure variables. For the participation data, n = 440 patients, 10 counselors, 14 rolling groups, 1030 specific group sessions (tapes), 27 raters, 1.5 sessions attended on average per patient. For the cohesion data, n =10 counselors, 14 groups, two raters, 76 specific group sessions (tapes).

Table III

Variance Components due to Patients, Counselors, Sessions, Specific Groups (tapes), and Raters on Alliance, Feedback, Self-Disclosure, Participation, and Cohesion Ratings

Process Variable	Patient	Counselor	Session	Specific group	Rater	Patient × Session	Patient × Group	Patient × Rater	Group × Rater	Residual
Alliance – WAI total score	**122.4 (13)**	**125.7 (14)**	5.9 (1)	**62.0 (7)**	80.6 (9)	20.3 (2)	**139.8 (15)**	**32.7 (4)**	**179.2 (20)**	**130.9 (15)**
Number of positive learning statements										
Received from counselor	.005 (1)	**.022 (5)**	.0003 (0)	.014 (3)	.022 (5)	.035 (7)	**.067 (14)**	.003 (1)	**.091 (19)**	**.220 (46)**
Received from patients	**.0027 (2)**	**.0022 (2)**	.00001 (0)	**.0039 (4)**	.0109 (10)	.0048 (4)	.0060 (6)	**.0074 (7)**	**.0055 (5)**	**.0654 (60)**
Number of non-positive learning statements										
Received from counselor	**.359 (13)**	**.240 (8)**	0 (0)	**.146 (5)**	.092 (3)	.034 (1)	**.800 (28)**	**.073 (3)**	**.186 (7)**	**.907 (32)**
Received from patients	**.154 (9)**	**.046 (3)**	0 (0)	.018 (1)	.078 (4)	**.302 (17)**	**.323 (18)**	**.057 (3)**	**.094 (5)**	**.678 (39)**
Number of self-disclosures										
Here-and-now	**.0136 (6)**	.0024 (1)	.0003 (0)	.0052 (2)	.0146 (7)	.0142 (7)	**.0220 (10)**	**.0124 (6)**	**.0232 (11)**	**.1042 (49)**
Past	**6.58 (23)**	**2.25 (8)**	.004 (0)	**4.08 (15)**	1.27 (5)	1.53 (5)	**5.38 (19)**	**.44 (2)**	**3.26 (12)**	**3.21 (11)**
Participation										
Percent time speaking	**.184 (19)**	**0.052 (5)**	.011 (1)	**.3840** (40)	-	.017 (2)	**.218 (23)**	-	-	**.085 (9)**
Number of turns at talk	**36.19 (16)**	**29.39 (14)**	0.72 (.3)	**78.53 (37)**	-	1.07 (1)	0 (0)	-	-	**68.06 (32)**
Group cohesion									Session by Rater	
Focus	.567 (19)	-	0	-	.376 (13)	.201 (7)	-	.582 (19)	.113 (4)	**1.151 (38)**
Interest/involvement	.748 (27)	-	0	-	.403 (14)	.231 (8)	-	.069 (2)	.034 (1)	**1.300 (47)**
Trust	.011 (21)	-	.001 (1)	-	0	.005 (10)	-	.008 (15)	0	**.028 (53)**
Facilitative behavior	1.476 (34)	-	.093 (2)	-	.112 (3)	0	-	.307 (7)	0	**2.349 (54)**
Bonding	**.540 (21)**	-	0	-	.152 (6)	.099 (4)	-	0	0	**1.730 (69)**
Global cohesiveness	.782 (29)	-	.012 (1)	-	0	.467 (17)	-	0	0	**1.442 (53)**
Global quality	.977 (27)	-	0	-	.206 (6)	.438 (13)	-	.318 (9)	.062 (2)	**1.569 (44)**
Affective intensity	.762 (32)	-	0	-	0	.251 (11)	-	0	.036 (2)	**1.314 (56)**
Conflict	.062 (16)	-	0	-	.049 (12)	.025 (6)	-	.006 (2)	0	**.254 (64)**

