Working Alliance Predicts Symptomatic Improvement in Public Hospital-Delivered Psychotherapy in Nairobi, Kenya

Fredrik Falkenström, Mary Kuria, Caleb Othieno and Manasi Kumar

The self-archived postprint version of this journal article is available at Linköping University Institutional Repository (DiVA):
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-153648

N.B.: When citing this work, cite the original publication.

Original publication available at:
https://doi.org/10.1037/ccp0000363

Copyright: American Psychological Association
http://www.apa.org/
Working Alliance Predicts Symptomatic Improvement in Public Hospital Delivered Psychotherapy in Nairobi, Kenya

Fredrik Falkenström¹, Mary Kuria², Caleb Othieno², Manasi Kumar²

¹ Department of Behavioral Sciences and Learning, Linköping University, Sweden
² Department of Psychiatry, University of Nairobi, Kenya

Author Note
The authors would like to thank the therapists and patients at Kenyatta National Hospital and Mathare National Teaching and referral Hospital who made this study possible.

The project was supported in 2015-2016 by MEPI/Prime-K seed grant covered under award 1R24TW008889 from the US National Institutes of Health to MK, and by Sörmland County Council grant DLL-514111 in 2015 and DLL-569981 in 2016 to FF.

Corresponding author: Fredrik Falkenström, Department of Behavioral Sciences and Learning, Linköping University, SE-581 83 Linköping, email: fredrik.falkenstrom@liu.se.
Abstract

Objective: Although the working alliance has been studied extensively in Europe and America, it is unknown to what extent the importance of working alliance for psychotherapy outcome generalizes to Lower- and Middle-income countries. Additionally, there is a need for more studies on the alliance using methods that are robust to confounders of its effect on outcome.

Method: Three hundred and forty-five outpatients seeking care at the two public psychiatric hospitals in Nairobi, Kenya filled out the Session Alliance Inventory and the Clinical Outcomes in Routine Evaluation – Outcome Measure each session. The effect of alliance on next-session psychological distress was modeled using the Random Intercept Cross-Lagged Panel Model, which estimates a cross-lagged panel model on within- and between patient disaggregated data.

Results: Changes in the working alliance from session to session significantly predicted change in psychological distress by the next session, with an increase of one point of the SAI in a given session resulting in a decrease of 1.27 points on the CORE-OM by the next session (se = 0.60, 95% CI -2.44, -0.10). This finding represents a medium-sized standardized regression coefficient of between .16 and .24. Results were generally robust to sensitivity tests for stationarity, missing data assumptions, and measurement error.

Conclusion: Results affirm cross-cultural stability of the session-by-session reciprocal effects model of alliance and psychological distress/symptoms as seen in a Kenyan psychiatric outpatient sample, using the latest developments in cross-lagged panel modeling. A limitation of the study is its naturalistic design and lack of control over several variables.

Keywords: Working alliance, outcome, cross-cultural psychology, cross-lagged panel model
Public Health Statement: This study shows that the quality of patient-therapist collaboration (working alliance) in a given session contributes to the improvement of symptomatic distress during the week(s) following that session, for patients treated with talking therapy in two outpatient psychiatric hospital samples in Nairobi, Kenya. Our work can be seen as a cross-cultural replication of prior findings on the alliance-outcome relationship in psychotherapy from European and North American contexts. It also offers some insights about challenges in the delivery of mental health services in resource-constrained African contexts.
Working Alliance Predicts Symptomatic Improvement in Public Hospital Delivered Psychotherapy in Nairobi, Kenya

The working alliance, defined as patient-therapist agreement on goals and tasks of therapy in the context of a positive emotional bond (Bordin, 1979) is the most extensively researched psychotherapy process (Crits-Christoph, Connolly Gibbons, & Mukherjee, 2013). However, amongst the almost 300 studies included in the most recent meta-analysis of the alliance-outcome correlation (Flückiger, Del Re, Wampold, & Horvath, 2018), all were conducted either in North America, Europe, or Australia, i.e. in Western higher-income contexts. There is a paucity of evidence on alliance-outcome relationship from non-Western settings. In particular, studies on working alliance in Sub-Saharan Africa, where psychotherapies have been delivered since the last three to four decades, remain unknown. There is a need for cross-cultural validation of the working alliance concept to better understand the mechanisms of psychotherapies offered in Sub-Saharan African contexts, where huge technological and demographic changes are taking place. Here lies a glaring knowledge gap in the evaluation of cross-cultural psychotherapy practice.

Psychotherapy in Kenya

Psychotherapy in Kenya is offered by counselors (2 years of training at Diploma level), counseling psychologists (3-year training at degree level), clinical psychologists (3-year training at Masters level) or psychiatrists (medical doctors with 3 years postgraduate training in psychiatry). The total number of the above cadres of therapists is less than 400 for a population of 47 million, and the majority of these professionals work in private practice due to better remuneration than in public health contexts. Therapeutic models in Kenya are often eclectic, influenced by Western psychological models, pastoral counselling, and HIV counselling.
programs. Sociocultural and economic factors, including the availability of training, acceptability of appointments and time boundaries, transport availability, and treatment costs, modify patients’ and therapists’ behaviors to create a contextualized, pragmatic style of therapy, e.g. giving advice on how to retain jobs, help with food aid or antiretrovirals etc. or provide relaxation therapy in a very busy clinic where privacy is limited.

**Moderators of the working alliance effect on outcome**

As conceptualized by for example Bordin (1979), the working alliance should be universally important for psychotherapy outcome. However, different aspects of the alliance might be important in different therapy forms, and the alliance might impact outcome in different ways for different patients. A few studies have looked at moderators of the alliance effect (e.g. Falkenström, Ekeblad, & Holmqvist, 2016; Falkenström, Granström, & Holmqvist, 2013; Flückiger et al., 2018; Lorenzo-Luaces, DeRubeis, & Webb, 2014; Zilcha-Mano & Errázuriz, 2015), but there is a need for further studies in this regard.

