Development and Validation of a 6-item Working Alliance Questionnaire for Repeated Administrations During Psychotherapy

Fredrik Falkenström, Robert L. Hatcher, Tommy Skjulsvik, Mattias Holmqvist Larsson and Rolf Holmqvist

Original Publication:

Fredrik Falkenström, Robert L. Hatcher, Tommy Skjulsvik, Mattias Holmqvist Larsson and Rolf Holmqvist, Development and Validation of a 6-item Working Alliance Questionnaire for Repeated Administrations During Psychotherapy, 2015, Psychological Assessment, (27), 1, 169-183.
http://dx.doi.org/10.1037/pas0000038

Development and validation of a 6-item working alliance questionnaire for repeated administrations during psychotherapy

Fredrik Falkenström[1,2], Robert L. Hatcher[3], Tommy Skjulsvik[2], Mattias Holmqvist Larsson[2] & Rolf Holmqvist[2]

[1] Center for Clinical Research Sörmland, Uppsala University

[2] Department of Behavioural Sciences and Learning, Linköping University

[3] Graduate Center, City University of New York

Author Note

Corresponding author: Fredrik Falkenström, Lustigkullevägen 17, SE-616 33 Åby, Sweden.

Email: Fredrik.Falkenstrom@liu.se

Abstract

Recently, researchers have started to measure the working alliance repeatedly across sessions of psychotherapy, relating the working alliance to symptom change session-by-session. Responding to questionnaires after each session can become tedious, leading to careless responses and/or increasing levels of missing data. Therefore, assessment with the briefest possible instrument is desirable. Because previous research on the Working Alliance Inventory has found the separation of the Goal and Task factors problematic, the present study examined the psychometric properties of a 2–factor, 6-item working alliance measure, adapted from the Working Alliance Inventory, in three patient samples (N = 1095, 235, and 234). Results showed that a bifactor model fit the data well across the three samples, and the factor structure was stable across ten sessions of primary care counseling/psychotherapy. Although the bifactor model with one general and two specific factors outperformed the one-factor model in terms of model fit, dimensionality analyses based on the bifactor model results indicated that in practice the instrument is best treated as unidimensional. Results support the use of composite scores of all six items. The instrument was validated by replicating previous findings of session-by-session prediction of symptom reduction using the Autoregressive Latent Trajectory model. The 6-item working alliance scale, called the Session Alliance Inventory, is a promising alternative for researchers in search for a brief alliance measure to administer after every session.

Keywords: Working alliance, Psychotherapy, Process research, Longitudinal research, Measurement invariance, Confirmatory Factor Analysis, Structural Equations Modeling

Development and validation of a 6-item working alliance questionnaire for repeated administrations during psychotherapy

The working alliance concerns the quality of the collaborative relationship between therapist and patient in the process of psychotherapy, and has been studied extensively as a predictor of psychotherapy outcome (e.g. Horvath, Del Re, Fluckiger, & Symonds, 2011). Most studies to date have relied upon a single alliance measurement early in treatment to predict outcome in the form of symptom relief or improvement in functioning at the end of treatment. Lately, however, more complex models of the alliance-outcome relationship have been used: in a small number of studies, the alliance has been measured after each psychotherapy session and this measure was used to predict symptom change to the next session (Crits-Christoph, Gibbons, Hamilton, Ring-Kurtz, & Gallop, 2011; Falkenström, Granström, & Holmqvist, 2013; Hoffart, Øktedalen, Langkaas, & Wampold, 2013; Tasca & Lampard, 2012).

The working alliance is usually measured using self-report instruments, most commonly filled out by the patient. In contrast to observer methods for measuring the working alliance, questionnaires tend to focus not just on the current session, but on the quality of the alliance so far in treatment. This means that when the alliance is measured after session three, questions focus not just on the quality of the alliance in session three, but on the quality of the alliance from treatment start up to and including session three. This makes sense in studies where the alliance is measured just once early in treatment, because then researchers want to capture the alliance not just in that session but the patient's sense of an emerging alliance during the early phase of treatment. However, in studies where researchers measure the alliance in each session, an instrument is needed that focuses on the alliance in that particular session regardless of how the alliance has been in previous sessions.

One of the most common instruments used when measuring the working alliance is the Working Alliance Inventory (WAI; Horvath & Greenberg, 1989). Two brief versions of this instrument have been published; the WAI-S (Tracey & Kokotovic, 1989) and the WAI-SR (Hatcher & Gillaspy, 2006), both reduced to 12 items compared to the original 36. The WAI-SR, which has showed the best psychometric properties, has an overall coefficient alpha of around .95, which may indicate item redundancy (Falkenström, Hatcher, & Holmqvist, submitted). Measurement after every session is best accomplished with the briefest possible instrument to avoid increased attrition and/or careless responses.

According to the most influential alliance theory which also guided the construction of the WAI (Bordin, 1979; Horvath & Greenberg, 1989), the working alliance consists of three components; agreement on treatment goals, agreement on therapeutic tasks, and a positive emotional bond between patient and therapist. All versions of the WAI accordingly attempt to capture these three components. Factor analyses, however, have more or less consistently failed to differentiate the goal and task factors, with factor correlations usually being around .90 (Falkenström et al., submitted; Munder, Wilmers, Leonhart, Linster, & Barth, 2010). Most applied studies use a composite aggregate of all WAI items as a general alliance measure, citing the Tracey and Kokotovic (1989) finding that a general alliance factor explains most of the common variance among the items.

Most factor analytic studies of the WAI have compared one-, two-, and three factor models (e.g. Falkenström et al., submitted.; Hatcher & Gillaspy, 2006; Munder et al., 2010). However, the first factor analytic study of the WAI (Tracey & Kokotovic, 1989) found that a “bilevel hierarchical” model best represented the covariances among the WAI items, a model nowadays usually referred to as “bifactor” (e.g. Reise, 2012). In this model all of the items load on one general factor, and relevant item subsets load on specific “group” factors. The group factors are specified as orthogonal to the general factor, and usually also to each other. Thus, the group factors are interpreted as representing the variance not accounted for by the general factor (i.e. the residuals from a one-factor model). Although the bifactor model has been criticized, mainly because the orthogonality constraints among factors make interpreting the factors difficult (e.g. Vanheule, Desmet, Groenvynck, Rosseel, & Fontaine, 2008), there are several advantages of the bifactor model: it usually fits data better than competing models and it permits calculating measures of the degree of unidimensionality, and model-based indices of the reliability of general and subscale composite scores (Reise, 2012). Because of this, the bifactor model may be useful in assessing whether interpretation of total and/or subscale scores is possible.

An often-overlooked issue (at least within clinical psychology) that arises with repeated administration of any instrument is whether the factor structure stays the same over time. This is referred to as longitudinal measurement invariance, and is as important to establish as reliability and validity (Vandenberg & Lance, 2000). Specifically, if the factor structure changes with time, aggregate composite scores cannot be compared across time because participants change their interpretation of item content. The change in means of observed variables may then not reflect change in the latent construct, but rather change in the way participants relate to item content. In principle, any parameter of a Confirmatory Factor Analysis can be subjected to measurement invariance tests, but factor loadings and indicator intercepts are the most important ones. This is because these are the parameters that represent the measurement part of the CFA, which should stay the same over time so that the researcher can interpret change in the latent means that represent “true” change. However, if the measurement part changes with time, it is impossible to interpret change in latent factors.

Steinmetz (2013), using Monte Carlo simulations, tested the amount of bias in composite scores due to violation of measurement invariance. He found that inequality of indicator intercepts creates the most bias in composite scores. Indicator intercepts are constants added to the measurement equation in order to set the metric of the scale, just like intercepts in regular regression analysis, while factor loadings are slopes or weights for how much each indicator is influenced by the latent factor. Steinmetz simulations showed that just one unequal indicator intercept across groups increased probability of Type I error substantially when composite scores were used to test mean differences, and with half the intercepts unequal the probability of spurious mean differences was increased by as much as 60.7%. Unequal factor loadings, however, had negligible effects on composite scores.

The aim of the present study was to develop a brief measure of the working alliance that can be used for session-wise administration over the course of psychotherapy. In addition, we wanted to test the factor structure of this brief version using Confirmatory Factor Analysis, and to use the bifactor model to assess the degree of multidimensionality and reliability of subscale scores. We also wanted to test the stability of the factor structure over time using longitudinal measurement invariance analyses. Finally, should the measure hold up to these stringent tests, we wanted to test its validity in predicting symptom improvement session-by-session.

Methods

Participants

Sample 1. This sample, which was the primary sample for this study, consisted of patients attending primary care counseling and psychotherapy of different orientations (most were versions of CBT or Psychodynamic therapy) at two service regions in Sweden. The sample consisted of altogether 1061 patients who filled out the Swedish translation of the WAI-SR at least once during any of the first ten sessions. At Session 1 there were 1006 complete questionnaires (i.e. 55 missing). With each additional session the number of patients declined at a negatively accelerating rate until at Session 10 there were 120 patients who filled out both WAI-SR and CORE-OM. This pattern is typical of naturalistic data from clinical settings, with most patients having very brief therapeutic contacts (e.g. Hansen, Lambert, & Forman, 2002; Stiles, Barkham, Mellor-Clark, & Connell, 2008). Some patients continued up until a maximum of approximately 30 sessions, but due to the complexity of the models that were to be tested a limit was set at ten sessions because of the likelihood that estimation would not work properly with the small sample that continued beyond session ten. Demographic information was available for between 75 to 80 percent of the patients. The mean age was 37.3 years (median 35, *SD* = 14.3, range 14-88), 74 % were women and 92 % were born in Sweden. More details are available in Falkenström, Granström and Holmqvist (2013; 2014) and in Holmqvist, Ström, & Foldemo (2014).

