NIH Public Access
Author Manuscript
J Consult Clin Psychol. Author manuscript; available in PMC 2012 June 1.

Published in final edited form as:
J Consult Clin Psychol. 2011 June ; 79(3): 267–278. doi:10.1037/a0023668.

The Dependability of Alliance Assessments: The Alliance–Outcome Correlation is Larger than You Might Think

Paul Crits-Christoph,
Department of Psychiatry, University of Pennsylvania

Mary Beth Connolly Gibbons,
Department of Psychiatry, University of Pennsylvania

Jessica Hamilton,
Department of Psychiatry, University of Pennsylvania

Sarah Ring-Kurtz, and
Department of Psychiatry, University of Pennsylvania

Robert Gallop
Department of Mathematics, West Chester University

Abstract

Objective—To examine the dependability of alliance scores at the patient and therapist level, to evaluate the potential causal direction of session-to-session changes in alliance and depressive symptoms, and to investigate the impact of aggregating the alliance over progressively more sessions on the size of the alliance-outcome relationship.

Method—We used data from a study (*N*=45 patients; *N*=9 therapists) of psychotherapy for major depressive disorder in which the alliance was measured at every treatment session to calculate generalizability coefficients and to predict change in depressive symptoms from alliance scores. Two replication samples were also used.

Results—At the therapist level, a large number of patients (about 60) per therapist is needed to provide a dependable therapist-level alliance score. At the patient level, generalizability coefficients revealed that a single assessment of the alliance is only marginally acceptable. Very good (> .90) dependability at the patient level is only achieved through aggregating four or more assessments of the alliance. Session-to-session change in the alliance predicted subsequent session-to-session changes in symptoms. Evidence for reverse causation was found in later-in-treatment sessions, suggesting that only aggregates of early treatment alliance scores should be used to predict outcome. Session 3 alliance scores explained 4.7% of outcome variance but the average of sessions 3 to 9 explaining 14.7% of outcome variance.

Conclusion—Adequately assessing the alliance using multiple patients per therapist and at least 4 treatment sessions is crucial to fully understanding the size of the alliance-outcome relationship.

Correspondence concerning this article should be addressed to Paul Crits-Christoph, Department of Psychiatry, University of Pennsylvania, Room 650, 3535 Market Street, Philadelphia, PA. 19104. crits@mail.med.upenn.edu.

Publisher's Disclaimer: The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at www.apa.org/pubs/journals/ccp

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Keywords

alliance; outcome; generalizability theory; therapist

The therapeutic alliance, described by Bordin (1979) as composed of the emotional bond between patient and therapist, agreement on tasks, and agreement on goals, is a widely-investigated aspect of psychotherapy. Despite the extent of research on this facet of therapy, divergent views exist on the relative importance of the alliance in relation to treatment outcome. One view of the role of the alliance has tended to minimize its relative contribution to treatment outcome. This view has been supported by correlational studies which found that technical factors predicted outcome more strongly than the alliance, or that the alliance did not predict outcome once other variables have been controlled (DeRubeis, Brotman, & Gibbons, 2005; DeRubeis & Feeley, 1990; Gaston, 1990; Feeley, DeRubeis, & Gelfand, 1999; Kokotovic & Tracey, 1990). Although meta-analytic reviews of studies of the alliance in relation to outcome have concluded that there is substantial evidence for a link between the alliance and treatment outcome (Horvath & Symonds, 1991; Martin, Garske, & Davis, 2000), the results of these reviews actually can be viewed as consistent with a relatively minor role for the alliance. Martin et al. (2000), reviewing 79 studies, and Horvath and Bedi (2002), reviewing 90 studies, obtained an average correlation between the alliance and outcome of .22 and .21, respectively, indicating that less than 5% of the variance in outcome is due to the alliance. Commenting on these results, Crits-Christoph, Connolly Gibbons, and Hearon (2006) state that, “We would hope that the other 95.16% of the outcome variance that is not related to the alliance is also considered when attempting to understand the central features of the change process in psychotherapy” (p. 281).

A second view of the role of the alliance is that it is a key aspect of treatment. Bordin (1979) viewed the effectiveness of a therapy as partly, if not entirely, related to the strength of the working alliance. Several authors have labeled the alliance a “strong” or “robust” predictor of outcome (e.g., Barber, Connolly, Crits-Christoph, Gladis, & Siqueland, 2000; Carroll, Nich, & Rounsaville., 1997; Castonguay, Constantino, & Holtforth, 2006; Connors, Carroll, DiClemente, Longabaugh, & Donovan, 1997; de Roten et al., 2004; Gaston, 1991; Klein et al., 2003; Krupnick et al., 1994; Marmar, Weiss, & Gaston, 1989; Piper, Boroto, Joyce, McCallum, & Azim, 1995), despite the less than 5% of outcome variance explained by the alliance in meta-analytic reviews. An American Psychological Association Division 29 Task Force formulated an entire edited volume around the concept of “psychotherapy relationships that work,” with the primary focus on the alliance, concluding that “the therapy relationship…makes substantial and consistent contributions to psychotherapy outcome independent of the specific type of treatment” (Steering Committee, 2002, p. 441). Based on this substantial contribution of the therapy relationship, the Task Force made recommendations that practitioners be encouraged to make the creation and cultivation of a therapy relationship a primary aim in the treatment of patients and that training programs in psychotherapy be encouraged to provide explicit and competency-based training in the effective elements of the therapy relationship (of which the alliance is the major focus) (Steering Committee, 2002). Practitioners of psychotherapy have also long held that the alliance is a central feature of psychotherapy, dating back to Freud’s (1912/1966) description of the therapy relationship as having two parts, one that interferes with cooperation and one that induces cooperation, with the aspects of the relationship that induce cooperation being “the vehicle of success” (p. 105) in therapy.

Is the alliance relatively weakly associated with outcome, explaining less than 5% of the outcome variance, or a strong, robust finding that is central to psychotherapy and should be a primary focus of treatment and training? While it is certainly possible that these polarized

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

positions on the value of the alliance reflect pre-existing biases by each camp, with the available research or clinical evidence molded to fit such biases, another possibility is that the "data" that generates the clinical view of the overriding importance of the alliance is different from the "data" that ends up in meta-analyses. Clinicians, for example, might be attuned to micro level changes in the therapeutic alliance that alliance questionnaires may not assess. In addition, when clinicians think back on a clinical case and recall their impression of the alliance for patients that had favorable or unfavorable outcomes, they likely think about and integrate information across the entire therapy, recognizing that the alliance might have fluctuated over time but was generally positive (for good outcome cases) or generally negative (for poor outcome cases), or became relatively positive after initially being negative or fluctuating. However, therapists might also selectively recall positive aspects of their alliance, discounting ruptures and negatives aspects of the alliance, when treatment was successful (and vice versa for treatment failures). In contrast to a clinicians' integration of information across therapy, most research studies typically only use a limited sample of therapy sessions. If the alliance varies to a certain degree from session to session, sampling a single session (or even two) may provide a relatively unreliable assessment of the general status of the alliance across the entire treatment and therefore may yield results inconsistent with a clinical view of a particular therapy patient that takes into account information across all sessions.

To evaluate this issue among studies that have reported correlations between the alliance and outcome relationship, we examined all of the 90 studies contained in the Horvath and Bedi (2002) review of alliance studies. For 84 of these studies (relevant data could not be extracted from 6 of the studies), we found that most (n=65; 77.4%) either assessed the alliance at a single session or did not calculate an average alliance over sessions if more than one session was assessed. Of the 19 studies that did calculate an average alliance over sessions, only two studies used a patient-report measure of the alliance based on treatment sessions and averaged the alliance over more than 3 sessions: 6 sessions in Marziali (1984); 20 sessions in Ogrodniczuk, Piper, Joyce, & McCallum (2000). The Marziali (1984) study reported relatively strong correlations of patient report of positive alliance indicators with outcome (squared r's of .32, .18, and .14 with patient, therapist, and clinical evaluator outcome ratings). The Ogrodniczuk et al. (2000) study also reported some higher correlations (e.g., squared r of .14 with change in general symptoms in supportive therapy) than the average correlation (squared r = .04) between the alliance and outcome found by Horvath and Bedi (2002).