Culture is a potential moderator that has so far received little research attention. For a long time, cultural and psychological anthropologists have been questioning the intercultural application of psychotherapies developed in Western contexts (Littlewood, 2001), and this critical lens is important to ensure that psychotherapy is meaningful to people of different cultures and offered in ways that are culturally congruent. As material and social conditions in different cultural contexts vary we cannot take the universality of alliance in psychotherapy or mental health treatments for granted.

**Causality in alliance-outcome research**

Most of the existing studies relating alliance to outcome are simple pre-post correlational designs, which are relatively weak in terms of causal inference (Shadish, Cook, & Campbell,
2002). Stronger causal interpretations can be made from cross-lagged panel designs, especially if they use disaggregation of within- and between person effects (Falkenström, Finkel, Sandell, Rubel, & Holmqvist, 2017). These designs control for stable between-person differences and reverse causation, which are major threats to the validity of pre-post correlation designs. Although several studies have used such designs to study the prediction of outcome from alliance (see Zilcha-Mano, 2017, for a review), this research is still in its infancy. In addition, even these studies are based on assumptions that are seldom tested (Falkenström, Finkel, et al., 2017). Many studies ignore the problem of endogeneity associated with simultaneously separating between-and within-person effects while also adjusting for the effect of the prior value of the outcome variable (Falkenström et al., 2016). In addition, assumptions regarding missing data, measurement error, and stationarity are seldom tested.

Purpose of present study

The primary purpose of the present study was to test whether findings regarding the prediction of next-session psychological distress from working alliance ratings found in Europe and America would hold up in a Kenyan sample. In a practice-based psychotherapy research project on the process and outcome of talking therapy in Nairobi, Kenya (Falkenström, Gee, Kuria, Othieno, & Kumar, 2017), we had already found that the participants were, on average, improving in their psychological functioning as they received mental health care. Like several previous studies (Falkenström et al., 2016; Falkenström et al., 2013; Marker, Comer, Abramova, & Kendall, 2013; Rubel, Rosenbaum, & Lutz, 2017; Zilcha-Mano, 2017; Zilcha-Mano & Errázuriz, 2015; Zilcha-Mano et al., 2016), we were primarily interested in within-patient effects (i.e. session-by-session predictions) since these are more likely to represent causal effects than between-patient effects. Improvement in alliance quality in a given session was hypothesized to
predict improvement in symptoms/distress by the next session, as in prior studies in Western contexts. To our knowledge, this is the first study of the alliance-outcome relationship carried out in a Sub-Saharan African public health context.

A secondary purpose of the study was to demonstrate cutting-edge statistical analysis of panel data, using longitudinal Structural Equation Modeling. Models for studying within-patient effects are getting increasingly popular among psychotherapy process-outcome researchers, but misconceptions abound and many studies use faulty methods and/or do not realize the full potential for these methods (e.g. Falkenström, Finkel, et al., 2017).

**Method**

**Setting**

This was an observational study on regular clinical treatments. Data was collected at two public hospitals, Mathare National Teaching and Referral Hospital (MNTRH) – the only psychiatric and mental health teaching and referral hospital in Kenya – and the outpatient psychiatric clinic, mental health department and youth clinic of Kenyatta National Hospital (KNH) – a 1500-bed national teaching and referral hospital, both in Nairobi, Kenya. KNH and MNTRH are teaching hospitals where trainees are posted for clinical experience. Both hospitals have several free services to patients exposed to gender-based violence, affected with HIV, combat veterans and their families, those who are unemployed or disabled and youth under the age of 24 years. Services for these groups are highly sought after and a large number of people with low socioeconomic status bank on these national hospitals for specialist care. These were the major factors making the patients in our study visit the two hospitals. The psychotherapy offered in the two hospitals is largely unstructured, and the patient may be attended to by a different therapist in different sessions. Most of the patients with mental health problems receive
some form of talking therapy from the attending specialist, but those that require intense psychotherapy are referred to the counseling or clinical psychologist. Most people with mental health problems prefer to be attended to by a counselor or psychologist due to stigma associated with consulting a psychiatrist.

**Procedure**

Three research assistants working part time in the project approached treatment-seeking patients assigned to any intervention described by clinicians as ‘counselling’ or ‘psychotherapy’ (the only inclusion criterion in this study), described the study and asked if they were willing to participate in the study by filling out questionnaires when coming for their sessions. Patients who gave written informed consent filled out the Clinical Outcomes in Routine Evaluation – Outcome Measure (Evans et al., 2002) before their sessions and the Session Alliance Inventory (Falkenström, Hatcher, Skjulsvik, Larsson, & Holmqvist, 2015) after sessions. The project was approved by the Kenyatta National Hospital/University of Nairobi Ethics and Research Committee (P85/02/2014). Data collection started in March 2015 and continued to June 2016.

**Participants**

Three hundred and forty-five participants were recruited by the research assistants. Very few patients declined participation (< 3%), although a few were unable to fill out questionnaires due to intoxication or psychotic states. The participants attended one of three KNH clinics; Youth clinic \((n = 140)\), Department of Mental Health \((n = 14)\) and Psychiatric Clinic \((n = 11)\) and Mathare Hospital \((n = 180)\). The participants’ ages ranged from 18 to 60 years old \((M = 28.9, SD = 9.8)\). Table 1 shows demographic information about the participants. The majority of the participants were male (72.6%), while only 27.4% of the sample were female. Diagnoses were established by the intake clinicians using the International Classification of Diseases, version 10.
(WHO, 2004). The most common disorders that patients were seeking treatment for were addictions (54.8%), psychosis (17.5%), depression (16.9%) and anxiety/stress (12.0%). Other identified problems that were less common included interpersonal problems, physical problems, work/academic problems, self-esteem problems or trauma/abuse. The patients were treated with a variety of therapies, for example, Cognitive Behavioral Therapy, Interpersonal Psychotherapy, Addiction Counseling, Supportive Therapy, Family Therapy, Psychoeducation and Brief Solution Focused Therapy.