Sample 2. This sample was the second sample used by Hatcher and Gillaspy (2006) in the development of the Working Alliance Inventory – Short form Revised (WAI-SR, see below) for cross-validating the factor structure derived from an exploratory factor analysis of their first sample. The sample consisted of 235 adult outpatient clients (71% women, 24% men, and 5%

unidentified as to gender) from a number of counseling centers and outpatient facilities primarily from the southwestern United States filling out the WAI at session three. The clients were treated using different psychotherapeutic approaches, although most common were cognitive-behavioral (33%) and psychodynamic (25%) treatments. Age ranged from 18 to 64 (M=28.4, SD=9.9). In this sample the alliance was only measured in Session 3. More details are available in Hatcher and Gillaspy (2006).

Sample 3. A third sample was composed of 234 patients from specialist psychiatric departments throughout Sweden, from an ongoing naturalistic study of routine psychotherapy delivered in psychiatric care. The patients were treated using psychotherapy of different orientations, and the WAI-SR data were taken from session three in order to be comparable to the other two samples. Demographic information was unavailable at the time the present analyses were done.

Measures

Working Alliance Inventory – Short form Revised (WAI-SR; Hatcher & Gillaspy, 2006). The Working Alliance Inventory is one of the most common alliance measures (Horvath et al., 2011). The WAI-SR was developed from the original 36-item Working Alliance Inventory (Horvath & Greenberg, 1989) using Exploratory and Confirmatory Factor Analyses together with Item Response Theory modeling in order to reduce the number of items. The factor structure of the WAI-SR has subsequently been tested in at least two studies (Falkenström et al., submitted; Munder et al., 2010). In the present study, patients in Sample 2 filled out the original English language version of the WAI-SR, while Samples 1 and 3 used a Swedish translation made by Holmqvist and Skjulsvik (2013). The translation was done using back-translation and modifications in several steps. Although Hatcher and Gillaspy (2006), based on their IRT analyses, recommended a five-point response scale, in the present study the original seven-point response scale was used in all three samples. For cross-sectional analyses, data from Session 3 was used. The reasons for this choice was 1) that this is a common choice in alliance research since researchers want to allow enough time for the alliance to develop but not too much time so that most of outcome has already occurred, and 2) in Sample 2 the alliance was only measured in Session 3 and it was deemed preferable to use the same session in all three samples.

Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM; Evans et al., 2002). The CORE-OM is a self-report measure consisting of 34 items measuring psychological distress experienced during the preceding week, on a five-point scale ranging from “Not at all” to “Most or all the time”. The items cover four major problem areas: subjective wellbeing, problems/symptoms, life functioning, and risk (to self or others). Higher scores indicate greater distress. The present study used the Swedish version (Elfström et al., 2012). This version has shown excellent internal (.93-.94) and test-retest (.85) reliability, convergent validity and sensitivity to change (Holmqvist et al., 2014). The CORE-OM was completed immediately before each session. In the present study, only the total (composite) score, which has a possible range between 0 and 40, was used. This total score was calculated as the mean of all 34 items multiplied by 10, according to standard CORE procedure.

Item reduction

Because of the high intercorrelation between Task and Goal factors in previous studies, it was deemed unrealistic that a 6-item measure would be capable of distinguishing between them.

These two scales were therefore combined into one Task/Goal factor, while trying to keep at least one item from each original scale to preserve some of the content from the original scales. Since three items are usually seen as providing enough information to estimate a latent variable (e.g. Kline, 2011; Little, 2013), we wanted to have three Bond items and three items from the combined Goal and Task scales.

We used our previous CFA of the WAI-SR (Falkenström et al., submitted) to select items that would be suitable to include in an ultra-brief version of this scale. Residual correlation analyses in that study found items 1 ("As a result of these sessions I am clearer as to how I might be able to change") and 2 ("What I am doing in therapy gives me new ways of looking at my problem") to have relatively large residual correlations in two of the three samples studied. Content analysis of these items showed that they may be interpreted as asking for how well therapy is going (i.e. outcome so far in treatment) rather than alliance specifically. When looking at other items this way, item 11 ("/My therapist/ and I have established a good understanding of the kind of changes that would be good for me") might also be interpreted in a similar way. Because of this we decided to exclude items 1 and 2 (from the Task scale) and 11 (from the Goal scale).

Due to similarity in content between items 3 ("I believe /my therapist/ likes me") and 7 ("I feel that /my therapist/ appreciates me") from the Bond scale, we decided to keep only item 7 for the ultra-brief version. The choice between these two items was more or less arbitrary (see Hatcher, 2010, for further discussion). Items 4 ("/My therapist/ and I collaborate on setting goals for my therapy") and 6 ("/My therapist/ and I are working towards mutually agreed upon goals") from the Goal scale were also considered relatively similar in content, so only item 6 was included since it was deemed more general (item 4 may be most relevant in the first few sessions, while item 6 may be relevant throughout treatment). Finally, we decided not to include item 10 (“I feel that the things I do in therapy will help me to accomplish the changes that I want”), because it may not be specific to the alliance but may reflect hope, positive spirit, or the fact that the therapist uses a good therapy method.

This item reduction process left us with the following items for the Bond scale:

5. My therapist and I respect each other.

7. I feel that /my therapist/ appreciates me.

9. I feel that /my therapist/ cares about me even when I do things that he/she does not approve of.

The following items were included for the combined Task/Goal scale:

6. /My therapist/ and I are working towards mutually agreed upon goals.

8. /My therapist/ and I agree on what is important for me to work on.

12. I believe the way we are working with my problem is correct.

Statistical analyses

We first compared four models; a one-factor model, a two correlated factors model, and bifactor models with one general alliance factor and two “group” (i.e. specific) factors corresponding to the Bond and Task/Goal dimensions. Two versions of the bifactor model were tested, one with orthogonal group factors and one with correlated group factors. Figure 1 shows the four models.

In preliminary analyses we found that the bifactor models did not converge when using Maximum Likelihood estimation. This most likely reflected identification problems. We therefore chose to use Bayesian estimation with informative priors for the factor loadings (e.g. B.

O. Muthén & Asparouhov, 2012). In Bayesian estimation, prior information is used to inform estimation, which allows estimation of many models that are not identified using other estimation methods. For more information on Bayesian statistics, see Appendix A.

Priors for the loadings on the general alliance factor were set to a mean of .7, because previous analyses (Falkenström et al., submitted) showed that the amount of shared variance in a two-factor model was about 50% (i.e. $.7^2 = .49 \approx .50$). Priors for the loadings on the Bond and Goal/Task factors in two correlated factors model were also set to a mean of .7. For the bifactor models the loadings for the group factors were set to a mean of .5, reflecting our belief that these loadings would be smaller than the ones on the general factor.

The variances for the informative priors, which indicate the degree of certainty the researcher has about the prior means (and the degree of influence the priors will have on the posterior distribution) were chosen to be as large as possible while still allowing for reasonably efficient estimation. When prior variance gets too large, models will not converge, and when prior variance is too low the prior will dominate the posterior distribution (B. O. Muthén & Asparouhov, 2012) which in this case would be undesirable. After some initial experimentation with different variances for the factor-loading priors, the variances were set to .02, which corresponds to a 95% confidence interval of ± .28. These priors were the same in the one-factor model and the bifactor models. When factor means were estimated, informative priors for the means and intercepts were used instead of the usual methods of scaling (i.e. marker variable, fixed factor, or effects coding; Little, 2013). The normal distribution prior for the latent means was set at mean zero, and for the intercepts at 5.5 – a value taken from descriptive statistics in previous analyses of the WAI-SR (Falkenström et al., submitted). After initial exploration, the variances for the priors for the factor means and intercepts were set at .5 (the highest value that permitted estimation to converge), corresponding to a 95% confidence interval of ±1.4.

As an initial step, we estimated separate CFA:s for the first ten sessions of Sample 1. Although we had data on more than ten sessions, the sample size decreased with each additional session due to fewer patients attending longer therapies, so we chose ten sessions as a reasonable limit because the sample size was below 100 subjects for sessions 11 and higher. When no latent means were estimated the models converged relatively rapidly (i.e. around 20-30,000 iterations), but estimation was allowed to continue up until at least 50-60,000 iterations to ensure convergence. When latent means with large variance informative priors were estimated, convergence seemed to take until around 80,000 – 100,000 iterations, and estimation was continued until 150,000 - 200,000 iterations to ensure convergence.

For the longitudinal measurement invariance analyses, the mean of the general factor at Session 1 was constrained to zero and the variance to 1 in order to facilitate model identification. All other means and variances were estimated freely. However, informative priors were needed for the factor variance-covariance matrices in order to get the MCMC estimator to converge. To this end, Inverse Wishart (IW) distribution priors were used. Muthén and Asparouhov's (2012) Method 2, which performed best of the three methods tested in their simulation study, was used. Because the general factor was not allowed to correlate with the group factors, covariances for the general factor at different sessions was estimated as a separate IW block. Covariances among the group factors were estimated for all combinations of sessions (within and between factors). The prior for the factor variances was set at a mean of 1. The prior for the covariances among the same factor at different sessions, and between the two group factors at the same session, was set at a mean of .50, while covariances between the group factors at different sessions was set at a mean of .30. These values were deemed plausible, since a relatively strong autocorrelation component would be expected from prior research (e.g. Falkenström et al., 2013). Less is known about the cross-factor covariances at different sessions, but since the correlations at the same sessions were relatively strong, a positive, slightly smaller value than the autocorrelation, was expected.