Variability in the alliance across treatment sessions has been evident in a number of studies. Several studies, for example, have found that the alliance tended to increase over treatment sessions (Golden & Robbins, 1990; Joyce & Piper, 1990; Kivlighan & Shaughnessy, 1995; Patton, Kivlighan, & Multon, 1997; Kramer, de Roten, Beretta, Michel, & Despland, 2009; Paivio & Patterson, 1999; Piper et al., 1995; Piper, Ogrodniczuk, Lamarche, Hilscher, & Joyce, 2005; Sauer, Lopez, & Gormley, 2003; Stiles, Agnew-Davies, Hardy, Barkham, & Shapiro, 1998), although other studies have not found linear increase in the alliance over sessions (Eaton, Abeles, & Gutfreund, 1988; Gaston, Piper, Debbane, Bienvenu, & Garant, 1994; Klee, Abeles, & Muller, 1990; Hartley & Strupp, 1983; Hilsenroth, Peters, & Ackerman, 2004; Marmar et al., 1989; Morgan, Luborsky, Crits-Christoph, Curtis, & Solomon, 1982; Sexton, Hembre, & Kvarme, 1996). Kivlighan and Shaughnessy (2000) identified three patterns of patient-rated alliance development—linear, stable, and quadratic growth—in two separate patient samples. Stiles et al. (2004) found four distinct patterns of changes in the alliance over time, including two which closely resembled Kivlighan and Shaughnessy's linear growth and stable alliance patterns and two new patterns: one characterized by a negative slope, a slightly positively accelerated curve and high variability and one with a positive slope, a negatively accelerated curve and low variability. These

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

types of individual differences in patterns of change in the alliance over time work against the likelihood that a sample of a session or two of alliance scores adequately captures the overall alliance for the case as a whole and therefore attenuate the relation of the individual differences in alliance level to outcome.

Most relevant to the issue of the number of sessions needed to obtain a stable estimate of the typical level of alliance are studies that examine the correlation of alliance scores from session to session. Such correlations have varied across studies, with some studies reporting correlations of alliance scores over sessions in the range of .50 to .65 (Paivio & Bahr, 1998; Paivio & Patterson, 1999; Bachelor & Salamé, 2000; Luborsky, Crits-Christoph, Alexander, Margolis, & Cohen, 1983; Hersoug, Høglend, Monsen, & Havik, 2001). However, other studies have reported lower correlations, particularly when non-adjacent sessions were examined. For example, Brossart, Willson, Patton, Kivlighan, and Multon (1998) found relatively small correlations when early treatment sessions were correlated with later treatment sessions (patient-rated measure: r^2= .04, therapist-rated measure: r^2= .12).

To directly answer the question of how many treatment sessions are needed to create a stable individual difference measure of the alliance, generalizability theory (Cronbach, Rajaratnam, & Gleser, 1963; Shavelson & Webb, 1991; Wasserman, Levy, & Loken, 2009) can be used. 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. In the current context, generalizability theory can be used to examine the adequacy, or "dependability," of generalizing from a limited sample of sessions to the universe of sessions from which the sample of sessions was selected. To index this dependability quantitatively, a generalizability coefficient can be calculated that assesses 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 (e.g., sessions). Thus, the generalizability coefficient can tell you whether scores from one session are stable, given any session-to-session variability on the alliance, in measuring individual alliance differences between patients. Furthermore, additional generalizability coefficients can also be calculated that would evaluate the extent to which averaging two (or any number of) sessions creates a more dependable score. In terms of acceptable magnitude, generalizability coefficients can be interpreted like standard reliability coefficients. In fact, for a simple design with one context assessed (e.g., raters), the generalizability coefficient is equal to an intraclass correlation coefficient (Hoyt & Melby, 1999). However, with more complex designs that include more than one context (these contexts, such as sessions, therapists, and patients, are referred to as "facets" in generalizability theory), standard reliability analyses using classical test theory will often overestimate reliability because potential interactions between the facets are neglected in classical test theory.

In addition to the influence of dependability of measurement, there are potential confounds that may produce larger correlations of the alliance with outcome. One issue is the problem of reverse causation: the alliance being a function of earlier symptom change rather than the alliance leading to subsequent symptom change. If the level of alliance is simply a marker for improvement to date, as further improvement occurs or doesn't occur over the course of treatment, higher correlations between the alliance and termination outcome will be evident as alliance assessments get closer to the final termination assessment of outcome and therefore a better marker of those final outcomes. These progressively higher correlations with outcome will then increase the correlation of an average alliance score containing later-in-treatment alliance assessments with outcome compared to the similar correlation using an average alliance score that contains fewer later-in-treatment sessions. However, the alliance-final outcome correlation may also increase over sessions even without reverse causation if

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the alliance is strengthening over time for some patients, leading to good outcomes, and weakening for other patients, leading to relatively poorer outcomes.

The purpose of the current study was three-fold: (1) to conduct a generalizability theory analysis of the alliance, (2) to investigate potential confounds of the alliance-outcome relation by examining the causal direction of associations between session-to-session changes in the alliance and session-to-session changes in symptoms, and (3) to examine the relation of alliance scores aggregated over various numbers of sessions to treatment outcome at termination. The generalizability theory analysis followed the lead of recent multilevel studies of the alliance-outcome relationship (Baldwin, Wampold, & Imel, 2007; Crits-Christoph et al., 2009) by incorporating patient and therapist levels, in addition to variability over sessions, so that generalizability coefficients relevant to variability due to the patient and variability due to the therapist could be calculated. At the patient level, we hypothesized that generalizability coefficients would indicate that larger numbers of sessions than typically used in previous research would be needed to measure the alliance dependably and to produce correlations with outcome more consistent with clinical views of a relatively robust alliance-outcome relationship. At the therapist level, we hypothesized that a greater number of patients per therapist than typically used in previous research would be needed to measure the alliance dependably. We also conducted replications of the patient and therapist level generalizability coefficients using two additional samples.

Method

Design

The primary sample for the analyses presented here was drawn from a previous study that focused on training therapists to improve their alliances. The previous study (Crits-Christoph et al., 2006) involved five psychotherapists, each of whom treated patients with major depressive disorder (MDD) before, during, and after training in alliance-fostering therapy. The five therapists were assigned nine patients each, three patients for each study phase. The study was reviewed and approved by an Institutional Review Board.

Participants

A total of forty-five patients were recruited through newspaper advertisements as well as through the outpatient psychiatric referral system at the (institution). To be eligible, subjects (ages 18–60) had to have a primary diagnosis of major depressive disorder, be available for the 16 sessions of study treatment, and provide written, informed consent. Exclusion criteria included current or past history of schizophrenic disorders, bipolar disorders, or cluster “A” Axis II personality disorders (schizoid, schizotypal, or paranoid). Subjects were also excluded from the study if, in the past 12 months, they met criteria for alcohol or substance dependence, obsessive-compulsive disorder, eating disorder, or borderline personality disorder and any acute, unstable, or severe Axis III medical disorder that might interfere with the safe conduct of the study.

Patients were initially screened by phone to determine possible eligibility. If patients met the eligibility criteria, they were scheduled for a diagnostic evaluation that consisted of a Structured Clinical Interview for DSM-IV for Axis I (SCID-I; First, Spitzer, Gibbon, & Williams, 1994) and Axis II (SCID- II; First, Spitzer, Gibbon, Williams, & Benjamin, 1994).