**Therapists**

The therapists were either full-time mental health professionals employed by the hospitals to offer mental health services, or postgraduate interns from clinical psychology, psychiatry, nursing and counseling fields assisting in delivering clinical services. Therapists wanted to be anonymous, so there was no information on therapists except their gender and mental health professional/resident status. About two thirds of patients (67%) were treated by professional psychotherapists, and 70% of the sessions were conducted by male therapists.

**Measures**

English versions of self-report measures were used in the study. English is one of the two official languages of Kenya, and most people in Nairobi speak English well. Kiswahili versions were prepared but did not need to be used.

**Session Alliance Inventory** (SAI; Falkenström, Hatcher, Skjulsvik, et al., 2015) is a 6-item measure of working alliance, based on items taken from the Working Alliance Inventory (WAI; Hatcher & Gillaspy, 2006; Horvath & Greenberg, 1989). Similar to the WAI, the SAI contains items reflecting the three theoretical aspects of alliance; agreement on Goals and Tasks of treatment, and a positive emotional Bond. In contrast to the WAI, the SAI items do not reflect
the three dimensions equally (i.e. there are not equally many items for Goal, Task, and Bond), but half of the items are Bond items and half the items are Goal or Task items. Although this may seem to put undue emphasis on Bond compared to Goal and Task, it has been empirically shown that Goal and Task are indistinguishable while Bond constitutes a separate but correlated factor (e.g. Falkenström, Hatcher, & Holmqvist, 2015; Falkenström, Hatcher, Skjulsvik, et al., 2015). Still, this difference should be borne in mind when interpreting results based on the SAI.

In contrast to most alliance measures, the SAI asks for the patient’s experience of the alliance in the most recent session, while most other alliance measures ask for a more global experience of the alliance. This is based on the idea that the SAI should be used for repeated administrations during therapy. Another difference to the WAI is that the SAI is scored on a scale from 0 (“Not at all”) to 5 (“Completely”), rather than 1 (“Never”) to 7 (“Always”) as in the WAI. Initial psychometric evaluation of the SAI has been positive (Falkenström, Hatcher, Skjulsvik, et al., 2015). In particular, the SAI has shown measurement invariance (Falkenström, Hatcher, Skjulsvik, et al., 2015), i.e. stable factor structure across repeated administrations, which is crucial for a measure that is to be applied repeatedly over time. If measurement invariance does not hold, observed scores from different time-points cannot be compared, since the factor structure is different for the different measurements. The SAI is intended to be used as a unidimensional measure, i.e. only the total score should be used. The SAI total score showed good internal consistency ($\alpha = .88$).

**Clinical Outcomes in Routine Evaluation – Outcome Measure (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 0 (“Not at all”) to 4 (“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 CORE-OM has shown good internal and test-retest reliability (0.75-0.95), convergent validity, large differences between clinical and non-clinical samples, and good sensitivity to change. A factor analysis on the present sample showed that the CORE-OM has a strong general distress factor, and that the only subscale that added anything on top of that was the risk scale (Falkenström, Kumar, Sahid, Kuria, & Othieno, in press). Thus, only the total score was used, which had excellent internal consistency ($\alpha = .94$).

**CORE therapy assessment form (CORE-A; Evans et al., 2002).** The CORE- A consists of patient demographic information, identified problems/concerns, assessments of risk to self/others, etc. Since it was developed mainly as an information gathering form to be completed by practitioners, not as a psychometric measure, there are no conventional psychometric studies on it. A slightly shortened version was used, which can be obtained from the authors at request.

**Statistical analysis**

The data was analyzed using cross-lagged panel analysis, to be able to correctly model the time-structure of the data. The structure of the data was complicated due to the fact that a large number of patients attended only one or a few sessions, with N diminishing rapidly with each session. This structure is very similar to naturalistic data from other countries (e.g. Falkenström et al., 2013). There are several advantages of using Structural Equation Models for panel data, especially when the time-series dimension (T) is short but sample size (N) is fairly large. The most important advantage compared to alternative modeling strategies (e.g. multilevel modeling) is the possibility of separating within- and between person variances at the same time as including a lagged dependent variable as covariate without inducing endogeneity problems (e.g. Falkenström, Finkel, et al., 2017). Other advantages include the possibility of relaxing
stationarity assumptions, i.e. assumptions regarding the stability of the means, variances, and covariances across time. In addition, standard SEM model fit indices can be used to evaluate model fit. We decided to use the Random Intercept Cross-Lagged Panel model (RI-CLPM; Hamaker, Kuiper, & Grasman, 2015), since it is explicitly developed to separate within- from between-person variance in contrast to the traditional cross-lagged panel model and some other variations of this model that exist. This is done by estimating the cross-lagged regressions on latent within-person deviation scores, represented in the model by residuals from the random intercept model. This means that cross-lagged paths can be interpreted as the effects of changes over time in one variable on changes over time in the other variable. The random intercept captures average between-person differences, blocking potential confounding with effects of unobserved higher-level (e.g. patient-, therapist- or clinic) variables as long as their effects are constant over time. This strengthens the possibility for causal interpretation considerably compared to traditional statistical models. In the present study we used data from the first five sessions, since there was too sparse data in sessions six, seven and eight.

Initial model tests and modifications were done using Maximum Likelihood (ML) estimation with robust standard errors, with missing data handled by Full Information Maximum Likelihood (FIML). More complex models had to be estimated using Bayesian estimation, which performs better than ML in complex models with relatively sparse data (Gelman et al., 2014). Additional advantages of Bayesian estimation include that error variances (and other model parameters) are not assumed to be normally distributed, and that there is no risk for inadmissible estimates (e.g. negative error variances). For more information on Bayesian estimation, see online supplement.
Statistical power was determined by running Monte Carlo simulations. The model with five time-points was run with several combinations of effect and sample sizes. The simulations, using 1000 simulated samples, showed that for a medium-sized effect (standardized coefficient = .30), adequate power could almost be obtained (power = 78%) with a sample as small as N = 50 (with a balanced design and no missing data). Going up to N = 100, similar result (power = 75%) was obtained for a small-to-medium sized effect (standardized coefficient = .20). Finally, using the data pattern of the present study (i.e. the proportions of missing data for each session from Table 2, e.g. N = 345 for CORE-OM and N = 325 for SAI at session 1, then N = 271 for CORE-OM at session 2, and so on), it was found that for a medium-sized effect (standardized coefficient = .30), power was high (94%) while for a small-to medium sized effect (standardized coefficient = .20) power was only 62%. Importantly, regression coefficients and standard errors were estimated with minimal bias in all analyses (e.g. relative coefficient bias < 1.2% bias, relative standard error bias < 2.5% for the small-to-medium effect size condition with unbalanced design). All analyses were done using Mplus version 8.1 (Muthén & Muthén, 1998-2017).