The variance for the IW prior is set using degrees of freedom, which is the sum of the number of diagonal elements in the covariance matrix (p) plus a number chosen for the degree of informativeness. Initial explorations indicated that a prior with df = p+30 had to be used for estimation to converge with reasonable estimates (e.g. factor variances below 5, non-negative correlations among factors) while still not influencing the posterior distribution too much (as seen for instance when all factor variances are estimated as exactly 1). An IW prior with p+30 degrees of freedom corresponds to a variance of 0.045 (i.e. SD=0.21) for the covariances, and 0.074 (i.e. SD=0.27) for the variances. Another way of evaluating the degree of informativeness for the IW distribution prior is that df=p+30 is equivalent to the effect of adding 30 cases to the dataset (B. O. Muthén & Asparouhov, 2012).

In the longitudinal measurement model, correlations among residuals for the same items at different sessions (so-called correlated uniqueness) were estimated freely, as is usually recommended (e.g. Little, 2013). Measurement invariance was first tested using the traditional way, postulating exact invariance of factor loadings and intercepts (e.g. Little, 2013; van de Schoot, Lugtig, & Hox, 2012). If exact invariance did not hold, BSEM “approximate invariance” (B. O. Muthén & Asparouhov, 2013; van de Schoot et al., 2013) was tested. Convergence for the longitudinal models seemed to be achieved by around 100,000 – 120,000 iterations, but all models were subsequently run for between 200,000 – 250,000 iterations to ensure convergence.

Missing data was handled by including all available information in the analyses, an approach that has been shown to yield more or less equivalent results as Multiple Imputation (e.g. Graham, Olchowski, & Gilreath, 2007). The argument for including all available information rather than deleting cases with missing data is that the less restrictive assumption of Missing-At-Random applies for cases included in the analysis, while for deleted cases Missing-Completely-At-Random is assumed (Enders, 2011; Rubin, 1976).

All analyses were done with the software Mplus 7.2 (L. K. Muthén & Muthén, 2012), using tests for continuous scales

Results

Descriptive statistics

Table 1 shows means, standard deviations, range, skewness, and kurtosis for Session 3 of all three samples. Descriptive statistics for the other sessions used in longitudinal CFA models, and correlation matrices can be obtained from the authors.

Confirmatory Factor Analysis

Table 2 shows model fit information for the four models estimated at Session 3. In the largest sample, only the bifactor model with correlations between group factors showed adequate model fit in terms of a non-significant posterior predictive *p*-value. In addition, this model outperformed the other models in terms of relative model fit according to the information criteria DIC and BIC. In the two smaller samples the bifactor model with orthogonal group factors showed almost as good model fit as the bifactor model with correlated group factors. When correlations between group factors were allowed, the size of these was .59 in Sample 1, .53 in Sample 2, and .78 in Sample 3.

Nesting within therapists

In all three samples there was nesting of patient data within therapists, i.e. most therapists saw more than one patient. In Samples 2 and 3 there was no information on which patients saw the same therapists, but in Sample 1 it was possible to estimate the degree of statistical dependency due to therapists. In Session 3 of Sample 1, the IntraClass Correlations (ICC) ranged between .03 for item 5 to .14 for item 6. Thus, statistical dependency was not alarmingly high, but possibly high enough to influence results for some of the items. For this reason a two-level CFA was estimated, in which therapist-level random effects were estimated for all six items. The within-therapist model was the bifactor model with correlated group factors, while on the therapist level only means and variances were estimated since we were only interested in this level as a way of controlling for dependency (and the number of therapists (N = 69) was deemed too small to enable a separate CFA to be estimated for the between-therapist level). Since no model fit indices are available for Bayesian two-level models, we compared standardized factor loadings between the two-level model and the single-level model to estimate the degree of bias in parameters due to ignoring clustering. Factor loadings were highly similar, with the largest difference in standardized loading being .02. We therefore concluded that results of single-level models were likely to be unbiased.

The use of composite sum scores

Despite the fact that the bifactor model outperformed the one-factor model in terms of model fit, it is possible that composite scores of all six items can be used. In order to test this, the amount of parameter bias introduced by treating the scale as unidimensional has to be determined. Reise et al. (2012) showed that when the Percentage of Uncontaminated Correlations (PUC), defined as the ratio of correlations between items belonging to different group factors to the total number of unique inter-item correlations, is higher than .80, bias in structural coefficients due to multidimensionality is negligible. When the PUC is lower than .80, the proportion of explained variance for the general factor to group factors also has to be weighed in. For the 6-item alliance measure, the PUC is only .60[1]. Because of this, estimates from the bifactor model with correlations between all three factors constrained to zero were used to calculate the Explained Common Variance statistic (ECV; Reise, 2012), which is defined as the ratio of explained variance (sum of squared loadings) for the general factor divided by the sum of explained variance for general plus group factors. The ECV was .80 in Sample 1, .76 in Sample 2, and .85 in Sample 3, indicating that around 75-85% of variance was due to the general factor. Also, the Omega Hierarchical coefficient (Reise et al., 2012; Reise, 2012) was calculated for the general factor. Omega Hierarchical can be interpreted as an estimate of how much variance in summed scores can be attributed to a single general factor. In the three samples, Omega Hierarchical was .84, .78, and .84, respectively.

In cases when the PUC is lower than .80, Reise et al. (2012) proposed ECV > .60 and Omega Hierarchical > .70 as a tentative benchmarks for when structural coefficients are estimated with negligible bias. This means that the composite score of all six items of the 6-item alliance instrument is likely to be unbiased. The reliability of the total composite score, calculated as coefficient Omega (Lucke, 2005), which is analogous to coefficient alpha, for the full scale was .94 in Sample 1, .89 in Sample 2, and .92 in Sample 3.

Comparisons with the full WAI and with the WAI-SR

The total (composite) score of the 6-item alliance instrument was correlated with the total score for the full WAI (Horvath & Greenberg, 1989) and for the WAI-SR (Hatcher & Gillaspy, 2006). Data on the full WAI was only present in Sample 2, while WAI-SR data was present in all three samples. Results for Sample 2 showed that the 6-item version correlated .91 with the full WAI and .95 with the WAI-SR. In Session 3 of the two Swedish samples, the correlation between the 6-item version and the WAI-SR was .96. This means that the total score of the 6-item alliance instrument measures more or less the same concept as the full WAI and (perhaps especially) the WAI-SR.

The use of subscale scores

For both subscales, coefficient Omega was high, ranging from .85 for the Bond scale in Sample 2 to .90 for the Goal/Task scale in Sample 1. However, when the variance due to the general factor was removed, reliability dropped to between .13 for both Bond and Goal/Task scales in Sample 3 to .23 for the Bond scale in Sample 2. These latter figures should be interpreted as the reliability for residualized subscale scores from which the variance due to the general factor has been removed. Thus, little reliable variance seems to be left in the subscales beyond that attributable to the general factor.

Longitudinal measurement invariance

Table 3 shows model fit and standardized factor loadings for all ten sessions when separate CFA:s were estimated for each session. As can be seen, model fit was good to excellent for all sessions, and factor loadings were high except for the loading of item 5 on the Bond group factor which was low (<.40) in most sessions. The intercorrelation between the group factors (not shown in the table) ranged between .75 in Session 1 and .43 in Session 9. Estimated indicator intercepts and factor means are shown in Table 4. These estimates seem to indicate invariance at least of intercepts across sessions, although this remains to be tested. Factor means tended to increase slightly over time – at least for the general factor.

In order to formally test measurement invariance, a longitudinal CFA model was set up for the first ten sessions of Sample 1. Model fit information for different models is presented in Table 5. First, a Configural model, in which all factor loadings and intercepts are allowed to differ, was estimated. Model fit for this model showed some evidence of lack of fit, with a posterior predictive *p*-value of .02 (95% CI: 4.96, 360.0). However, with an effective sample size of 1061[2], and 1027 degrees of freedom, the amount of misspecification indicated by a posterior predictive *p*-value of .02 might very well be trivial. Still, because model fit was excellent when the bifactor models were estimated for each session separately, the misfit for the longitudinal CFA had to be due to misfit in the modeling of relationships between indicators across sessions. Since the correlations among latent factors were estimated more or less freely (i.e. freely but with large variance informative priors), the most plausible explanation for the lack of model fit would be a large number of small residual correlations among different item indicators at different sessions (residual correlations among same item indicators at different sessions were already estimated freely in the Configural model). This latter hypothesis can be tested using BSEM with informative priors for residual correlations (B. O. Muthén & Asparouhov, 2012).

We therefore re-estimated the Configural model, but this time with all possible residual covariances estimated using informative priors for the residual variance-covariance matrix, with a mean of .5 for the diagonal (i.e. residual variances) and a mean of zero for the off-diagonal elements (i.e. residual covariances). The variances of the priors were set so that the 95% confidence intervals would span approximately from .20 above to .20 below the mean. Model fit for this model turned out to be excellent, with a Posterior Predictive p-value of .39 (95% CI: -152.07, 200.69), thus confirming that misfit for the original Configural model was due to residual correlations assumed to be exactly zero in this model. Inspection of estimates for the residual correlations showed that the BSEM assumption of trivially small residual correlations (B. O. Muthén & Asparouhov, 2012) held; the standardized residuals ranged from -.24 to .41 with a mean of .01 (absolute mean = .09) and a standard deviation of .09. In total, 1770 residual correlations were estimated, and 117 were statistically significant at the .05 level. Of the residual correlations that were statistically significant, 77 were between the same items measured at different sessions – which are expected to correlate. Counting only correlations between different items, only 2.7% were statistically significant, which of course is less than would be expected by chance at the 5% significance level. In conclusion, it seems safe to say that model misfit of the Configural model could be explained by many trivially small residual correlations.