Five therapists (3 women, 2 men) were trained, and all therapists were relatively inexperienced Ph.D. or Psy.D. psychologists with one to three years post-degree experience to preclude the possibility that experience would result in the therapists having developed their own ways of achieving high alliances and therefore having little variability in the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

alliance. In terms of therapeutic orientation, two of the therapists identified themselves as primarily cognitive-behavioral, two were psychodynamic, and one was primarily trained in family systems therapy.

Treatment

Treatment consisted of 16 weekly, 50-minute individual therapy sessions for patients in all three phases of treatment. In the pre-training phase of the study, each therapist treated three patients using his/her usual approach to psychotherapy. In the training phase of the study, a manual-based alliance-fostering therapy (Crits-Christoph et al., 1998) method was taught. Therapists were supervised utilizing this method while treating an additional three patients. In the post-training study phase, therapists each treated an additional three patients using the alliance-fostering psychotherapy, but without intensive supervision. For the purposes of the current report, all patients were included in the analyses, regardless of training phase. Details of the alliance-fostering treatment approach are given elsewhere (Crits-Christoph et al., 2006).

Assessments

Alliance—The alliance was evaluated at the end of every session (for all three phases) using the California Psychotherapy Alliance Scale - Patient Version (CALPAS; Gaston, 1991). Research assistants collected the CALPAS and patients were informed that therapists would not have access to the forms.

The CALPAS is comprised of 24 items rated on a 7-point Likert scale. This study focused primarily on the total CALPAS score, which has been reported to have an internal consistency of .83 in previous studies (Gaston, 1991). In this study, the internal consistency of the CALPAS varied from .86 to .95 across the 16 session assessments.

Patient Outcomes—The Hamilton Depression Rating Scale (HAM-D; Hamilton, 1960) and the Beck Depression Inventory (BDI-II; Beck, Steer, & Brown, 1996) were used to measure depressive symptoms. The 17-item version of the HAM-D was given to patients using the Structured Interview Guide to enhance the reliability of the measure (SIGH-D; Williams, 1988). Total scores on the 17-item HAM-D ≥ 23 are considered “very severe” depression, 19–22 “severe” depression, 14–18 “moderate” depression, and 8–13 “mild” depression. The HAM-D is widely used in treatment studies of depression and has good reliability and adequate convergent and discriminant validity, though it also has weaknesses (e.g., multidimensionality) (Bagby, Ryder, Schuller, & Marshall, 2004). This scale was administered by trained clinical evaluators at pretreatment and post-treatment (week 16). In the current study, internal consistency reliability (Cronbach’s alpha) of the HAM-D total score among all patients who were assessed at baseline was .70; internal consistency for patients assessed at week 16 in the current study was .81.

The BDI-II is a 21-item self-report measure that assesses common symptoms of depression, focusing on cognitions, on a 4-point rating scale. Total scores range from 0 to 63, with scores above 28 considered “severe,” 20 to 28 “moderate,” and 14 to 19 “mild” levels of depressive symptoms. Patients completed the BDI-II at baseline and at the beginning of every treatment session. The median internal consistency (Cronbach’s alpha) of the BDI-II across the 17 BDI administrations (baseline and one at each session) was .93 (range = .76 to .96).

Replication Samples

Two additional samples were used to see whether similar results for estimates for generalizability coefficients at the patient and therapist level would be obtained. Both of

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

these studies also used the patient version of the CALPAS administered at several treatment sessions. The first sample was a Center-wide pooled-study database that combined data from eight pilot studies conducted from 1990 to 1995 with an additional four studies conducted from 1996 to 2002 at the University of Pennsylvania to examine the efficacy of cognitive and psychodynamic therapies for various disorders, including generalized anxiety disorder, panic disorder, chronic depression, borderline personality disorder, avoidant personality disorder, and obsessive compulsive personality disorder. Details of these studies are presented in Crits-Christoph et al. (2001) and Connolly Gibbons et al. (2009). All studies administered the CALPAS at sessions 2, 5, 10, and 15. From these studies, only therapists who treated at least three patients were included in the final sample, and only patients with at least one CALPAS assessment were selected. This resulted in a sample of 236 patients and 30 therapists in the combined study database.

The second replication sample was data from the NIDA Cocaine Collaborative Study (NCCS; Crits-Christoph et al., 1999). This study involved random assignment of cocaine dependent individuals to four treatment groups: cognitive therapy plus group drug counseling, supportive-expressive therapy plus group drug counseling, individual drug counseling plus group drug counseling, and group drug counseling alone. Of the full randomized sample of patients assigned to an individual therapist (*N*=364), we selected only those patients with at least one CALPAS assessment (*N*=300 patients; *N*=37 therapists; number of patients per therapist ranged from 4 to 15). The CALPAS was administered at sessions 2, 5, and 10.

Statistical Analyses

Calculations of generalizability coefficients were based on a model that included random effect terms for session, patient, and therapist. The design specified that patients were nested within therapists and crossed with sessions. Using CALPAS total scores from all available treatment sessions for each patient, variance components were estimated from a mixed effects model using SAS Proc Mixed (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). Using equations for generalizability coefficients (GC) given in Webb, Shavelson, and Haertel (2000) (equations 46 and 49 on pages 28 and 29, respectively), we calculated the patient-level generalizability coefficients as follows:

$$GC_p = \frac{\sigma^2_{p:t} + \sigma^2_t}{\sigma^2_t + \sigma^2_{p:t} + \frac{\sigma^2_{ts}}{n_s} + \frac{\sigma^2_{ps:t,e}}{n_s}}$$

The therapist generalizability coefficient was calculated as follows:

$$GC_t = \frac{\sigma^2_t}{\sigma^2_t + \frac{\sigma^2_{p:t}}{n_p} + \frac{\sigma^2_{ts}}{n_s} + \frac{\sigma^2_{ps:t,e}}{n_p n_s}}$$

where p = patients, t = therapists, s = sessions, and e = residual error.

Using these formulas, estimates of generalizability coefficients for averaging across various numbers of sessions ranging from one session to the average of 14 were calculated by altering the number of sessions in the denominator of the formulas. The number of treatment sessions was varied from 1 to 14 (a maximum of 14 sessions was used because the regression analyses predicting outcome described below excluded sessions 1 and 2). When this number was set to “1,” the generalizability coefficient for an alliance score based on a

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

single session is calculated; when this number is set to "2," the generalizability coefficient for an alliance score based on the average of two sessions is calculated; and so on up to the average alliance score across 14 sessions. The impact of varying the number of patients per therapist on therapist generalizability coefficients was also examined by changing the N in the terms in the denominator of the equation.

Mixed model longitudinal analyses were conducted to investigate the possible confounding effect of reverse causation in which prior improvement in symptoms influences the alliance. These causal direction questions are best addressed by examining change in one measure (e.g., depressive symptoms) in relation to change in the other measure (e.g., alliance) incorporating a time lag so that temporal precedence is established (e.g., does prior symptom change lead to subsequent change in alliance?). To examine this reverse causation, change in alliance scores from session X-1 to session X was the dependent variable and prior change in the BDI from session X-1 to session X was the time-varying predictor variable, using all data from sessions 3 to 16. Because the BDI was administered at the beginning of the session and the CALPAS at the end of the session, change in the BDI likely occurred prior to change in the alliance in this analysis. We also examined change in the alliance (from session X-1 to X) as the time-varying predictor in relation to change in BDI scores from session X to X+1 as the dependent variable. In all of these analyses time was a covariate, allowing the amount of change in alliance and BDI scores to vary from session to session.