Results

Descriptive statistics

Table 2 shows means, standard deviations, ranges, skewness and kurtosis statistics for the CORE-OM and SAI from sessions 1 to 5. As can be seen from the Table, the SAI exhibited strong skewness and (especially) kurtosis, with high mean values (> 4 “Very much” at each session). The average CORE-OM scores went down from 14.69 (SD = 7.95) in the first session to 8.89 (SD = 5.67) in session five, corresponding to an average decrease of 1.16 CORE-OM scores each session. On average, the time interval between sessions was 23.76 days (SD =
25.48); however, this was skewed with about half the sessions having less than two weeks in
between them. There was no intermittently missing data but the dataset was strongly unbalanced
with respect to time, with N diminishing rapidly for each extra session (see Table 2).

**Test of the effect of alliance quality on next-session symptom level**

An initial RI-CLPM model was set up for the first five sessions (see Figure 1), in which
all parameters were constrained to be equal across time, thus assuming a completely stationary
process. This model did not fit well to the data according to any model fit criteria ($\chi^2$ (54) =
213.40, $p < .001$; RMSEA = .09, 90% CI .08, .11, probability RMSEA < .05: < .001; CFI = .64;
SRMR = .28; AIC = 7764.63). Stationarity constraints were removed one by one, guided by
modification indices and the AIC as criterion for improving model fit. The constraints that were
removed were 1) stability of means over time ($\Delta$AIC = -86.07), 2) equality of autoregressions of
the SAI ($\Delta$AIC = -16.64), 3) equality of residual variances of CORE-OM ($\Delta$AIC = -19.87), 4)
equality of residual variances of SAI ($\Delta$AIC = -16.20), and 5) equality of reverse causation over
time ($\Delta$AIC = -8.46). The final model fit reasonably well according to most fit indices ($\chi^2$ (31) =
46.64, $p = .04$; RMSEA = .04, 90% CI .01, .06, probability < .05: .80; CFI = .96; SRMR = .11;
AIC = 7614.39).

Figure 1 shows parameter estimates of the final model. The paths from the latent within-
patient centered alliance variables (represented in Figure 1 as ws1-ws4) to next-session latent
within-patient centered distress (represented in Figure 1 as wc2-wc5) were statistically
significant ($p = .03$). The effect was such that an increase of one point on the SAI in a given
session would result in a decrease of 1.27 points on the CORE-OM in the next session (se = 0.60,
95% CI -2.44, -0.10, standardized coefficients shown in Figure 1), after controlling for stable
between-patient differences and holding the prior CORE-OM level constant. The paths from
CORE-OM to SAI differed from session to session, with significant paths at sessions 2, 4 and 5. Essentially, the results replicate a reciprocal causation model of the relationship between alliance and symptoms as shown in prior research (e.g. Xu & Tracey, 2015).

Sensitivity tests to violations of model assumptions

There are several potential threats to the validity of these findings. Some were considered in the model modifications described in the previous section, while others are better explored through sensitivity analyses (i.e. testing the robustness of findings in models that incorporate these threats). Three issues were considered; 1) bias due to including cases with fewer than the minimum number of observations required for estimating the RI-CLPM, 2) bias due to differing time intervals between sessions, 3) bias due to measurement error in the variables used, 4) bias due to missing data, and 5) bias due to shared linear trajectories between variables. Each of these are considered below. Due to the complexity of these models, they were all estimated using Bayesian estimation (see online supplement).

Including only cases with at least three data points. The final model was re-estimated on a reduced sample with only patients having at least three data points for both CORE-OM and SAI (N = 147). Model fit for this model was slightly worse than for the full data, although still reasonable according to most criteria ($\chi^2 (31) = 60.01, p = .001; \text{RMSEA} = .08, 90\% \text{CI} .05, .11, \text{probability} < .05: .06; \text{CFI} = .93; \text{SRMR} = .13$). Parameter estimates were very similar to the model based on the full data, with the effect of SAI on next session CORE-OM estimated to -1.34 (se = 0.68, $p = .05$).

Varying time between sessions. Since the models used assume equal intervals between sessions, the number of days between sessions was added as a covariate predicting both CORE-OM and SAI. Number of days between sessions was not significantly related to CORE-OM or
SAI at any session, and the estimate of the effect of SAI on next-session CORE-OM was unaffected by this (coefficient = -1.37, se = 0.66, p = .04).

**Measurement error.** In simple linear regression analysis with only one predictor, it is known that measurement error in the predictor causes negative bias in the estimated regression coefficient, that is, the estimate is closer to zero than the true effect. However, in more complex models measurement error can cause bias in any direction, and should therefore ideally be separated from true variance (Cole & Preacher, 2014). In the context of RI-CLPM with at least three waves it is possible to estimate residual variances both for the latent within-patient deviation variables and for the observed variables (Kenny & Zautra, 1995), as long as error variances are assumed equal across occasions. The residual variances of the observed variables are interpreted as measurement errors, and they affect the observed variable only at a single occasion. In contrast, the residual variances of the latent within-patient centered variables are interpreted as “dynamic errors” or “innovations” (Schuurman, Houtveen, & Hamaker, 2015) since, due to autoregression, they feed forward and affect also subsequent scores. The model is shown in Figure S1 (online supplement).