Note. Values in parentheses are the percent of total variance due to each effect in the model. $n = 417$ patients; 10 counselors; 14 groups (with rolling membership); 387 specific group sessions (tapes); 3.86 sessions attended on average per patient; five raters (two per session) for alliance, feedback, and self-disclosure variables. For the participation data, $n = 440$ patients, 10 counselors, 14 groups (with rolling membership); 1030 specific group sessions (tapes), 27 raters, 5.90 sessions attended on average per patient. For the cohesion data, $n = 10$ counselors, 14 rolling groups, two raters, 76 specific group sessions (tapes). Variance components shown in bold are significantly different from zero at $p < .05$; those underlined are significant at $p < .01$.

Table IV

Interjudge Reliability for Process Ratings

Process variable	Perjudge *ICC* (A, 1)	Two judges pooled *ICC* (A, 2)
Alliance – WAI total score	.71	.85
Number of positive learning statements		
Received from counselor	.30	.65
Received from other patients	.18	.59
Number of non-positive learning statements		
Received from counselor	.56	.78
Received from other patients	.48	.74
Number of self-disclosures		
Here-and-now	.27	.64
Past	.71	.85
Participation		
Percent time speaking	.92	.96
Number of turns at talk	.88	.94
Group cohesion		
Focus	.64	.82
Interest/Involvement	.82	.91
Trust	.85	.93
Facilitative behavior	.90	.95
Bonding	.94	.97
Global cohesiveness	1.00	1.00
Global quality	.84	.92
Affective intensity	.98	.99
Conflict	.86	.93

Note. n = 417 patients; 10 counselors; 14 groups (with rolling membership); 387 specific group sessions (tapes); 3.86 session per patient; five raters (two per session) for alliance, feedback, and self-disclosure variables. For the participation data, n =119 patients, 10 counselors, 57 specific groups (tapes), 27 raters, 1.5 sessions per patient. For the cohesion data, n =10 counselors, two raters, 76 specific groups (tapes). *ICC* = intraclass correlation coefficient calculated from generalizability model, using the McGraw and Wong (1996) models in which *ICC* (A, k) is the degree of absolute agreement for measurements that are the average of k independent measurements (judges).

Table V

Generalizability Coefficients for Potential Study Designs

	Number of sessions							
	2		4		8		12	
Number of raters:	**2**	**4**	**4**	**8**	**4**	**8**	**4**	**8**
Variable								
Alliance – WAI total score	.82	.87	.90	.93	.92	.95	.92	.95
Number of positive learning statements								
Received from counselor	.20	.21	.34	.35	.48	.50	.55	.58
Received from other patients	.30	.38	.46	.55	.51	.63	.53	.66
Number of non-positive learning Statements								
Received from counselor	.86	.90	.92	.95	.93	.96	.94	.96
Received from other patients	.46	.48	.63	.65	.74	.77	.79	.82
Number of self-disclosures								
Here-and-now	.50	.57	.68	.72	.73	.80	.76	.83
Past	.87	.88	.93	.94	.95	.96	.96	.97
Participation								
Percent time speaking	.95	.95	.97	.98	.99	.99	.99	.99
Number of turns at talk	.98	.98	.99	.99	.99	.99	.99	.99
Group cohesion								
Focus	.50	.66	.72	.84	.76	.86	.77	.87
Interest/Involvement	.68	.81	.88	.94	.93	.96	.94	.97
Trust	.51	.67	.75	.86	.80	.89	.82	.90
Facilitative behavior	.67	.80	.87	.93	.91	.95	.92	.96
Bonding	.56	.71	.83	.91	.91	.95	.94	.97
Global cohesiveness	.68	.81	.90	.95	.95	.97	.96	.98
Global quality	.64	.78	.85	.92	.88	.94	.90	.95
Affective intensity	.70	.82	.90	.95	.95	.97	.97	.98
Conflict	.48	.65	.78	.88	.87	.93	.90	.95