The next step in a measurement invariance analysis involves constraining the factor loadings to be equal across sessions, a model that is called variously Weak Factorial Invariance (Little, 2013) or Metric Invariance (van de Schoot et al., 2012). This model fit worse than the Configural model according to the DIC, but better according to the BIC (see Table 5, $\Delta DIC_{metric-configural} = 87.2$, $\Delta BIC_{metric-configural} = -511.2$). Because the BIC penalizes more for estimating more parameters than the DIC does, the evaluation of the invariance restriction depends on the relative weight one puts on model fit (DIC) or parsimony (BIC). In this respect, the DIC makes for a more stringent test of measurement invariance than the BIC; possibly too stringent since the DIC has been reported to favor overparameterized models (Ando, 2011)[3].

Next, indicator intercepts were constrained to be equal across sessions. This model, with both factor loadings and intercepts constrained to be equal over time, is called Strong Factorial Invariance (Little, 2013) or Scalar Invariance (van de Schoot et al., 2012). Model fit for this model was worse than for the Configural model and slightly worse than the Metric model according to the DIC ($\Delta DIC_{scalar-metric} = 6.0$, $\Delta DIC_{scalar-configural} = 93.2$), but better than both Configural and Metric models according to the BIC ($\Delta BIC_{scalar-metric} = -345.6$, $\Delta BIC_{scalar-configural} = -856.8$).

Because the exact invariance tests failed (at least according to the DIC), we went on by testing BSEM approximate measurement invariance (B. O. Muthén & Asparouhov, 2013), in which informative priors with zero means and small variances are assigned for the differences between parameters across time. This model has particular advantages when many groups or many time points are compared, as in the present case, because it allows identification of invariance of particular parameters at specific time points in a manner resembling the use of modification indices in Maximum Likelihood SEM. Thus, the BSEM invariance model is useful for diagnosing if particular items or particular sessions are causing parameter variance. Instead of assigning informative priors for the *values* of the factor loadings as in the previous models, informative priors are set for the *differences* among estimates of the same parameters measured at different time points. This was done for factor loadings and intercepts. We started with a normal prior with zero mean and variance .01 (i.e. 95% confidence interval spanning from -.20 to .20) for the differences, and then we re-estimated the model using larger variance priors as sensitivity analyses, as recommended by Muthén and Asparouhov (2013). Although it may seem obvious that assigning larger variances to the difference priors will result in more parameters being estimated with larger differences, it is not clear how the significance of these differences is affected because standard errors also tend to increase when the prior variances increase.

Model fit indices for models with variances of .01 and .05 for the differences between parameters at different sessions are shown in Table 5. Both models fit slightly better than the Configural model according to the DIC. The BIC favored the BSEM model with prior variance .05, while the BSEM model with prior variance .01 had the highest BIC of these three models. A model with a variance of .10 for the difference prior was not even close to convergence at 200,000 iterations, so attempts to estimate such a model were abandoned. For the loadings on the general factor, there were five loadings (out of a total of 60) that differed significantly from the mean of loadings for the respective indicator across sessions. Of these, three were from Session 1 (items 6, 7, and 9), and the fourth and fifth were item 5 at Session 8 and item 6 at Session 9. When the prior variance was increased from .01 to .05, items 8 and 12 from Session 1 and item 5 at Session 5 also differed significantly from the mean. For the Goal/Task factor, there were two loadings (out of 30) that differed significantly from the mean; item 8 from Session 1 and item 12 from Session 9. When the prior variance for the differences was increased to .05, the first of these was no longer significant. For the Bond factor there were six loadings (out of 30) that differed significantly. Two of these were from Session 1 (items 5 and 7), one from Session 3 (item 5), and all three items from Session 8 (items 5, 7, and 9). When the variance for the difference prior was increased to .05, item 5 at Session 8 was no longer significant; otherwise the results were the same. Deviations were generally larger in the model with prior variance .05, as expected. In the model with prior variance .05, the largest deviation from the mean of standardized loadings across sessions was .23, which was for item 9 on the general factor at Session 1. If Session 1 loadings were not counted, the largest deviation from the mean across sessions was .08 – a deviation that is unlikely to have any substantial impact on measurement.

No intercept differed significantly from the mean across sessions, regardless of prior variance for the difference. A model with .01 variance for the loadings and .10 for the intercepts was also tested, to see if increasing the variance specifically for the intercept priors would yield more significant deviations for these parameters. This model converged, but still no intercept differed significantly from the mean. The largest (non-significant) deviation from the mean of intercepts across sessions was .12 for item 9 at Session 1.

To sum, the pattern that could be discerned was that factor loadings at Session 1 may differ from the other sessions, especially for the general factor. The other significant factor loading differences most likely were due to random error, since there was no pattern to them and their variation seemed small in magnitude. With 180 differences estimated, nine would be expected to be significant at p <.05 due to chance alone. This is more than the number of significances we found, if the ones for Session 1 are not counted. In addition, with the sample size used, statistical power was high to find small deviations from invariance.

Because the BSEM invariance model is designed to search for a solution in which the variance across time points for each parameter is small, it can yield biased results if the data do not conform to the BSEM structure (i.e. many small deviations from invariance rather than a few large deviations). Due to this “alignment issue”, Muthén and Asparouhov (2013) recommended a two-step procedure: after identification of non-invariant parameters, the model should be re-estimated with the non-invariant parameters estimated freely while all other parameters are constrained to exact equality. Model fit for this model is reported in Table 5 as Partial invariance model. As shown in Table 5, both the DIC and the BIC favored this model to the Configural model ($\Delta DIC_{\text{partial-configural}} = -6.6$, $\Delta BIC_{\text{partial-configural}} = -875.6$), indicating that Partial measurement invariance held.

Validation: prediction of symptom reduction from session to session

A previous study on the same dataset as used in this paper (Sample 1) showed that the WAI-SR predicted change in symptom scores from session to session, even after controlling for the “reverse causation” effect of prior symptom change on alliance scores (Falkenström et al., 2013). The method used in that study was multilevel path analysis with lagged response variables, using composite means of all 12 items of the WAI-SR. In the present paper, we wanted 1) to replicate this analysis using the 6-item alliance measure, in order to test the validity of this scale, and 2) to use a more established statistical method, since there are some potential problems with the use of lagged response variables (Rabe-Hesketh & Skrondal, 2012) [4].

The previous analysis used a two-step procedure, with the first step being to estimate separate OLS regressions for each variable (WAI-SR and CORE-OM) on session number, saving residuals from these analyses as variables to use in the second step, which was a cross-lagged path analysis with variables in the Long format of Multilevel Modeling. The two-step procedure was used in order to isolate the within-patient effects and to control for non-stationarity, issues that may bias results if not accounted for (Curran & Bauer, 2011). In the present study, we instead used the Autoregressive Latent Trajectory model (Bollen & Curran, 2004; Curran & Bollen, 2001), which combines Latent Growth Curve Modeling with autoregressive cross-lagged modeling. Using the ALT model (with variables in the Wide format of Structural Equations Modeling), Working Alliance and CORE-OM scores are decomposed into within- and between patient components by estimating random intercepts at the patient level (Curran, Lee, Howard, Lane, & MacCallum, 2012); and control for non-stationarity in the time-series is accomplished by estimating random slopes for the linear trends over time. The previous analyses were restricted to those patients who provided at least three CORE-OM and WAI-SR questionnaires, because this is required for the residualized centering approach. In contrast, the ALT model can incorporate all patients who provided at least one questionnaire. On the other hand, the ALT model was restricted to the first ten sessions, for the same reason as the longitudinal CFA:s.

In the ALT model, the first measurement occasion is typically treated as "pre-determined" and is thus not part of the linear trajectory, but its mean and variance are estimated separately. Initial analyses with ALT models estimated separately for the two variables indicated that for the CORE-OM, the autoregressive effect and residual variances could be constrained to be equal across sessions without significant loss of model fit (Satorra-Bentler ΔChi-square = 3.44 (8), *p* = .90 for autoregression, and ΔChi-square = 5.50 (8), *p* = .70 for residuals)[5]. For Working Alliance, autoregression and residuals were estimated separately for each occasion. Cross-lagged effects were first estimated separately for each session for both variables, but since a Chi-square difference test showed non-significant reduction in model fit when constraining the cross-lag for the regression of CORE-OM on Working Alliance to equality (Satorra-Bentler

ΔChi-square = 5.23(8), *p* = .73) this was done, although the regression of Working Alliance on CORE-OM was estimated separately for each session (constraining these to equality led to Satorra Bentler ΔChi-square = 21.13(9), *p* = .01).

The random effects in this model, representing the between-patient estimates, were not of primary interest in this analysis since we were mostly interested in the within-patient effects. We chose to freely estimate variances for all random effects, and correlations between random effects within each variable. Correlations between random effects of different variables were only estimated for similar random effects, i.e. the Intercept of CORE-OM was allowed to correlate with the Intercept of Working Alliance, and the Slope of CORE-OM was allowed to correlate with the Slope of Working Alliance. However, no correlations were estimated between the Intercept of CORE-OM and the Slope of Working Alliance and vice versa. Similarly, the initial observation of each measure was allowed to correlate with the random effects of the same variable (i.e. CORE-OM at Session 1 with Intercept and Slope of CORE-OM and vice versa), but no cross-variable correlations were estimated for these either. Comparing this model with a model with all correlations between random effects estimated freely showed non-significant reduction in model fit (Satorra Bentler ΔChi-square = 20.65(14), *p* = .11.