Regression analyses were used to predict treatment outcome from the alliance. Initially, a single level model was conducted so that results could be compared to the type of analysis that is typically used in the literature. In this single level model, the HAM-D total score at week 16 was used as the dependent variable, and baseline HAM-D total score was a covariate. In order to control for the influence of early symptomatic improvement on the alliance (which can cause a spurious relationship between the alliance and final outcome), we also covaried early symptomatic change (BDI scores at baseline and week 3 entered analyses as covariates). By Week 3, 50% of the change from baseline to week 16 had occurred in average BDI scores (baseline mean = 35.2; session 3 mean = 25.7; week 16 mean = 16.2). In addition, we made the decision to examine the alliance beginning at session 3 so that there was an increased likelihood that patient ratings of the alliance were influenced more by the evolving therapeutic interactions (as relevant to agreement on goals, agreement on tasks, and therapeutic bond) rather than predisposing patient factors and reporting biases (tendencies to be positive or negative in self-report) that might be more evident in session 1 and session 2 ratings. We report squared partial correlations from the regression analyses.

Following Baldwin et al. (2007), we also implemented a multilevel modeling approach adjusting for the hierarchy of clustering with nested random effects(Goldstein, 1987; Bryk & Raudenbush, 1992) to predict outcome from the alliance. These hierarchical linear models (HLM) were implemented using SAS Procedure Proc Mixed (Littell et al., 2006). In this model, patient was specified at Level 1 and therapist at Level 2. The HAM-D total score at post-treatment was the dependent variable; baseline HAM-D, baseline BDI, session 3 BDI, and alliance were predictor variables. These multilevel models are well suited to address the aims of our analysis for four reasons: (a) they take into account the hierarchical nature of this data structure with patients nested within therapist, (b) they are able to consider therapists as a random factor, and this corresponds to an additional source of variability, (c) they allow between- and within-therapist correlations to be modeled simultaneously, and (d) they use estimation procedures that are robust for unequal sample sizes within therapists (Baldwin, et al. 2007). Separating the patient-to-patient differences and therapist-to-therapist differences in the alliance scores was accomplished through centering. In general, predictors can be centered at the mean within-group (i.e., group mean centering) or centered at the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

overall mean (i.e. grand mean centering). The difference between patients within the same therapist was quantified by patient differences from their respective therapist's mean (i.e., group mean). The difference between therapists was quantified by therapist's differences from the overall mean (i.e. grand mean) of the respective alliance term. Whether differences between patients within therapist or differences between therapists explained the outcome was determined by the statistical significance of the regression coefficients corresponding to the two parts (i.e., within patient and within therapist). The corresponding model was as follows:

$$HAMD_{ij}=\gamma_{00}+\alpha_1 BaselineHAMD_{ij}+\alpha_2 BaselineBDI_{ij}+\alpha_3 Session3BDI_{ij}+\underbrace{\left(\gamma_{10}\left(Alliance_{ij}-TM\ Alliance_{.j}\right)\right)}_{\text{within Therapist}}+\underbrace{\left(\gamma_{11}\left(TM\ Alliance_{.j}-GM\ Alliance_{..}\right)\right)}_{\text{between Therapist}}+\underbrace{U_{0j}}_{\text{Therapist random intercept}}+\underbrace{R_{ij\cdot}}_{Error}$$

Notation is as follows: Terms with the subscript ij corresponds to value for the i^{th} patient within the j^{th} therapist; the subscript •j coupled with the TM prefix corresponds to the mean for the j^{th} therapist (group mean); the subscript •• coupled with the GM prefix corresponds to the grand mean. The therapist random effect, U_{0j}, is distributed normally with mean of 0 with a source of variability due to therapist. The residual effect, R_{ij}, is distributed normally with mean of 0 with a source of variability due to error.

This multilevel modeling allows for examining the extent to which (1) patient differences in alliance scores within a therapist predict outcome, determined by the statistical significance of γ_{10} regression coefficients, and (2) therapist differences in average alliance (averaging across patients within a therapist) are predictive of therapists' average patient outcomes, determined by the statistical significance of γ_{11} regression coefficients. Effect sizes (Cohen's *d*, which was then converted to partial *r's* to compare to the literature), derived from the *F*-test for the compound symmetry design, created by the specification of therapist as a random effect, were calculated as $d=2\sqrt{\frac{F}{df}}$, where *F* is the *F*-test statistic for the regression coefficient of the interaction term(Rosenthal & Rosnow, 1991).

Results

Characteristics of Sample and Treatment Duration

About half (56%) of the sample of 45 patients was women, 47% had completed college or post-graduate training, 51% were employed full-time, 42% were married, and about 25% were members of a minority racial/ethnic group (7% African American; 2% Latino; 2% Native American; 4% Asian; 9% Other). Patients ranged in age from 20 to 59 years, with an average of 42.8 years. Concurrent Axis I diagnoses were evident in 69% of patients, and 58% had a concurrent Axis II diagnosis. The most common additional Axis I diagnoses were dysthymia (20%), social anxiety disorder (16%), posttraumatic stress disorder (13%), and generalized anxiety disorder (13%). The average HAM-D score at baseline was 17.2 (*SD* = 4.4; *N*=45) and the average BDI score was 35.1 (*SD* = 7.7; *N*=45).

The average number of treatment sessions (out of 16) attended was 14.7 (*SD* =2.32; range 6–16; median = 15; mode = 16).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Dependability of Alliance Scores

Patient-level generalizability coefficients for CALPAS scores are given in Figure 1. Plotted are the patient-level generalizability coefficients for alliance scores based on averaging various numbers of sessions. The patient-level generalizability coefficient based on a single session was .77. A generalizability coefficient of .77 indicates that 77% of the variance in overall mean alliance scores is attributable to individual differences in patients' "true scores," where the remaining 23% is attributable to other effects that reduced the precision of patient-level scores. A generalizability coefficient ≥ .80 is generally viewed as acceptable (Cardinet, Johnson, & Pini, 2010), although similar to reliability coefficients, a higher coefficient is better in order to minimize error variance and maximize true score variance. As can be seen in Figure 1, aggregating over an increasing number of sessions results in an increase in generalizability coefficients. Most of the increment in the generalizability coefficient occurs with adding a second, third, and fourth session, with little further increase (beyond .96) after seven sessions have been aggregated.

Figure 1 also presents therapist-level generalizability coefficients based on various numbers of sessions. The number of patients per therapist (9) is held constant for these calculations. The therapist-level generalizability coefficient is low (.34) when the alliance is assessed at only one session, indicating that nine patients per therapist and one assessment of the alliance is not sufficient to create a mean therapist alliance score that dependably differentiates therapists in terms of their typical levels of average alliances (across patients in their caseload). However, as seen in Figure 1, increasing the number of sessions assessed for the alliance has little impact on the therapist-level generalizability coefficient. With nine patients per therapist, increasing the number of sessions assessed for each patient from one to 14 raises the therapist generalizability coefficient from .34 to .40.

Table 1 presents further results of the current study in conjunction with generalizability coefficients from the two replication samples. Very similar results were apparent in all three studies. In order to achieve consistently very good (near .90 or above) patient-level generalizability coefficients across the studies, the alliance needed to be assessed at four occasions. When equating for the same number of patients per therapist (set at 9, as in the current study), therapist-level generalizability coefficients were low (ranging from .30 to . 34). In order to achieve an adequate (.80) therapist-level generalizability coefficient, the necessary number of patients per therapist varied from 54 to 70 across the three studies when four occasions were used to assess the alliance. If the alliance is assessed at only a single session, as is done in many studies, the number of patients per therapist needed to achieve a minimally adequate (.80) therapist-level generalizability coefficient would be 70 based on the current study, 78 based on the NIDA CCTS data, and 97 based on the Center-wide pooled database.