A model in which measurement errors in CORE-OM and SAI were assumed independent was compared to a model in which errors were allowed to correlate. Both models fit the data well (posterior predictive p-values were .46 for both models). Models were compared using the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002). The model with correlated measurement errors fit slightly worse according to the DIC (correlated errors DIC = 7619.00; uncorrelated errors DIC = 7617.70). In addition, the covariance between errors was not statistically significant (coefficient -.31, p = .11). Measurement error variances were statistically significant for both CORE-OM and SAI; however, separating out measurement
error had little effect on the estimate of the effect of SAI on subsequent CORE-OM (coefficient = -1.30, se = 0.67, p = .02, 95% CI -2.65, -0.03).

**Missing data analysis.** The models used include all available data, an approach that assumes that data is Missing At Random (MAR; Rubin, 1976). MAR is a less restrictive assumption than Missing Completely At Random that listwise deletion requires, since in longitudinal data MAR allows missing data to be correlated with any variable at other occasions but the value of the dependent variable at the occasion of missingness is assumed independent of missingness (Enders, 2011). In the present context, this means that the value of CORE-OM should not be the cause of data missingness. This assumption is unlikely to hold since it is reasonable to think that patients are more inclined to not fill out measures when they are more distressed (or when the alliance is bad). In such a case, it is important to do sensitivity analyses of the MAR assumption. In cross-lagged models, the model that seems most fitting is the Diggle-Kenward selection model (Diggle & Kenward, 1994; see also Falkenström et al., 2013). In this model, the probability of missingness is predicted from the repeated measures of the outcome variable at the previous occasion and at the same occasion (see Figure S2, online supplement). Because the value of the outcome variable at the dropout occasion is unknown, this model relies on the assumption of multivariate normality for the repeated measures variables. This assumption makes it possible to estimate the probability of missingness depending on the value of CORE-OM at the dropout occasion (i.e. the MAR assumption) even though that value is unknown (Enders, 2011).

The selection model fit well (posterior predictive p-value = .40), and parameter estimates showed that missingness depended on the value of the SAI at the dropout occasion and at the occasion prior to dropout (but the values of CORE-OM were unrelated to missingness).
However, as emphasized by Enders (2011), the missingness predictions should not be interpreted substantively since they depend strongly on distributional assumptions. The effect of SAI on subsequent CORE-OM was unaffected by the modeling of missing data (coefficient = -1.31, se = 0.40, \( p < .001 \), 95% CI -2.08, -0.54), although the prediction was more precise due to decreased standard errors. This means that, although the MAR assumption was likely to be violated, the estimated effect of SAI on subsequent CORE-OM was unaffected by modeling this.

**Correlated linear trends.** It is generally recognized that correlated trends over time in two time-series may seriously bias cross-lagged coefficients. The extent to which the correlated trends represent bias is due to the extent these trends represent other causes than the ones tested by the cross-lagged coefficients (Falkenström, Finkel, et al., 2017). If this is the case, time-trends should be estimated separately from cross-lagged coefficients (detrending). On the other hand, detrending in a situation when trends are generated by the phenomenon under interest (likely when there is reciprocal causation between the two variables) is a conservative strategy which is likely to yield downwardly biased estimates for the cross-lagged coefficients. Since it is generally unknown whether time-trends represent external causes, Falkenström, Finkel, et al. (2017) recommend testing both detrended and not detrended models. It should be noted that the model used so far estimated unrestricted means over time for both SAI and CORE-OM, which means that in one sense the model is already detrended. Specifically, the model separates average changes in means over time for the group as a whole from the covariances used for the cross-lagged part of the model. Thus, any linear or non-linear average change pattern for the group as a whole is accounted for in the model separately from the cross-lagged path parameters.

However, the model used so far does not consider *individual differences* in change over time. Stated differently, our model is adjusted for shared trajectories between SAI and CORE-
OM across the whole sample, but not for individual-specific trajectories. To model individual trajectories requires change to be estimated as linear, since a non-linear change model with individual differences would not be possible to estimate. Adding a random linear growth factor to the RI-CLPM results in a model described by Curran, Howard, Bainter, Lane, and McGinley (2013, see Figure S3, online supplement). This model fit the data well (posterior predictive $p = .32$). Estimates showed significant fixed and random slopes for both CORE-OM and SAI. In addition, the slopes correlated strongly (standardized coefficient = -.73, $p = .004$). The effect of detrending was that the effect of SAI on subsequent CORE-OM disappeared completely (coefficient = 0.05, $p = .48$). This may mean that the effect of alliance on subsequent symptom level is due to some confounding variable that is changing in a linear fashion across sessions. However, it should be noted that there were no significant effects at the within-patient level, that is, neither autoregression for CORE-OM or SAI nor the effect of CORE-OM on SAI were significant. This most likely means that there was too little data to separate out within-person effects from between-person time-trends in this sample. In addition, as mentioned, detrending is a conservative strategy that may result in Type-II errors.

To sum up, results were robust to all sensitivity tests except detrending against individual trajectories. This means that findings can be interpreted with more confidence, knowing that at they were unlikely to be affected by the inclusion of patient with fewer than three measurements, variable time between sessions, measurement error, violation of the MAR assumption, or spontaneous improvements affecting the whole sample equally. The result for the individual trajectory detrending is a cause for some concern, though, even if it seems likely that the data was too sparse for this highly complex model.
Moderators of alliance effect on outcome

To explore whether any patient-level variable moderated the alliance-outcome relationship, i.e. whether the alliance is a more important mechanism for some patients than others, a model with a latent interaction term between the alliance and various predictors was tested. The predictors were tested one at a time, to enable separate tests of each. The predictors that were tested were the four major problem areas that therapists identified in the CORE-A (addictions, psychosis, depression and anxiety) plus personality problems since that variable had been found to predict the alliance effect in previous studies, patient gender and age. However, none of the predictors reached statistical significance (all $p > .14$).