The final model, shown in Figure 2, was estimated using Maximum Likelihood estimation with robust standard errors and showed excellent model fit (Likelihood Ratio Chi-square = 200.88, df = 187, *p* = .23, N = 1095). The cross-lagged effect of the alliance on subsequent symptoms was very similar to the previous study; a small but statistically significant negative value indicating that higher alliance scores in one session predicted less severe symptoms in the next session (b = -.36, se = .12, 95% CI -0.59, -0.13, *p* = .002) while controlling for the "reverse causation" of symptoms predicting alliance scores (which was also statistically significant). Previous research on unbalanced naturalistic psychotherapy data has shown the slopes of symptom change to be related to treatment length, thus violating the Missing-At-Random assumption. Since the cross-lagged effects of primary interest in the present analyses were defined as deviations from the patient-specific intercepts and slopes, biased estimates of random effects would also bias within-patient estimates. For this reason, the model was re-estimated with treatment length included as a fixed predictor of all between-patient variables (random intercepts and slopes plus Session 1 CORE-OM and Working Alliance). This had negligible impact on the cross-lagged effect of Working Alliance on CORE-OM change.

Discussion

This study tested the factor structure of a 6-item alliance measure based on the Working Alliance Inventory – Short form Revised. Confirmatory Factor Analysis showed that a bifactor structure with one general alliance factor and two group factors fit the data well in three relatively large independent samples. In the primary (largest) sample, a bifactor model with correlations between group factors fit the data best. When correlations between group factors were allowed, these were moderate to large (around .53, .59, and .78 in the respective samples). Dimensionality analyses showed that despite some degree of multidimensionality, the instrument is in practice best treated as unidimensional since most of the variance is explained by the general factor and very little reliable variance is added by the group factors. Reliability for the composite sum or mean of the six items was excellent (between .89 and .94 in the three samples).

The longitudinal measurement invariance analyses for the first ten sessions of primary care counseling/psychotherapy indicated that strong measurement invariance generally held. Indicator intercepts were remarkably stable across the first ten sessions, with none of the six indicator intercepts deviating significantly from the average at any of the ten sessions analyzed despite statistical power to detect small differences. The possible exception from measurement invariance was factor loadings at Session 1, which showed a pattern of deviating from the mean of loadings at all occasions (although even these deviations were not very large). It is possible that this was due to larger power to find statistically significant differences for Session 1, since the sample size was largest at that session. However, there was no pattern to the other significant differences indicating overrepresentation of sessions with larger N. It thus seems more likely that the invariance found for Session 1 was due to actual differences in factor loadings at this time point. This finding is not surprising, because immediately after the first meeting with a therapist it is likely that the working alliance is not yet fully formed (Hatcher & Gillaspy, 2006), making it hard for patients to rate it. Although it is likely that factor loading differences at Session 1 were “true” differences, it is unlikely that these make any substantial difference in practice. In Steinmetz (2013) simulations, differences in factor loadings had minimal impact on composite scores. Apart from Session 1, factor loadings were stable except for a few minor deviations most likely attributable to random error given how many differences were tested and the statistical power to detect small differences.

The fact that indicator intercepts were stable over ten measurement occasions is strong evidence for the stability of the scale, and supports the use of composite scores for this instrument. Factor loadings were slightly less stable, but simulations have shown that the most important aspect of measurement invariance – especially if composite scores are to be used – is invariance of indicator intercepts, with factor loadings having negligible influence on composites (Steinmetz, 2013). Taking the results of the dimensionality analyses together with measurement invariance analyses, our results indicate that the composite mean of all six items of the 6-item alliance scale can be used in longitudinal analyses. Although statistically, latent variable models are preferable to models based on composite scores (e.g. due to the possibility to model measurement errors), using composites is simpler and can greatly reduce computational time for complex models. For instance, each longitudinal CFA model in the present study took several hours to run, in comparison with a few seconds for models based on observed composite scores.

Results of the ALT model showed that previous findings of a session-to-session effect of the alliance on symptom reduction could be replicated using the 6-item alliance measure, thus supporting the validity of this measure. This analysis showed that previous results held even when a more established statistical model was used. An advantage of the ALT model was that it was possible to use information from all patients who had filled out at least one CORE-OM or WAI-SR, making the sample size almost twice as large (N = 1094 compared to N = 636) as the previous analysis, which required at least three filled out measures for both instruments. On the other hand, the ALT model was limited to the first ten sessions while the MLM could use information also from later sessions since it accommodates strongly unbalanced data with greater ease (Raudenbush, 2001).

Our findings showed very strong correlations across sessions for all three factors (between approximately .45 and .80, with most being between .60 and .80). This may be partly accounted for by the wording of the items in the WAI, from which the items were taken. Specifically, the WAI asks the patient about his/her *general* experience of the alliance with the therapist, e.g. "I feel /my therapist/ appreciates me" or "/My therapist/ and I agree on what is important for me to work on". This makes sense in a context when the alliance is measured once early in treatment, because then the researcher is interested in the emerging alliance not just in the present session, but throughout the early phase of treatment. When measuring the alliance each session, however, the researcher is interested in the alliance in the present session only. It is likely that if the wordings of items are changed, correlations over time will be smaller. Although large correlations across sessions is not a problem per se, they may lead to attenuated within-patient variation in alliance scores over time, resulting in reduced power to find significant relationships with other time-varying variables.

Because of this, we have re-formulated the items in 6-item alliance measure to show clearly to the patients that the researchers are only interested in the patient's experience of the present session. The resulting instrument is shown in Appendix B[6]. Except for re-formulating the items to focus on the present session only, there are two other changes from the WAI-SR (and other versions of the WAI):

1) Especially when focusing on the present session only, the frequency-based response scale in the original WAI and WAI-SR (e.g. ranging from “seldom” or “never” to “always”) becomes awkward. We therefore decided on another response scale, focusing on the extent to which the statement of the item is true, ranging from 0 = “Not at all” to 5 = “Completely”. This scale consists of six levels, in contrast to the original seven-level scale which has been shown to be problematic because most patients do not endorse the lowest levels (Hatcher & Gillaspy, 2006).

2) The original design with a blank line in which the patient is supposed to mentally insert the therapist’s name (e.g. “I feel _____ appreciates me”) was abandoned. Although this design may prime the patient to think about his/her therapist’s name, thus possibly making the response more personal, we believe that especially when the alliance measure is filled out each session this format becomes tedious.

Finally, we wanted to find a name of the scale that simultaneously reflected both its affinity with the WAI, from which the items were taken, as well as its differences from this measure (Goal and Task dimensions collapsed into one, focus on the current session). The name chosen was “Session Alliance Inventory” (SAI). Separate psychometric analyses will be carried out on this measure, since it has been slightly changed from the version tested in this paper.

An important advantage of longitudinal measurement is the possibility of disaggregating stable person-level attributes from time-specific deviations from these (e.g. Curran & Bauer, 2011). This means that some response-set biases that are stable over time but vary among patients, such as social desirability responding and acquiescence (Podsakoff, MacKenzie, Lee, &

Podsakoff, 2003) can be separated from fluctuations in alliance quality from one session to another within the same therapy. The time-specific deviations may be markers of episodes of rupture-and-repair of the alliance (Safran & Muran, 2000), processes that are increasingly recognized as important for therapists to attend to. A possibility when using very brief questionnaires repeatedly over time is that respondents may remember item responses from previous sessions. It might be argued that this would create ”memory effects”, where some respondents routinely mark the same responses as they did last time without taking time to reflect on the present session. Although this is entirely possible, it is not obvious that remembering would necessarily lead to more bias. For example, for some participants the memory of previous responses may be used as a comparison when deciding upon the rating for today’s session –leading to more precise responses. We also believe that the emphasis in the instructions on responding with today’s session only in mind will diminish the risk for routinely using last session’s response instead of reflecting on the present session.

The SAI is a promising alternative for researchers who want a brief instrument for getting a global working alliance score for each session. We know of two other “ultra-brief” alliance inventories, the Agnew Relationship Measure - 5 (ARM-5; Cahill et al., 2012) and the Session Rating Scale (Duncan et al., 2003). Like the SAI, both of these are very brief (5 and 4 items, respectively) and focus on the current session, although the Session Rating Scale is intended more as a clinical tool than as a research instrument. Compared to the ARM-5, the SAI is more closely linked to Bordin’s alliance theory and to the Working Alliance Inventory, the most widely used alliance measure. The two instruments were also developed in slightly different ways; while the SAI was developed using Confirmatory Factor Analysis, the ARM-5 was developed using Principal Component and Rasch analyses. For the ARM-5 parallel patient- and therapist forms were developed in the same analyses, which may be an advantage if researchers are interested in conducting dyadic analyses (e.g. studying congruence between patient- and therapist ratings) but possibly a disadvantage if patients and therapists interpret the same items differently (in that case a common factor structure would represent a compromise between two distinct structures).