Direction of Causal Influence: Session-to-Session Changes in Alliance and Depressive Symptoms

Longitudinal mixed model analyses revealed no significant impact of prior symptom improvement (session X-1 to session X, with symptoms measured before the session) on session-to-session changes in the alliance (session X-1 to session X, with the alliance measured after the session) across sessions 3 to 16 ($t(44)= -.64$, $p=.53$, $r= - .09$). Because the impact of prior symptom change on the alliance may be particularly evident on later-in-treatment scores, we also conducted these analyses by separately examining sessions 3 to 9 and sessions 10 to 16. While the impact of prior symptom change on change in the alliance was non-significant (and a positive correlation) in the earlier (3 to 9) sessions ($t(44) = 1.50$, $p = .13$, $r=.22$), there was evidence for reverse causation in the later (10 to 16) sessions ($t(42) = - 2.47$, $p = .014$, $r=-.36$; negative correlation indicating decreases [improvement] in

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

BDI associated with increases in alliance). Similar analyses were also conducted with change in BDI scores (from session X to X+1) as the dependent variable and change in alliance from the prior session (X-1) to the current session (X) as the time-varying predictor variables. These analyses revealed an impact of change in the alliance on subsequent change in depression across sessions 3 to 16 ($t(44)=-3.37$, $p=.0008$, $r=-.45$), as well as sessions 3 to 9 ($t(44)=-2.56$, $p=.011$, $r=-.36$) and 10 to 16 ($t(42)=-2.53$, $p=.012$, $r=-.36$) taken separately.

Prediction of Outcome from Alliance Averaged over Various Numbers of Sessions

Based on our results showing that later-in-treatment alliance scores were confounded with the influence of prior symptom change, we examined the impact of aggregating alliance scores using only earlier treatment scores. Seven early sessions (3 to 9) were chosen as the maximum number of sessions to aggregate because the generalizability theory analysis suggested that dependability of measurement of the alliance reached an asymptote at about 7 sessions. Figure 2 presents the results of regression models predicting termination (week 16) HAM-D score from alliance, controlling for baseline HAM-D, baseline BDI, and session 3 BDI scores. Plotted is the percent of outcome variance accounted for by the alliance when alliance scores are averaged over increasingly larger number of sessions. Using alliance at session 3 only, the percent variance explained was 4.7% (r = -.22; value is negative because higher alliance scores are associated with lower levels of depressive symptoms at termination). Averaging over increasingly larger numbers of sessions resulted in a linear increase in percent of outcome variance explained, with a score based on the average alliance across sessions 3 to 9 (7 sessions averaged) explaining 14.7% of the outcome variance. With the current sample size (N=45), 9% or greater percent variance explained is statistically significant at .05 (two-tailed), although degrees of freedom are slightly reduced in the multilevel models with covariates.

Predictive results for a multilevel model (patients nested within therapists; both crossed with sessions) are also given in Figure 2. At the patient level, the percent variance explained ranged from 1.8% when session 3 CALPAS was used as the predictor to 10.8% when the average of sessions 3 to 9 was used as the predictor. As expected based on the generalizability coefficient analyses, the therapist-level contribution to the alliance-outcome relationship shows little increment with increasing numbers of sessions averaged. A more dependable therapist-level that contains more “true score” therapist level variance, and therefore increased potential prediction of outcome, can largely only be achieved by increasing the numbers of patients per therapist in a study.

Among the early sessions (3 to 9), alliance at session 3 explained the least (4.7%) variance in change in HAM-D total scores from intake to termination. Outcome variance explained by the alliance varied among the other early sessions (from 8.0% at session 8 to 13.7% at session 7). Because of this variability in the predictive success of the alliance at a given session (whether systematic or random), the pattern of increasing amount of variance explained as shown in Figure 2 will be affected in part by the starting point in such a sequence. To address this, we calculate average alliance scores over various four sessions blocks – four sessions was chosen because that is an adequate number to achieve good (>. 90) dependability. Percent variance in change in HAM-D total scores from intake to termination (covarying BDI change from intake to session 3) was 10.1% for the average of sessions 3 to 6 alliance scores; 14.3% for the average of sessions 4 to 7; and 16.7% for the average of sessions 5 to 8. Although to a certain extent these increasing numbers (from 10.1% to 16.7%) were a function of a tendency for higher correlations of the alliance with outcome as sessions progressed from session 3 to 8, the impact of aggregation was apparent even if sessions are aggregated progressively backwards beginning at session 8: the alliance at session 8 alone explained 7.8% of the variance in outcome, while the average of the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

alliance at sessions 6 to 8 explained 16.7% of the variance in outcome (though further backward aggregating did not increase the percent variance explained).

Discussion

The results of this study may help to sort out the discrepancy concerning two views of the importance of the alliance to psychotherapy outcome. Measures of the alliance that take into account the typical level of the alliance across multiple sessions were substantially better predictors of outcome at termination than was the alliance measured at a single session. The fact that a single session is inadequate to measure the individual patient differences in the alliance was evident in the generalizability coefficients that we calculated. At least two treatment sessions were needed to arrive at an alliance score with a minimally acceptable(.80) generalizability coefficient. However, ideally, research instruments have a higher generalizability coefficient than .80. Very good (.90 or above) patient-level generalizability coefficients were only consistently achieved across the current sample and two replication samples when the alliance was aggregated over four or more occasions. Apparently, the reduction in error variance by going from 23% error (generalizability coefficient of .77) to 10% or 5% error has an appreciable impact on the relation of the alliance to outcome. Relatively few studies in the alliance literature have aggregated the alliance over four or more assessments. This suggests that the 4.8% of outcome variance associated with the alliance reported in the meta-analysis by Martin et al. (2000), and the 4.4% reported by Horvath and Bedi (2002), may underestimate the size of the alliance-outcome relationship compared to an adequate patient-level measurement of the alliance (aggregated over multiple sessions).

Although aggregating more sessions increases the size of the alliance-outcome relation, the use of late-in-treatment sessions can introduce confounds to alliance-outcome relation. Our results indicated that, from session 10 to 16, session-by-session changes in depressive symptom were significantly predictive of subsequent session-to-session changes in the alliance. This reverse causation can lead to the alliance being a marker for outcome. Studies that average the alliance across treatment, including late sessions (e.g., Ogrodniczuk, Piper, Joyce, & McCallum, 2000) might be reporting alliance-outcome correlations that are biased upwards because of this confound. To adequately understand the scope of the impact of prior symptom change on the alliance, further studies using large sample sizes and employing structural equation modeling are needed to tease out the direction of causality and in particular examine the possibility of reciprocal influences between the alliance and outcome. In addition, future research can explore whether or not the clinician's view of the importance of the alliance in relation to outcome is biased by selective recall, is in part an illusion created by the impact of prior symptom change on the alliance, or is an intuitive integration of these reciprocal influences over time. Until such studies are conducted, however, researchers who examine the alliance-outcome correlation should be aware that, while averaging over several sessions yields a more dependable alliance score, averaging over a large number of sessions (particularly later in treatment sessions) may well increase the influence of prior symptomatic improvement (or other third variables) on the alliance-outcome relation.

In the current study, we aggregated alliance over four consecutive early-in-treatment sessions and found that the relation of these average alliance scores to outcome was a larger effect than found when using single session assessments of the alliance. However, it should be noted that the generalizability coefficients simply indicate that more assessments are better, not that averaging, for example, sessions 3, 4, 5, and 6 is optimal. Aggregating any random sample of sessions would likely yield similarly improved generalizability coefficients and better prediction of outcome compared to a single session. From a clinical

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

point of view, however, a primary interest in the alliance is in how it sets the stage for the ongoing work of therapy. Thus, assessment of the alliance early in treatment is most relevant to this clinical view of the role of the alliance. Moreover, as mentioned, late-in-treatment assessments of the alliance are more likely to be influenced by prior symptom improvement.

Several studies (e.g., Barber et al., 2000; Crits-Christoph et al., 2009; Klein et al., 2003), like the current one, have used early symptom improvement as a covariate when examining the alliance-outcome relation. Although it is useful to do this in order to rule out the impact of early symptom change on the alliance, the determinants of such early symptom change may be of particular interest in themselves. Moreover, removing this early symptom change from final symptom change may reduce variability in final outcomes, thereby artificially limiting correlations with outcome.