Discussion

The results of this study showed that changes in the working alliance from session to session predicted reduction in psychological distress/symptoms to the following session. This was an outpatient sample from mostly lower socioeconomic backgrounds seeking treatment at the two public hospitals in Nairobi, Kenya. This is the first study, to our knowledge, that studied alliance-outcome relationships on the African continent. The study can in part be seen as a cross-cultural replication of findings from other countries (e.g. Falkenström et al., 2016; Falkenström et al., 2013; Falkenström, Granström, & Holmqvist, 2014; Marker et al., 2013; Tasca & Lampard, 2012; Xu & Tracey, 2015) showing a reciprocal influence model between alliance and outcome, that is, distress influencing alliance while alliance simultaneously influencing distress. Relief from psychological distress leads to strengthening of the working alliance between the therapist and client, and as the alliance is improved psychological distress is lessened in the course of therapy. This finding shows that the work done in public hospitals in low resource
contexts of Nairobi follows similar processes as have been shown in psychotherapy in more developed, higher-income countries.

The study was based in the two national referral hospitals of Nairobi who rely on public funds, with patients from very low to low income groups including the large slum areas of Nairobi and also referral patients from the countryside and other rural pockets of Kenya. There are many great challenges in such a setting, from the very basic material issues such as covering for transport to the hospital to more socioeconomic issues such as social inequalities inherent in Kenyan society and strong stigma and prejudice associated with mental illness. The traditional cultures also do not recognize mental illness as such; e.g. most tribal languages have no word for depression. Moreover, in a low-income group a mentally ill adult becomes economically unviable. A significant number of technological and social class changes are taking place in Kenya; however, mental health is a highly under-resourced and poorly serviced sector (Bitta, Kariuki, Chengo, & Newton, 2017; Jenkins et al., 2013). As public hospitals barely manage to provide cover for all conditions, at the cost of consistency and maintaining the same therapeutic team, there is no guarantee that the patients see the same therapist each session. A change of therapist means that the alliance has to be established anew, which is another great challenge. The very brief treatments and long average time between sessions seen in our data is most likely due to financial issues making patients wait until absolutely necessary to revisit their therapist. Another issue explaining the long gaps between sessions is that many patients at the youth clinic are college students studying at boarding schools, who can only come to the clinic during school holidays.

It may be, however, that these difficult circumstances mean that establishing a working alliance is even more critical to treatment success. Stated differently, there are so many obstacles
to creating a good-enough working alliance in these contexts, so that doing so might constitute a large part of the work of therapy. If this is so, it is heartening to see that in most cases therapists and patients seem to be able to form a good alliance, and that symptoms decrease on average across therapy sessions. Qualitative studies from this project explore patients’ private theories and attributions of psychopathology and psychotherapy (Mbuthia, Kumar, Falkenström, Kuria, & Othieno, 2018), and therapists’ views on how Western psychotherapy models need to be adapted to be used with Kenyan patients (Kumar, Mita, Kuria, Othieno, & Falkenström, in preparation).

The path from symptoms/distress to alliance was in this study only significant for some sessions (sessions two, four and five). This is different from previous studies, although we are unaware of any previous study in which the homogeneity of this effect was tested (rather than assumed). We are unable to explain why the reverse effect was non-significant at some sessions and not others, however. The most likely explanation is simply random fluctuations due to a relatively small sample, with the loss of power from estimating separate coefficients at each session.

**Strengths and limitations**

These results are based on observational data, so causal conclusions cannot be made with certainty. However, a strength of this study is that the statistical models that were used enable us to rule out some of the potential alternative explanations that cannot be ruled out in traditional alliance-outcome studies based only on correlation between alliance in one session and pre-post outcome. The issue of reverse causation, i.e. psychological distress/symptoms impacting on alliance, is explicitly modeled within the RI-CLPM framework. Alternative explanations regarding confounders of the alliance-outcome correlation are important to consider. The separation of within- from between-patient variances enable us to rule out a whole class of
confounders that are stable over time, e.g. patient variables such as diagnosis, personality, temperament, genes etc., as well as average therapist effects, even though they are not included in the analysis (Moerbeek, 2004). Also, compared to most other studies of within-patient effects, we were able to test, and when necessary relax, stationarity constraints of similar variances and covariances across time (see Falkenström, Finkel, et al., 2017). Finally, the present study added the separation of measurement error from ‘true’ scores, showing that measurement errors did not affect results.

The possible threats to causal inference that remain are confounders that change between sessions, and that simultaneously affect both alliance and psychological distress. An example would be some kind of intervention from the therapist’s side that affect psychological distress directly while simultaneously improving the patient’s alliance ratings (but not affecting distress solely indirectly via the alliance, which is also possible but would not invalidate the causal status of the alliance). Results also held up to the sensitivity test for missing data. However, the unbalanced data might have caused a problem if patients who stayed longer in treatment where patients who were had stronger effects of changes in the alliance on symptoms, since these patients would then be weighted more heavily due to having more data. This could have been tested by a pattern-mixture model (Little, 1995). However, such a model was not estimable for this dataset, probably due to sparse data.

In addition, the results were robust for average group change across sessions, what in the time-series literature is called detrending (controlling for average trends over time). Detrending is important if it is likely that the cross-lagged relationships are affected by general effects of time. Psychological distress can of course change as an effect of time; this is what we mean by ‘spontaneous improvement’. It is not quite as obvious that the quality of the working alliance
would improve simply by the passage of time, but that may still be a possibility. In the present study, detrending against individual linear change made the working alliance effect, along with everything else on the within-patient level, disappear. The fact that the alliance effect disappeared in this model is a potential problem, that could indicate that the effect of the alliance on next-session symptoms is caused by a third variable that changes in a linear fashion across sessions. However, the fact that all effects at the within-patient level disappeared when individual detrending was applied, indicates that the data in the present study may have been too sparse to enable separation of within-person effects of alliance on symptoms from linear change over time.