The present study has several strengths; most prominently the use of three relatively large samples and replication of the same factor structure in two different language versions (Swedish and English). The use of advanced statistical modeling for testing longitudinal measurement invariance and dimensionality, issues that are often overlooked in clinical psychology, are also strengths. Limitations include the strongly unbalanced dataset used for measurement invariance tests and symptom prediction session-by-session, which may strain the missing data algorithms with unknown consequences (although nothing in our results indicate any problems). The use of Bayesian estimation with informative priors offers a lot of flexibility, which can be seen both as an advantage (models can be tested that are impossible to test using regular methods) and a disadvantage (e.g. the increased flexibility means that there are more possibilities to consider, with more things that can go wrong and with results sometimes being more difficult to interpret). In addition, the study is limited by the fact that patients filled out the full WAI-SR, and the six SAI items were taken from this context rather than patients responding to the SAI items only. It is possible that patients' responses differ when responding to the full WAI-SR compared to when only SAI items are used.

References

Ando, T. (2011). Predictive Bayesian Model Selection. *American Journal of Mathematical and Management Sciences*, *31*(1-2), 13–38. doi:10.1080/01966324.2011.10737798

Asparouhov, T., & Muthén, B. O. (2010). Bayesian Analysis Using Mplus : Technical Implementation, 1–38.

Bollen, K. a., & Curran, P. J. (2004). Autoregressive Latent Trajectory (ALT) Models: A Synthesis of Two Traditions. *Sociological Methods & Research*, *32*(3), 336–383. doi:10.1177/0049124103260222

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice*, *16*(3), 252–260.

Cahill, J., Stiles, W. B., Barkham, M., Hardy, G. E., Stone, G., Agnew-Davies, R., & Unsworth, G. (2012). Two short forms of the Agnew Relationship Measure: the ARM-5 and ARM-12. *Psychotherapy Research : Journal of the Society for Psychotherapy Research*, *22*(3), 241–55. doi:10.1080/10503307.2011.643253

Crits-Christoph, P., Gibbons, M. B. C., Hamilton, J., Ring-Kurtz, S., & Gallop, R. (2011). The dependability of alliance assessments: The alliance-outcome correlation is larger than you might think. *Journal of Consulting and Clinical Psychology*, *79*(3), 267–278. doi:http://dx.doi.org/10.1037/a0023668

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, *62*(1), 583–619. doi:doi:10.1146/annurev.psych.093008.100356

Curran, P. J., & Bollen, K. A. (2001). The best of both worlds: Combining autoregressive and latent curve models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change.* (pp. 107–135). Washington, DC US: American Psychological Association. doi:10.1037/10409-004

Curran, P. J., Lee, T., Howard, A. L., Lane, S., & MacCallum, R. (2012). Disaggregating within-person and between-person effects in multilevel and structural equation growth models. In J. R. Harring & G. R. Hancock (Eds.), *Advances in Longitudinal Methods in the Social and Behavioral Sciences* (pp. 217–253). Charlotte, NC: Information Age Publishing.

Duncan, B. L., Miller, S. D., Sparks, J. A., Claud, D. A., Beach, P., Reynolds, L. R., & Johnson, L. D. (2003). The Session Rating Scale: Preliminary Psychometric Properties of a “Working” Alliance Measure. *Journal of Brief Therapy*, *3*(1), 3–12.

Elfström, M. L., Evans, C., Lundgren, J., Johansson, B., Hakeberg, M., & Carlsson, S. G. (2012). Validation of the Swedish Version of the Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM). *Clinical Psychology & Psychotherapy*.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*(1), 1–16. doi:10.1037/a0022640 10.1037/a0022640.supp (Supplemental)

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE--OM. *British Journal of Psychiatry*, *180*(1), 51–60. doi:http://dx.doi.org/10.1192/bjp.180.1.51

Falkenström, F., Granström, F., & Holmqvist, R. (2013). Therapeutic alliance predicts symptomatic improvement session by session. *Journal of Counseling Psychology*, *60*(3), 317–328. doi:10.1037/a0032258

Falkenström, F., Granström, F., & Holmqvist, R. (2014). Working alliance predicts psychotherapy outcome even while controlling for prior symptom improvement. *Psychotherapy Research*, *24*(2), 146–59. doi:10.1080/10503307.2013.847985

Falkenström, F., Hatcher, R. L., & Holmqvist, R. (n.d.). *Confirmatory Factor Analysis of the Patient Version of the Working Alliance Inventory - Short Form Revised*.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). CRC press.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science : The Official Journal of the Society for Prevention Research*, *8*(3), 206–13. doi:10.1007/s11121-007-0070-9

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice*, *9*(3), 329–343. doi:10.1093/clipsy/9.3.329

Hatcher, R. L. (2010). Alliance theory and measurement. In J. C. Muran & J. P. Barber (Eds.), *The therapeutic alliance: An evidence-based guide to practice* (pp. 7–28).

Hatcher, R. L., & Gillaspy, J. A. (2006). Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy Research*, *16*(1), 12–25. doi:10.1080/10503300500352500

Hoffart, A., Øktedalen, T., Langkaas, T. F., & Wampold, B. E. (2013). Alliance and outcome in varying imagery procedures for PTSD: a study of within-person processes. *Journal of Counseling Psychology*, *60*(4), 471–82. doi:10.1037/a0033604

Holmqvist, R., Ström, T., & Foldemo, A. (2014). The effects of psychological treatment in primary care in Sweden – a practice-based study. *Nordic Journal of Psychiatry*, *68*(3), 204–12. doi:10.3109/08039488.2013.797023

Horvath, A. O., Del Re, A., Fluckiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy*, *48*(1), 9–16. doi:http://dx.doi.org/10.1037/a0022186

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology*, *36*(2), 223–233.

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equations Modeling* (pp. 650–673). The Guilford Press.

Kline, R. B. (2011). *Principles and practice of structural equation modeling (3rd ed.)*. *Principles and practice of structural equation modeling (3rd ed.).* New York, NY US: Guilford Press.

Little, T. (2013). *Longitudinal Structural Equations Modeling*.

Lucke, J. F. (2005). The α and the ω of Congeneric Test Theory: An Extension of Reliability and Internal Consistency to Heterogeneous Tests. *Applied Psychological Measurement*, *29*(1), 65–81. doi:10.1177/0146621604270882

Munder, T., Wilmers, F., Leonhart, R., Linster, H. W., & Barth, J. (2010). Working Alliance Inventory-Short Revised (WAI-SR): Psychometric properties in outpatients and inpatients. *Clinical Psychology & Psychotherapy*, *17*(3), 231–239.

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. doi:10.1037/a0026802

Muthén, B. O., & Asparouhov, T. (2013). *BSEM Measurement Invariance Analysis*.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide.* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *The Journal of Applied Psychology*, *88*(5), 879–903. doi:10.1037/0021-9010.88.5.879

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata. Volume I: Continuous responses* (3rd ed.). Texas: Stata Press Publication.

Raudenbush, S. W. (2001). Toward a coherent framework for comparing trajectories of individual change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change.* (pp. 35–64). Washington, DC US: American Psychological Association. doi:10.1037/10409-002

Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, *47*(5), 667–696. doi:10.1080/00273171.2012.715555

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2012). Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective. *Educational and Psychological Measurement*, *73*(1), 5–26. doi:10.1177/0013164412449831

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Safran, J. D., & Muran, J. C. (2000). *Negotiating the therapeutic alliance: A relational treatment guide* (p. Negotiating the therapeutic alliance: A relational). New York, NY: Guilford Press; US.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. Retrieved from http://projecteuclid.org/euclid.aos/1176344136

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. doi:10.1111/1467-9868.00353

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(1), 1–12. doi:10.1027/1614-2241/a000049

Stiles, W. B., Barkham, M., Mellor-Clark, J., & Connell, J. (2008). Effectiveness of cognitive-behavioural, person-centred, and psychodynamic therapies in UK primary-care routine practice: Replication in a larger sample. *Psychological Medicine*, *38*(5), 677–688. doi:10.1017/s0033291707001511

Tasca, G. A., & Lampard, A. M. (2012). Reciprocal Influence of Alliance to the Group and Outcome in Day Treatment for Eating Disorders. *Journal of Counseling Psychology*. doi:10.1037/a0029947

Tracey, T. J., & Kokotovic, A. M. (1989). Factor structure of the Working Alliance Inventory. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *1*(3), 207–210. doi:http://dx.doi.org/10.1037/1040-3590.1.3.207

Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*(Ml), 770. doi:10.3389/fpsyg.2013.00770

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. doi:10.1080/17405629.2012.686740

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–69. doi:10.1177/109442810031002

Vanheule, S., Desmet, M., Groenvynck, H., Rosseel, Y., & Fontaine, J. (2008). The factor structure of the Beck Depression Inventory-II: an evaluation. *Assessment*, *15*(2), 177–87. doi:10.1177/1073191107311261

Zyphur, M. J., & Oswald, F. L. (2013). Bayesian Estimation and Inference: A User's Guide. *Journal of Management*, (August). doi:10.1177/0149206313501200

Appendix A: Bayesian estimation

Bayesian statistics differs from the most commonly used Frequentist statistics primarily in two ways; 1) parameters are not considered fixed but are seen as random variables with distributions, and 2) prior information is used directly in model estimation, and the result of the analysis (called the posterior distribution due to estimates taking the form of distributions rather than point estimates) is the product of the prior information and the likelihood obtained for the data (e.g. Zyphur & Oswald, 2013). In addition, estimation of Bayesian models is usually done using simulation-based methods, called Markov Chain Monte Carlo (MCMC) estimation. Briefly, MCMC simulates samples from the posterior distribution, given the model and the data. This is done in a series of steps in which each step depends on the results of the previous one. Given a long enough chain, this procedure converges on the most likely parameter estimates. Usually more than one chain is run, in order to enable testing if the chains converge on similar distributions. In the present study two chains were used in all analyses.