A particularly important finding of the current study was the relatively low therapist-level generalizability coefficients found in all three samples examined. These findings are especially important because of investigations (Baldwin et al., 2007; Crits-Christoph et al., 2009) demonstrating stronger alliance-outcome relationships at the therapist level compared to the patient-level of analysis, despite the less than ideal therapist-level assessment in those studies. In the Baldwin et al. (2007) study, there were 4.1 patients per therapist; in the Crits-Christoph et al. (2009) study, there were 12.9 patients per therapist. The data of these previous studies, taken together with our current finding of low therapist generalizability coefficients when the ratio of patients per therapist is relatively low (e.g., below 50), implies that stronger relationships between the alliance and outcome would be evident at the therapist level if high numbers of patients per therapist were used in a study. Thus, an ideal study that would uncover an accurate estimate of the effect size for the impact of the alliance on outcome at both the patient and therapist level would include assessments of the alliance at multiple sessions and a large number of patients per therapist.

In designing an ideal study of the alliance in relation to outcome one other factor is relevant: statistical power. While incorporating a large number of patients per therapist is likely to guarantee that there is adequate power for examining effects at the patient level, to achieve statistical significance at the therapist level a study also would need to include a relatively large number of therapists. The relatively large number (80) of therapists in the Baldwin et al. (2007) study likely provided enough statistical power to detect a therapist-level effect despite the fact that the low number (4.1) of patients per therapists in that study provided a ceiling on the potential size of the alliance-outcome effect at the therapist level (i.e., low therapist-level generalizability coefficient). As unrealistic as it may sound, the results presented here along with statistical power considerations leads to the conclusion that a study designed to accurately estimate the size of the alliance – outcome relationship at both the patient and therapist level would include perhaps 50 therapists, each treating 60 patients, for a total sample size of 3000 patients, and assessment of the alliance at four or more occasions.

Beyond the practical considerations of such a study and the lack of previous knowledge about limited therapist-level generalizability coefficients, one reason why most investigators do not think in these sample size terms is that the goal of most research is simply to detect an effect (whether or not the detected effect is attenuated from the maximum possible effect), not to determine the accurate size of an effect. In fact, the majority of published studies of the alliance did indeed detect a statistically significant relation of the alliance to outcome. But a full understanding of the role of the alliance in psychotherapy, at both the scientific and clinical level, would answer the question of how large is the alliance-outcome relationship and whether the patient or therapist level (or both) are contributing. To answer

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

this question, issues related to adequacy of assessment (at relevant levels) and study design need to be taken into account.

We can speculate on the implications of these results for clinical practice and the training of psychotherapists. For ongoing monitoring of the alliance in clinical practice, it would obviously be useful to measure the alliance repeatedly. To the extent that this presents a burden on patients, particularly if an outcome instrument is already administered at treatment sessions, very brief measures of the alliance might have greater clinical utility, assuming the reliability and validity of such measures are adequate. In a training context where beginning therapists are being evaluated on their ability to form positive alliances with patients, the current results indicate that adequate evaluation of a therapist requires the assessment of the alliance across a relatively large number of patients. Whether extreme outliers (i.e., trainee therapists who consistently form very poor alliances) can be detected using a relatively small number of patients is a question that could be addressed in future research studies.

Several limitations of this study are important to note. For one, the size of the alliance-outcome relationship might vary by type of treatment. Within treatments in which the techniques might have a more potent impact (e.g., exposure therapy for certain anxiety disorders), the alliance might have less of an impact on outcome. Second, the alliance-outcome relationship might vary by disorder. Many alliance studies in the literature use samples of patients with a depressive disorder, or a mixed sample in which depression is common. The provision of a caring and empathic therapeutic relationship that facilitates a strong alliance may be especially important within the context of a disorder like depression often characterized by disconnection from other people, loneliness, and low self-esteem. Thus, generalizability of the current findings to various outpatient populations is uncertain. However, the fact that generalizability coefficients in our Center-wide pooled study database that incorporated a variety of treatments and disorders were quite similar at both the patient and therapist level to those generalizability coefficients found in the depression sample, and the cocaine study suggests that the results found here are not highly disorder-specific. The findings here may also not generalize to other alliance instruments. Other self-report alliance scales, and therapist or observer versions of the CALPAS and other scales, may yield different generalizability coefficients. Thus, with such other scales, a greater or lesser number of sessions and patients per therapist might be needed for adequate assessment of the alliance at the patient and therapist levels, respectively. Finally, as noted, prior symptomatic improvement and other potential third-variables need to continue to be addressed in future correlational studies of the alliance in relation to outcome.

Another concern that might be expressed about the results presented here is that our focus on the less than optimal dependability of measurement is a problem that applies to much of research. The effect size for almost any investigation reported in the scientific literature could be made larger if the reliability, or dependability, of the measures used was improved. However, as mentioned earlier, this concern reduces to the issue of whether an investigation is attempting to uncover whether an effect exists or attempting to accurately determine the size of an effect. We therefore conclude that the current results and the associated design implications should be considered during the debate of how important the alliance is to psychotherapy outcome. Beyond the alliance, other areas of research might benefit from the examination of generalizability coefficients relevant to different design facets and the implications of these coefficients for a full understanding of a phenomenon.

Acknowledgments

The preparation of this manuscript was funded in part by National Institute of Mental Health grant P50-MH-45178.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

References

Bachelor, A.; Salamé, R. Participants' perceptions of dimensions of the therapeutic alliance over the course of therapy; Journal of Psychotherapy Practice and Research. 2000. p. 39-53.Retrieved from http://jppr.psychiatryonline.org/cgi/content/full/9/1/39

Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? American Journal of Psychiatry. 2004; 161:2163–2177. doi: 10.1176/appi.ajp.161.12.2163. [PubMed: 15569884]

Baldwin SA, Wampold BE, Imel ZE. Untangling the alliance-outcome correlation: Exploring the relative importance of therapist and patient variability in the alliance. Journal of Consulting and Clinical Psychology. 2007; 75:842–852. doi: 10.1037/0022-006X.75.6.842. [PubMed: 18085902]

Barber JP, Connolly MB, Crits-Christoph P, Gladis M, Siqueland L. Alliance predicts patients' outcome beyond in-treatment change in symptoms. Journal of Consulting and Clinical Psychology. 2000; 68:1027–1032. doi: 10.1037/0022-006X.68.6.1027. [PubMed: 11142536]

Beck, AT.; Steer, RA.; Brown, GK. Manual for Beck Depression Inventory II (BDI-II). San Antonio, TX: Psychology Corporation; 1996.

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

Brossart DF, Willson VL, Patton MJ, Kivlighan DM, Multon KD. A time series model of the working alliance: A key process in short-term psychoanalytic counseling. Psychotherapy. 1998; 35:197–205. doi: 10.1037/h0087645.

Bryk, A.; Raudenbush, SW. Hierarchical Linear Modeling: applications and data analysis methods. Thousand Oaks, CA: Sage; 1992.

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

Carroll KM, Nich C, Rounsaville BJ. Contribution of the therapeutic alliance to outcome in active versus control psychotherapies. Journal of Consulting and Clinical Psychology. 1997; 65:510–514. doi: 10.1037/0022-006X.65.3.510. [PubMed: 9170775]

Castonguay L, Constantino M, Holtforth MG. The working alliance: Where are we and where should we go? Psychotherapy: Theory, Research, Practice, Training. 2006; 43:271–279. doi: 10.1037/0033-3204.43.3.271.