There are several limitations of the present study. An unknown number of patients (probably at least as many as we have data on) were never approached by the research assistants during the study time due to issues such as unavailability of any of the assistants on a particular day or that a patient was missed at the clinic. It seems likely, however, that the assumption of Missing Completely at Random that is required for listwise deletion may hold for these (although this is of course not testable since we have no data on these patients). In addition, the measures of patient-level moderators suffer from lack of reliability checks and there is no information on their validity, so the findings (or lack thereof) for moderators should be interpreted cautiously. Further limitations include lack of information on therapists, their orientation and training, and the fact that it was not possible to test for potential differences in alliance effect among treatments. The data did not conform to normality assumptions, which is also a limitation. With more data non-linear models could have been tested; however, with this sample size this would have been too much of a stretch. In addition, the Bayesian estimator does not rely as much on normality assumptions as Maximum Likelihood estimation.
Conclusions, clinical implications, and suggestions for future research

The findings of this study replicate prior research done in Western countries on the importance of monitoring the fluctuations in the working alliance from session to session. The clinical implication is not necessarily that the therapist should talk to the patient about the alliance in each and every session, but the therapist should be mindful that if the alliance quality goes down in a particular session, symptoms are likely to worsen to the next session and if the alliance can be improved during a session, symptoms are likely to improve to the next. Alliance ruptures can be dealt with by applying the rupture-resolution framework of Safran and Muran (2000). In addition, as noted by Falkenström et al. (2016), if a patient appears to have worse symptoms in a given session, it is likely that the working alliance quality will also suffer in that session. This means that the therapist can monitor symptom fluctuations across sessions, and when symptoms deteriorate it is also important to keep an eye on the alliance to prevent a negative spiral in which symptoms deteriorate, leading to worsening of the alliance, which in turn leads to further deterioration of symptoms.

Future research should attempt to study the interrelationships between alliance and therapeutic interventions, to see whether the alliance is a precondition for technique to work properly or a mediator of a technique-outcome relationship. Generally, further, more rigorous controlled research in African and other non-Western countries are needed. In addition, moderators of the alliance effect should be investigated further, since it is likely that alliance fluctuations are more important for some patients than for others or in some treatments compared to others. These considerations become important as the global mental health movement puts weight on task-sharing and task-shifting of mental health services in primary health care settings in regions like Kenya (Hanlon et al., 2014; Musyimi et al., 2017). The working alliance is partly
an immediate human empathic connection, but it is also a process that develops through training, rigor and supervision.
References


alliance and session outcome. *Journal of Consulting and Clinical Psychology, 84*(6), 484-496. doi:10.1037/ccp0000058
Table 1.

Demographic information about participating patients.

<table>
<thead>
<tr>
<th>Sex</th>
<th>Freq.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>249</td>
<td>72.59</td>
</tr>
<tr>
<td>Female</td>
<td>94</td>
<td>27.41</td>
</tr>
<tr>
<td>Previous therapy</td>
<td>55</td>
<td>16.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Presenting problem</th>
<th>Freq.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addictions</td>
<td>188</td>
<td>54.81</td>
</tr>
<tr>
<td>Psychosis</td>
<td>60</td>
<td>17.49</td>
</tr>
<tr>
<td>Depression</td>
<td>58</td>
<td>16.91</td>
</tr>
<tr>
<td>Anxiety</td>
<td>41</td>
<td>11.95</td>
</tr>
<tr>
<td>Interpersonal relations</td>
<td>22</td>
<td>6.41</td>
</tr>
<tr>
<td>Physical problems</td>
<td>22</td>
<td>6.41</td>
</tr>
<tr>
<td>Work/academic</td>
<td>22</td>
<td>6.41</td>
</tr>
<tr>
<td>Other problems</td>
<td>22</td>
<td>6.41</td>
</tr>
<tr>
<td>Self-esteem</td>
<td>17</td>
<td>4.96</td>
</tr>
<tr>
<td>Trauma/abuse</td>
<td>16</td>
<td>4.66</td>
</tr>
<tr>
<td>Personality problems</td>
<td>15</td>
<td>4.37</td>
</tr>
<tr>
<td>Eating disorders</td>
<td>11</td>
<td>3.21</td>
</tr>
<tr>
<td>Living/welfare</td>
<td>12</td>
<td>3.50</td>
</tr>
<tr>
<td>Bereavement/loss</td>
<td>8</td>
<td>2.33</td>
</tr>
</tbody>
</table>
Table 2.

*Descriptive statistics for the CORE-OM and SAI at sessions 1 to 5.*

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>CORE_t1</td>
<td>345</td>
<td>14.69</td>
<td>7.95</td>
<td>0.00</td>
<td>37.06</td>
<td>0.43</td>
<td>2.55</td>
</tr>
<tr>
<td>CORE_t2</td>
<td>271</td>
<td>11.78</td>
<td>7.04</td>
<td>0.00</td>
<td>31.47</td>
<td>0.51</td>
<td>2.50</td>
</tr>
<tr>
<td>CORE_t3</td>
<td>165</td>
<td>10.66</td>
<td>7.36</td>
<td>0.00</td>
<td>32.65</td>
<td>0.71</td>
<td>2.63</td>
</tr>
<tr>
<td>CORE_t4</td>
<td>71</td>
<td>9.12</td>
<td>5.26</td>
<td>0.00</td>
<td>22.06</td>
<td>0.39</td>
<td>2.62</td>
</tr>
<tr>
<td>CORE_t5</td>
<td>38</td>
<td>8.89</td>
<td>5.67</td>
<td>0.00</td>
<td>20.29</td>
<td>0.52</td>
<td>2.09</td>
</tr>
<tr>
<td>SAIP_t1</td>
<td>325</td>
<td>4.09</td>
<td>0.83</td>
<td>0.67</td>
<td>5.00</td>
<td>-1.33</td>
<td>5.38</td>
</tr>
<tr>
<td>SAIP_t2</td>
<td>258</td>
<td>4.27</td>
<td>0.78</td>
<td>1.17</td>
<td>5.00</td>
<td>-1.20</td>
<td>4.14</td>
</tr>
<tr>
<td>SAIP_t3</td>
<td>156</td>
<td>4.33</td>
<td>0.70</td>
<td>1.00</td>
<td>5.00</td>
<td>-1.33</td>
<td>5.60</td>
</tr>
<tr>
<td>SAIP_t4</td>
<td>66</td>
<td>4.24</td>
<td>0.80</td>
<td>0.00</td>
<td>5.00</td>
<td>-2.44</td>
<td>12.93</td>
</tr>
<tr>
<td>SAIP_t5</td>
<td>37</td>
<td>4.03</td>
<td>0.92</td>
<td>0.50</td>
<td>5.00</td>
<td>-1.84</td>
<td>7.47</td>
</tr>
</tbody>
</table>
Note. Circles represent latent variables and rectangles represent observed variables. Variables wc1-xc5 and ws1-ws5 are latent within-patient deviation variables (i.e. Level-1 variables), while Between_CORE and Between_SAI are the Level-2 random intercepts capturing the between-patient variances. Estimates are unstandardized regression coefficients, and the ones in bold going diagonally represent the effect of alliance on psychological distress at the next session while holding the present session psychological distress constant. The vertical paths from the wc to ws variables represent the effects of psychological distress experienced in the previous week on working alliance in the session. Horizontal paths represent autoregression.