The estimates from the simulated parameters of the Markov Chain(s) constitute the posterior distribution, although usually the first part of the chains are “burnt” because of the need to get rid of the influence of arbitrary starting values. The posterior distribution can be summarized in ways that may make the results of a Bayesian analysis look similar to a Frequentist analysis, for example using the mean and the 2.5-97.5% percentiles – a Bayesian version of the confidence interval that is usually termed the 95% credibility interval. This credibility interval can be interpreted straightforwardly as the probability that the parameter is within a certain interval, something that cannot be done with the Frequentist confidence interval because this interval is based on the idea of a large number of replications of a study and does not directly speak to the probability of a certain estimate (Kaplan & Depaoli, 2012).

Perhaps the most controversial aspect of Bayesian statistics is the use of informative priors, because of the subjectivity inherent in the choice of values for the prior. We would argue, along with for example B. Muthén and Asparouhov (2012), that subjectivity is used in regular SEM as well in the form of choices of model constraints. For example, when setting up a Confirmatory Factor Analysis, a model is used that specifies some paths that should be estimated freely while other paths are constrained to zero. The constraints used in regular CFA can be seen as particularly strongly informative priors that impose exactly zero correlations for certain paths. In regular CFA, the plausibility of these constraints is evaluated using model fit testing. The same can be done for the priors in Bayesian estimation. In a "truly" Bayesian approach, this might not be seen as necessary, since a compromise between the prior information and the data at hand is desired (resulting in an "update" of the prior information), but in the more pragmatic approach taken here model fit evaluation becomes important as a way of ensuring that the priors do not distort parameter estimates beyond the information inherent in the data. In sum, we believe that informative priors should be uncontroversial if 1) the variances of the priors are large enough so that these do not influence the posterior distribution unduly, and 2) close attention is given to model fit of the estimated model. The second point is especially important, because if the prior unduly influences estimates this will show up as a model with poor fit to data.

The absolute fit of Bayesian models can be tested using posterior predictive checking (Gelman et al., 2014). Posterior predictive checking is based on the idea that if future samples are simulated from the posterior distribution, these samples should be roughly similar to the observed data. The posterior predictive test value used in the current study was the probability that the discrepancy (i.e. Chi-square test value) between the model predicted and observed covariance matrices is smaller than the discrepancy between model predicted and simulated future samples' covariance matrices (Asparouhov & Muthén, 2010). This implies that a small value of the posterior predictive *p*-value indicates bad model fit, while a value close to .50 (i.e. 50/50 probability for observed and simulated data) indicates good fit. Relative model fit can be compared using the Deviance Information Criterion (Spiegelhalter, Best, Carlin, & van der Linde, 2002) and the Bayesian Information Criterion (Schwarz, 1978).

Appendix B: Session Alliance Inventory (patient version)

Below are sentences that describe some of the ways a person might think or feel about his or her therapist. When you read these descriptions, think about your last session only. Below each statement there is a six-point scale:

0	1	2	3	4	5
Not at all	A little	Moderately	Quite a bit	Very much	Completely

If the statement describes the way you felt or thought throughout your last session, circle the number 5; if the sentence describes a thought or feeling that did not occur at all to you circle the number 0. Use the numbers 1 to 4 to describe the variations between these extremes. You will be asked to complete this report a number of times during your therapy. Remember, each time we are interested in your impressions from the last session only. Work fast; your first impressions are the ones we would like to see. Please don’t forget to respond to every item. Thank you for your cooperation.

1. My therapist and I were working towards mutually agreed upon goals.

0	1	2	3	4	5
Not at all	A little	Moderately	Quite a bit	Very much	Completely

2. I felt that my therapist appreciated me.

0	1	2	3	4	5
Not at all	A little	Moderately	Quite a bit	Very much	Completely

3. My therapist and I respected each other.

0	1	2	3	4	5
Not at all	A little	Moderately	Quite a bit	Very much	Completely

4. We were in agreement on what is important for me to work on.

0	1	2	3	4	5
Not at all	A little	Moderately	Quite a bit	Very much	Completely

5. I felt that my therapist cared about me even if I had done things that he/she does not approve of.

0	1	2	3	4	5
Not at all	A little	Moderately	Quite a bit	Very much	Completely

6. I believe the way we were working with my problem(s) was correct.

0	1	2	3	4	5
Not at all	A little	Moderately	Quite a bit	Very much	Completely

Footnotes

[1] The PUC was calculated as follows (see Reise, 2012): with 6 items, there are (6×5)/2 = 15 unique correlations. Within each subscale, correlations are “contaminated” by both general and group variance, and there are [3×2)/2]×2 = 6 such correlations. The rest, i.e. 15-6 = 9, are “uncontaminated” correlations. The PUC is then calculated as the ratio of uncontaminated correlations to the total number of correlations, that is 9/15 = 0.6.

[2] The effective sample size is the number of patients included in the analysis. Since all available data was used, this is equal to the total number of patients who provided at least one WAI questionnaire at any session.

[3] Additional support for this can be seen in Table 5, where the two models with BSEM residual correlations have the lowest DIC values of all models. With 1770 residual correlations estimated, most of which were small and statistically non-significant, these models are clearly not parsimonious.

[4] Specifically, Rabe-Hesketh and Skrondal (2012) point out that the residuals are likely to be correlated between the lagged variable and the original (i.e. un-lagged) variable, thus violating one of the basic assumptions of regression analysis that is no correlation between errors of independent and dependent variables. This problem is avoided when using Structural Equations Modeling, because each occasion is treated as a separate variable.

[5] Because the Chi-square value for the robust Maximum Likelihood estimator can’t be used directly for Chi-square difference testing, the Satorra-Bentler scaled Chi-square difference test was used.

[6] A version in Swedish can be obtained from the authors upon request.

Table 1.

Descriptive statistics for the 6-item Working Alliance scale in Session 3 in the three samples.

Sample 1	N	Mean	SD	Min	Max	Skew	Kurtosis
Item 5	630	6.37	0.92	1	7	-1.66	3.74
Item 6	629	5.85	1.19	2	7	-0.95	0.46
Item 7	622	5.92	1.13	1	7	-0.99	0.86
Item 8	632	5.92	1.11	1	7	-0.97	0.75
Item 9	616	5.84	1.16	1	7	-0.90	0.44
Item 12	634	5.81	1.18	2	7	-.84	0.16
Sample 2							
Item 5	234	6.42	0.92	3	7	-1.68	2.52
Item 6	233	5.86	1.27	1	7	-1.22	1.22
Item 7	230	5.56	1.30	1	7	-0.88	0.42
Item 8	235	5.98	1.07	1	7	-1.23	2.13
Item 9	234	5.93	1.20	1	7	-1.21	1.50
Item 12	234	5.79	1.12	2	7	-0.85	0.27
Sample 3							
Item 5	234	6.32	0.99	1	7	-1.80	4.53
Item 6	227	5.89	1.24	1	7	-1.32	2.01
Item 7	230	5.77	1.26	1	7	-0.98	0.62
Item 8	228	5.79	1.20	1	7	-1.19	1.72
Item 9	230	5.73	1.26	1	7	-1.04	1.18
Item 12	231	5.72	1.26	1	7	-0.96	0.53

Table 2.

Model fit information for Confirmatory Factor Analyses of the 6-item Working Alliance scale, measured at Session 3 in 3 independent samples.

Sample 1 (*N*=635)	Chi^2 95% CI	PP p^1	DIC	BIC
One factor	231.8, 285.6	<.001	9008.7	9089.9
Two correlated factors	31.2, 78.1	<.001	8789.2	8872.3
Bifactor orthogonal groups	3.4, 43.9	.01	8758.0	8883.8
Bifactor correlated groups	-15.1, 25.4	.30	8725.6	8870.8
Sample 2 (*N*=235)				
One factor	29.6, 69.3	<.001	3707.7	3771.9
Two correlated factors	24.7, 65.7	<.001	3702.9	3773.3
Bifactor orthogonal groups	-22.0, 18.0	.55	3657.5	3759.2
Bifactor correlated groups	-22.0, 17.0	.59	3655.0	3764.2
Sample 3 (*N*=234)				
One factor	20.0, 61.5	<.001	3527.1	3591.0
Two correlated factors	14.8, 57.6	<.01	3521.6	3591.6
Bifactor orthogonal groups	-18.7, 23.0	.39	3489.6	3591.2
Bifactor correlated groups	-21.2, 18.5	.54	3478.8	3595.0

Note.

[1] Posterior Predictive *p*-value for the Chi-square test of model fit. Low values indicate poor fit, while values close to .50 indicate good fit.

Table 3.

Model fit information and standardized factor loadings for bifactor model with correlated group factors estimated separately for Sessions 1-10 in Sample 1.

				Standardized factor loading for item:											
				General alliance factor						Goal/Task factor			Bond factor		
Session	N[1]	Chi^2 95% CI	PP *p*[2]	5	6	7	8	9	12	6	8	12	5	7	9
1	1006	-19.7, 21.1	.47	.65	.78	.63	.86	.65	.65	.76	1.00	.82	.45	.84	.96
2	778	-21.6, 19.1	.54	.64	.73	.54	.94	.70	.60	.59	.82	.83	.36	.68	1.00
3	635	-15.1, 25.4	.30	.67	.76	.59	.86	.67	.74	.56	.78	.64	.32	.63	.95
4	507	-18.8, 20.2	.46	.66	.67	.68	.78	.73	.72	.56	.57	.64	.36	.66	.66
5	388	-17.5, 22.7	.39	.64	.78	.78	.75	.64	.74	.62	.73	.63	.32	.50	.47
6	299	-14.3, 27.4	.25	.59	.67	.77	.87	.75	.66	.60	.66	.74	.44	.74	.62
7	249	-19.8, 20.1	.47	.67	.67	.67	.83	.76	.73	.52	.59	.54	.30	.69	.72
8	192	-19.3, 23.0	.41	.68	.70	.80	.68	.60	.62	.72	.89	.60	.27	.51	.28
9	142	-16.2, 24.2	.34	.62	.73	.70	.69	.69	.72	.65	.75.	.46	.31	.47	.56
10	120	-18.1, 22.8	.41	.63	.79	.69	.77	.70	.74	.46	.56	.35	.44	.39	.50

Note.