Connolly Gibbons MB, Crits-Christoph P, Barber JP, Stirman SW, Gallop R, Goldstein L, Ring Kurtz S. Unique and common mechanisms of change across cognitive and dynamic psychotherapies. Journal of Consulting and Clinical Psychology. 2009; 77:801–813. doi: 10.1037/a0016596. [PubMed: 19803561]

Connors GJ, Carroll KM, DiClemente CC, Longabaugh R, Donovan DM. The therapeutic alliance and its relationship to alcoholism treatment participation and outcome. Journal of Consulting and Clinical Psychology. 1997; 65:588–598. doi: 10.1037/0022-006X.65.4.588. [PubMed: 9256560]

Crits-Christoph P, Connolly Gibbons MB, Crits-Christoph K, Narducci J, Schamberger M, Gallop R. Can therapists be trained to improve their alliances? A pilot study of Alliance-Fostering Therapy. Psychotherapy Research. 2006; 13:268–281. doi: 10.1080/10503300500268557.

Crits-Christoph, P.; Connolly, MB.; Gallop, R.; Barber, JP.; Tu, X.; Gladis, M.; Siqueland, L. Early improvement during manual-guided cognitive and dynamic psychotherapies predicts 16-week remission status; Journal of Psychotherapy Research and Practice. 2001. p. 145-154.Retrieved from http://jppr.psychiatryonline.org/cgi/content/full/10/3/145

Crits-Christoph P, Connolly Gibbons MB, Hearon B. Does the alliance cause good outcome? Recommendations for future research on the alliance. Psychotherapy: Theory, Research, Practice, Training. 2006; 43:280–285. doi: 10.1037/0033-3204.43.3.280.

Crits-Christoph, P.; Crits-Christoph, K. Alliance-fostering therapy for major depressive disorder. University of Pennsylvania; 1998. Unpublished manuscript

Crits-Christoph P, Gallop R, Temes CM, Woody G, Ball SA, Martino S, Carroll KM. The alliance in motivational enhancement therapy and counseling as usual for substance use problems. Journal of Consulting and Clinical Psychology. 2009; 77:1125–1135. doi: 10.1037/a0017045. [PubMed: 19968388]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Crits-Christoph P, Siqueland L, Blaine J, Frank A, Luborsky L, Onken LS, Muenz LR, … Beck AT. Psychosocial treatments for cocaine dependence: National Institute on Drug Abuse Collaborative Cocaine Treatment Study. Archives of General Psychiatry. 1999; 56:493–502. doi: 10.1001/archpsyc.56.6.493. [PubMed: 10359461]

Cronbach L, Rajaratnam N, Gleser GC. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology. 1963; 16:137–163.

de Roten Y, Fischer M, Drapeau M, Beretta V, Kramer U, Favre N, Despland J-N. Is one assessment enough? Patterns of helping alliance development and outcome. Clinical Psychology and Psychotherapy. 2004; 11:324–331. doi: 10.1002/cpp.420.

DeRubeis RJ, Feeley M. Determinants of change in cognitive therapy for depression. Cognitive Therapy Research. 1990; 14:469–482. doi: 10.1007/BF01172968.

DeRubeis RJ, Brotman MA, Gibbons CJ. A conceptual and methodological analysis of the nonspecifics argument. Clinical Psychology: Science & Practice. 2005; 12:174–183. doi: 10.1093/clipsy/bpi022.

Eaton TT, Abeles N, Gutfreund MJ. Therapeutic alliance and outcome: Impact of treatment length and pretreatment symptomatology. Psychotherapy: Theory, Research, Practice, Training. 1988; 25:536–542. doi: 10.1037/h0085379.

Feeley M, DeRubeis R, Gelfand L. The temporal relation of adherence and alliance to symptom change in cognitive therapy for depression. Journal of Consulting and Clinical Psychology. 1999; 67:578–582. doi: 10.1037/0022-006X.67.4.578. [PubMed: 10450629]

First, MB.; Spitzer, RL.; Gibbon, M.; Williams, JBW. Structured clinical interview for Axis I DSM-IV Disorders--Patient Edition. Washington, DC: American Psychiatric Press; 1994.

First, MB.; Spitzer, RL.; Gibbon, M.; Williams, JBW.; Benjamin, L. Structured Clinical Interview for DSM-IV Axis-II Personality Disorders (SCID-II, version 2.0). Biometrics Research Dept., New York State Psychiatric Institute; 1994.

Freud, S. The dynamics of transference. In: Strachey, J., editor. The standard edition of the complete psychological works of Sigmund Freud. Vol. Vol. 12. London, England: Hogarth Press; 1912/1966. p. 97-108.

Gaston L. The concept of the alliance and its role in psychotherapy: Theoretical and empirical considerations. Psychotherapy: Theory, Research, Practice, Training. 1990; 27:143–153. doi: 10.1037/0033-3204.27.2.143.

Gaston L. Reliability and criterion-related validity of the California Psychotherapy Alliance Scales--patient version. Psychological Assessment. 1991; 3:68–74. doi: 10.1037/1040-3590.3.1.68.

Gaston L, Piper WE, Debbane EG, Bienvenu J, Garant J. Alliance and technique for predicting outcome in short- and long-term analytic psychotherapy. Psychotherapy Research. 1994; 4:121–135. doi: 10.1080/10503309412331333952.

Golden BR, Robbins SB. The working alliance within time-limited therapy: A case analysis. Professional Psychology: Research and Practice. 1990; 21:476–481. doi: 10.1037/0735-7028.21.6.476.

Goldstein, H. Multilevel models in educational and social research. London, England: Griffin; 1987.

Hamilton MA. A rating scale for depression. Journal of Neurological and Neurosurgical Psychiatry. 1960; 23:56–62. doi: 10.1136/jnnp.23.1.56.

Hartley, D.; Strupp, H. The therapeutic alliance: Its relationship to outcome in brief psychotherapy. In: Masling, J., editor. Empirical studies of psychoanalytic theories. Hillsdale, NJ: Erlbaum; 1983. p. 1-27.

Hersoug, AG.; Høglend, P.; Monsen, JT.; Havik, OE. Quality of working alliance in psychotherapy: Therapist variables and patient/therapist similarity as predictors; Journal of Psychotherapy Research and Practice. 2001. p. 205-216.Retrieved from http://jppr.psychiatryonline.org/cgi/content/full/10/4/205

Hilsenroth MJ, Peters EJ, Ackerman SJ. The development of therapeutic alliance during psychological assessment: patient and therapist perspectives across treatment. Journal of Personality Assessment. 2004; 83:332–344. doi: 10.1207/s15327752jpa8303_14. [PubMed: 15548469]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Horvath AO, Symonds BD. Relation between working alliance and outcome in psychotherapy: A meta-analysis. Journal of Counseling Psychology. 1991; 38:139–149. doi: 10.1037/0022-0167.38.2.139.

Horvath, AO.; Bedi, RP. The alliance. In: Norcross, JC., editor. Psychotherapy relationships that work: Therapist contributions and responsiveness to patients. New York, NY: Oxford University Press; 2002. p. 37-69.

Hoyt WT, Melby JN. Dependability of measurement in counseling psychology: An introduction to generalizability theory. Counseling Psychologist. 1999; 27:325–352. doi: 10.1177/0011000099273003.

Joyce AS, Piper WE. An examination of Mann's model of time-limited individual psychotherapy. The Canadian Journal of Psychiatry. 1990; 35:41–49.

Kivlighan DM, Shaughnessy P. Analysis of the development of the working alliance using hierarchical linear modeling. Journal of Counseling Psychology. 1995; 42:338–349. doi: 10.1037/0022-0167.42.3.338.

Kivlighan DM, Shaughnessy P. Patterns of working alliance development: A typology of client's working alliance ratings. Journal of Counseling Psychology. 2000; 47:362–371. doi: 10.1037/0022-0167.47.3.362.

Klee MR, Abeles N, Muller RT. Therapeutic alliance: Early indicators, course, and outcome. Psychotherapy. 1990; 27:166–174. doi: 10.1037/0033-3204.27.2.166.

Klein DN, Schwartz JE, Santiago NJ, Vivian D, Vocisano C, Castonguay LG, … Keller MB. Therapeutic alliance in depression treatment: controlling for prior change and patient characteristics. Journal of Consulting and Clinical Psychology. 2003; 71:997–1006. doi: 10.1037/0022-006X.71.6.997. [PubMed: 14622075]

Kokotovic AM, Tracey TJ. Working alliance in the early phase of counseling. Journal of Counseling Psychology. 1990; 37:16–21. doi: 10.1037/0022-0167.37.1.16.

Kramer U, de Roten Y, Beretta V, Michel L, Despland J-N. Alliance patterns over the course of short-term dynamic psychotherapy: The shape of productive relationships. Psychotherapy Research. 2009; 19:699–706. doi: 10.1080/10503300902956742. [PubMed: 19606388]

Krupnick JL, Elkin I, Collins J, Simmens S, Sotsky SM, Pilkonis PA, Watkins JT. Therapeutic alliance and clinical outcome in the NIMH Treatment of Depression Collaborative Research Program: Preliminary findings. Psychotherapy. 1994; 31:28–35. doi: 10.1037/0033-3204.31.1.28.

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

Luborsky L, Crits-Christoph P, Alexander L, Margolis M, Cohen M. Two helping alliance methods for predicting outcomes of psychotherapy: counting signs vs. a global rating. Journal of Nervous and Mental Disease. 1983; 171:480–491. doi: 10.1097/00005053-198308000-00005. [PubMed: 6875532]

Marmar CR, Weiss DS, Gaston L. Towards the validation of the California Therapeutic Alliance Rating System. Psychological Assessment: A Journal of Consulting and Clinical Psychology. 1989; 1:46–52. doi: 10.1037/1040-3590.1.1.46.

Martin DJ, Garske JP, Davis MK. Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. Journal of Consulting and Clinical Psychology. 2000; 68:438–450. doi: 10.1037/0022-006X.68.3.438. [PubMed: 10883561]

Marziali EA. Three viewpoints on the therapeutic alliance: similarities, differences, and associations with psychotherapy outcome. Journal of Nervous and Mental Disease. 1984; 172:417–423. doi: 10.1097/00005053-198407000-00008. [PubMed: 6726213]

Morgan, R.; Luborsky, L.; Crits-Christoph, P.; Curtis, H.; Solomon, J. Predicting the outcomes of psychotherapy by the Penn Helping Alliance Rating Method; Archives of General Psychiatry. 1982. p. 397-402.Retrieved from http://archpsyc.ama-assn.org/cgi/reprint/39/4/397

Ogrodniczuk, JS.; Piper, WE.; Joyce, AS.; McCallum, M. Different perspectives of the therapeutic alliance and therapist technique in 2 forms of dynamically oriented psychotherapy; The Canadian Journal of Psychiatry. 2000. p. 452-458.Retrieved from https://ww1.cpa-apc.org/Publications/Archives/CJP/2000/June/June2000.asp

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Paivio SC, Bahr LM. Interpersonal problems, working alliance, and outcome in short-term experiential therapy. Psychotherapy Research. 1998; 8:392–407. doi: 10.1093/ptr/8.4.392.

Paivio SC, Patterson LA. Alliance development in therapy for resolving child abuse issues. Psychotherapy. 1999; 36:343–354. doi: 10.1037/h0087843.

Patton MJ, Kivlighan DM, Multon KD. The Missouri Psychoanalytic Counseling Research Project: Relation of changes in counseling process to client outcomes. Journal of Counseling Psychology. 1997; 44:189–208. doi: 10.1037/0022-0167.44.2.189.

Piper WE, Boroto DR, Joyce AS, McCallum M, Azim HF. Pattern of alliance and outcome in short-term individual psychotherapy. Psychotherapy. 1995; 32:639–647. doi: 10.1037/0033-3204.32.4.639.

Piper WE, Ogrodniczuk JS, Lamarche C, Hilscher T, Joyce AS. Level of alliance, pattern of alliance, and outcome in short-term group therapy. International Journal of Group Psychotherapy. 2005; 55:527–550. doi: 10.1521/ijgp.2005.55.4.527. [PubMed: 16232112]

Rosenthal, R.; Rosnow, RL. Essentials of behavioral research: Methods and data analysis. 2nd ed.. New York, NY: McGraw Hill; 1991.

Sauer EM, Lopez FG, Gormley B. Respective contributions of therapist and client adult attachment orientations to the development of the early working alliance: A preliminary growth modeling study. Psychotherapy Research. 2003; 13:371–382. doi: 10.1093/ptr/kpg027.

Sexton HC, Hembre K, Kvarme G. The interaction of the alliance and therapy micro process: a sequential analysis. Journal of Consulting and Clinical Psychology. 1996; 64:471–480. doi: 10.1037/0022-006X.64.3.471. [PubMed: 8698939]

Shavelson, RJ.; Webb, NM. Generalizability theory: A primer. Measurement methods for the social sciences series. Vol. Vol. 1. Thousand Oaks, CA: Sage; 1991.

Steering Committee. Empirically supported therapy relationships: conclusions and recommendations of the Division 29 Task Force. In: Norcross, JC., editor. Psychotherapy Relationships that Work: Therapist contributions and responsiveness to patients. New York, NY: Oxford University Press; 2002.

Stiles WB, Agnew-Davies R, Hardy GE, Barkham M, Shapiro DA. Relations of the alliance with psychotherapy outcome: Findings in the Second Sheffield Psychotherapy Project. Journal of Consulting and Clinical Psychology. 1998; 66:791–802. doi: 10.1037/0022-006X.66.5.791. [PubMed: 9803698]

Stiles W, Glick M, Osatuke K, Hardy G, Shaprio D, Agnew-Davies R, Barkham M. Patterns of alliance development and rupture-repair hypothesis: Are productive relationships U-shaped or V-shaped. Journal of Counseling Psychology. 2004; 51:81–92. doi: 10.1037/0022-0167.51.1.81.

Wasserman RH, 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]

Webb, NM.; Shavelson, RJ.; Haertel, EH. Reliability coefficients and generalizability theory. In: Rao, CR.; Sinharay, S., editors. Handbook of Statistics, 26: Psychometrics. Amsterdam, Netherlands: Elsevier; 2006. p. 81-124.

Williams, JBW. A Structured Interview Guide for the Hamilton Depression Rating Scale; Archives of General Psychiatry. 1988. p. 742-747.Retrieved from http://archpsyc.ama-assn.org/cgi/content/abstract/45/8/742

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

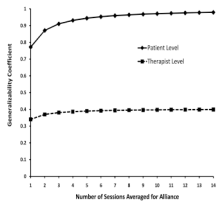

Figure 1.
Patient and therapist level generalizability coefficients for various numbers of sessions at which the alliance is assessed.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

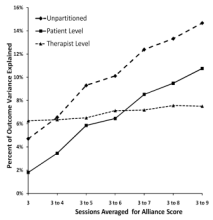

Figure 2.

Percent of outcome variance explained by alliance scores aggregated over various numbers of sessions. The line labeled “unpartitioned” is the percent of outcome variance explained by the alliance without partitioning into patient and therapist levels.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Patient-Level and Therapist-Level Generalizability Coefficients in Primary and Two Replication Samples

	Sample		
	Current Study	**NIDA CCTS**	**Center-Wide Pooled Database**
Patient-Level Generalizability Coefficient			
Based on 1 assessment of alliance	.77	.68	.75
Based on 4 assessments of alliance	.93	.89	.92
Therapist-Level Generalizability Coefficient			
Based on number of patients per therapist actually used in study (one assessment of alliance)	.34	.31	.27
Based on 9 patients per therapist (one assessment of alliance)	.34	.34	.30
Number of patients per therapist needed to achieve a .80 generalizability coefficient (four assessments of alliance)	58	54	70

Note: Current study had 45 patients and 9 therapists. NIDA CCTS had 300 patients and 37 therapists. Center-wide pooled database had 236 patients and 30 therapists. NIDA CCTS = National Institute on Drug Abuse Cocaine Collaborative Treatment Study.