[1] The total sample size was 1095, but some patients who didn't provide data for the first session did so for later sessions.

[2] Posterior Predictive *p*-value for the Chi-square test of model fit. Low values indicate poor fit, while values close to .50 indicate good fit.

Table 4.

Estimated intercepts and factor means for bifactor model estimated separately for Sessions 1-10 in Sample 1.

Session	Item 5	Item 6	Item 7	Item 8	Item 9	Item 12	General mean	Goal/Task mean	Bond mean
1	6.10	5.59	5.38	5.71	5.30	5.51	0.04	-0.22	0.34
2	6.06	5.32	5.44	5.50	5.42	5.44	0.25	0.08	0.15
3	6.02	5.42	5.47	5.56	5.44	5.44	0.45	0.18	0.25
4	6.00	5.49	5.61	5.54	5.60	5.47	0.44	0.24	0.16
5	6.01	5.40	5.59	5.49	5.57	5.36	0.39	0.51	0.33
6	5.98	5.52	5.64	5.59	5.55	5.59	0.50	0.08	0.29
7	6.00	5.46	5.62	5.50	5.58	5.51	0.52	0.22	0.32
8	5.96	5.52	5.69	5.58	5.69	5.58	0.53	0.26	0.34
9	6.00	5.52	5.69	5.46	5.58	5.58	0.50	0.40	0.34
10	5.90	5.46	5.65	5.44	5.68	5.41	0.50	0.41	0.39

Table 5.

Model fit indices for longitudinal CFA models, Sessions 1-10 in Sample 1.

	Chi^2 95% CI	PP p[1]	DIC	BIC
Configural model	4.96, 360.0	.02	53587.0	58183.1
Weak Factorial Invariance (Metric Invariance)	48.57, 407.61	<.01	53674.2	57671.9
Strong Factorial Invariance (Scalar Invariance)	62.89, 417.03	<.01	53680.2	57326.3
Configural model with BSEM residuals	-152.07, 200.69	.39	53402.7	67850.9
Strong Factorial Invariance with BSEM residuals	-142.96, 202.46	.37	53402.9	66793.6
BSEM Approximate Invariance, .01[2] prior	9.04, 361.16	.02	53581.4	58187.3
BSEM Approximate Invariance, .05[2] prior	6.72, 361.13	.02	53579.5	58160.0
Partial Invariance[3]	30.13, 384.35	.01	53580.4	57307.5

Note.

[1] Posterior Predictive *p*-value for the Chi-square test of model fit. Low values indicate poor fit, while values close to .50 indicate good fit.

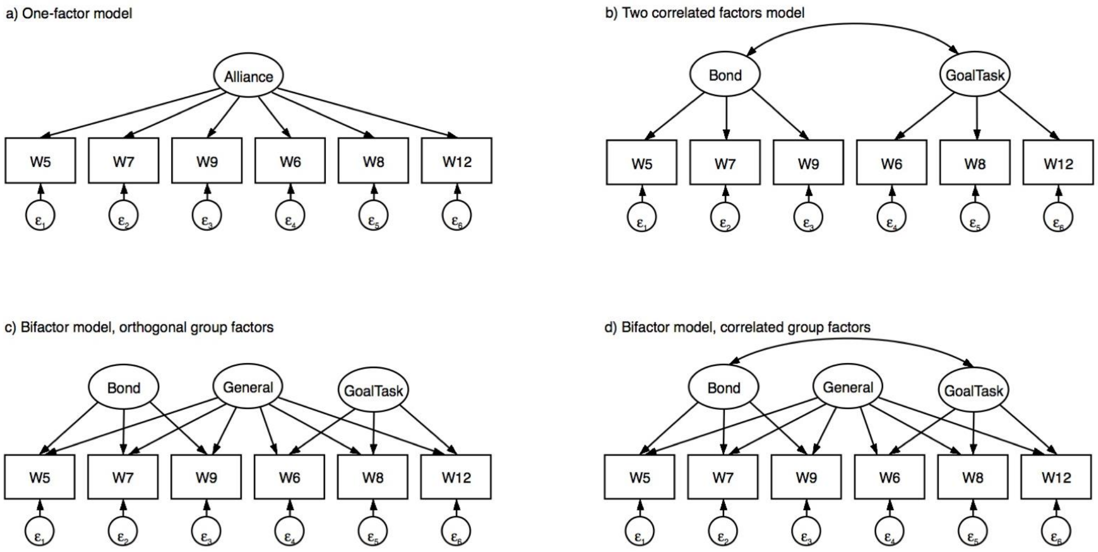
[2] Prior variance for the difference between parameter estimates (factor loadings and intercepts) at different sessions.

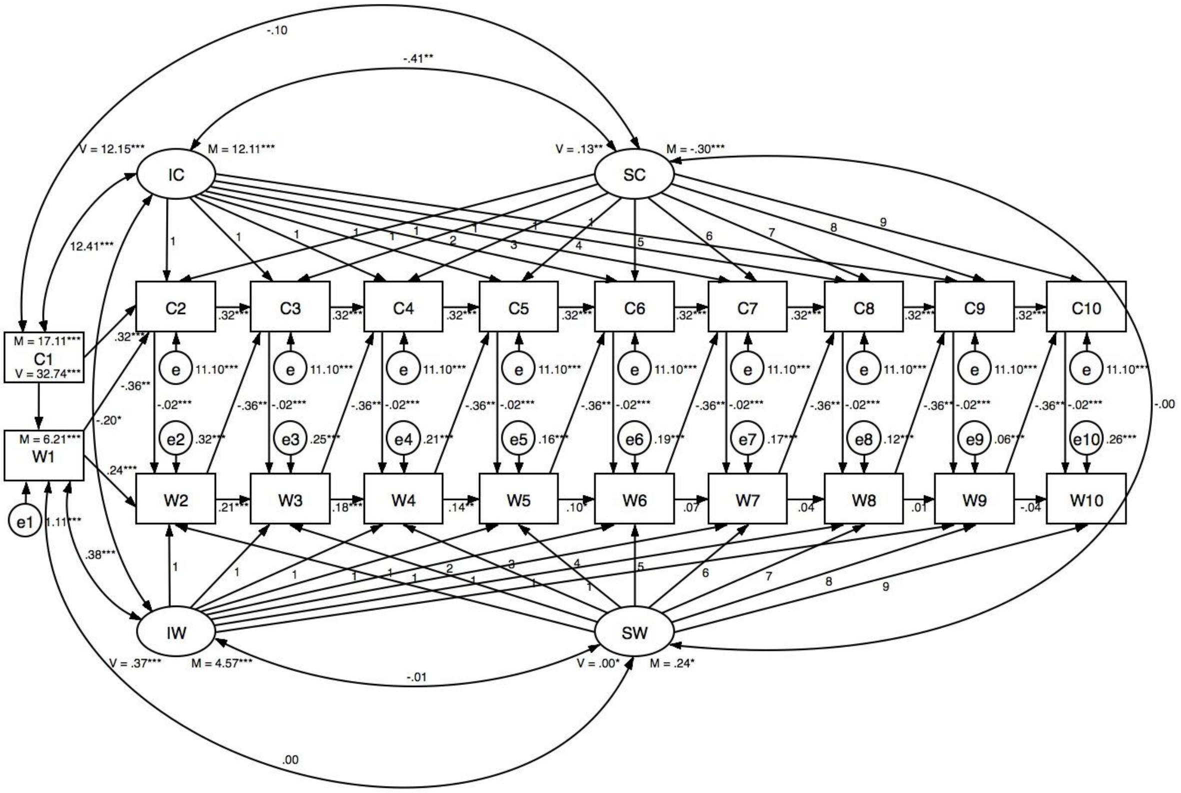
[3] All non-invariant loadings from BSEM invariance models freed, all other loadings and intercepts held exactly equal across sessions.

Figure Captions

Figure 1. Path diagrams for one-, two-, and bifactor models.

Figure 2. Bivariate Autoregressive Latent Trajectory Model.


a) One-factor model
Alliance
W5
W7
W9
W6
W8
W12
ε1
ε2
ε3
ε4
ε5
ε6
b) Two correlated factors model
Bond
GoalTask
W5
W7
W9
W6
W8
W12
ε1
ε2
ε3
ε4
ε5
ε6
c) Bifactor model, orthogonal group factors
Bond
General
GoalTask
W5
W7
W9
W6
W8
W12
ε1
ε2
ε3
ε4
ε5
ε6
d) Bifactor model, correlated group factors
Bond
General
GoalTask
W5
W7
W9
W6
W8
W12
ε1
ε2
ε3
ε4
ε5
ε6



Note. C1-C10 are observed CORE-OM scores for Sessions 1-10, and W1-W10 are means of the 6-item Working Alliance scale. IC and IW are the random intercepts for CORE-OM and Working Alliance, respectively, and SC and SW are the random slopes for the linear effect of Sessions for CORE-OM and Working Alliance.