* $p < .05$, ** $p < .01$, *** $p < .001$

Figure 1. Random Intercept Cross-Lagged Panel model of CORE-OM and SAI over the first five sessions of therapy.
Online supplement

Bayesian estimation

Estimation of Bayesian models is usually done using simulation-based methods, called Markov Chain Monte Carlo (MCMC) estimation. MCMC simulates values of parameters from the posterior distribution, given the model, the prior distributions, 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 should converge on the most likely parameter estimates. Usually more than one chain are 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 present study was not done using a fully Bayesian approach, which would have involved incorporating prior information in the estimation so that results represent a combination of prior knowledge and information in the data. Instead, a frequentist use of Bayesian estimation was used, and uninformative model priors were used to ensure that estimates were based on the data only. For instance, priors for variances were specified to follow an inverse Gamma distribution with alpha = -1 and Beta = 0, which means that the means and variances of the prior are infinite. This is a so-called improper prior, which is often used as an uninformative prior (Asparouhov & Muthén, 2010). Importantly, although this prior allows any positive variance estimate with equal probability, negative variance estimates cannot occur.

Model fit for 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 predicted and observed covariance matrices is smaller than the discrepancy between predicted and simulated covariance matrices for future samples (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 generated data) indicates good fit.

Convergence of the Markov Chains needs to be carefully assessed. In this study we used the Gelman-Rubin convergence criterion, which compares the estimated between- and within-chain variances for each model parameter. The Potential Scale Reduction (PSR) factor is defined as the ratio of the pooled between and within chain variances and the within-chain variance. The PSR should approach 1.00 when convergence has been achieved. We used the criterion of PSR < 1.05 (Asparouhov & Muthén, 2010), and the Kolmogorov-Smirnoff test of between-chain parameter differences (Zyphur & Oswald, 2013). The latter should ideally not be statistically significant for any parameter. After initial convergence was established, the estimation was re-run with at least twice the number of chains to ensure that PSR values remained below the convergence criterion.
Figures of models testing assumptions

1) Measurement errors (Figure 2), 2) missing data (Figure 3), and 3) correlated trends (Figure 4).

Note. Circles represent latent variables while rectangles represent observed variables. Variables wc1-xc5 and ws1-ws5 are latent within-person deviation variables (i.e. Level-1 variables), while Between_CORE and Between_SAI are the Level-2 random intercepts. The variables labeled ‘d’ represent so-called dynamic errors, or innovations, while the variables labeled ‘e’ represent measurement errors. The arced, double-headed arrow between the ‘e’ variables at the same occasion show the estimated correlation between measurement errors of SAI and CORE-OM.

*Figure S1. Random Intercept Cross-Lagged Panel model of CORE-OM and SAI over the first five sessions of therapy, including the separation of measurement error.*
Note. Circles represent latent variables while rectangles represent observed variables. Variables wc1-xc5 and ws1-ws5 are latent within-person deviation variables (i.e. Level-1 variables), while Between_CORE and Between_SAI are the Level-2 random intercepts. The variables labeled ‘Missing_2’ – ‘Missing_5’ are missing data indicators, which are coded ‘0’ for non-missing, ‘1’ the first time data is missing, and as missing the second time data is missing (Enders, 2011). Paths from SAI_t1 – SAI_t5 to Missing_2 – Missing_5 were included but are not shown in the diagram to avoid cluttering.

Figure S2. Diggle-Kenward selection model for the Random Intercept Cross-Lagged Panel model of CORE-OM and SAI over the first five sessions of therapy.
Note. Circles represent latent variables while rectangles represent observed variables. Variables wc1-xc5 and ws1-ws5 are latent within-person deviation variables (i.e. Level-1 variables), ‘Intercept_CORE’ and ‘Intercept_SAI’ are level-2 random intercepts representing varying initial status among patients, and ‘Slope_CORE’ and ‘Slope_SAI’ are random linear time-slopes representing variations in linear change over time.

*Figure S3. Structured Residuals Model (Curran, Howard, Bainter, Lane, & McGinley, 2014) of CORE-OM and SAI over the first five sessions of therapy.*
Table S1. Zero-order correlation matrix.

<table>
<thead>
<tr>
<th></th>
<th>CORE_t1</th>
<th>CORE_t2</th>
<th>CORE_t3</th>
<th>CORE_t4</th>
<th>CORE_t5</th>
<th>SAIP_t1</th>
<th>SAIP_t2</th>
<th>SAIP_t3</th>
<th>SAIP_t4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CORE_t2</td>
<td>0.50*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CORE_t3</td>
<td>0.37*</td>
<td>0.58*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CORE_t4</td>
<td>0.35*</td>
<td>0.48*</td>
<td>0.54*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CORE_t5</td>
<td>0.34*</td>
<td>0.27</td>
<td>0.50*</td>
<td>0.61*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAIP_t1</td>
<td>-0.11</td>
<td>-0.15*</td>
<td>-0.15</td>
<td>-0.15</td>
<td>-0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAIP_t2</td>
<td>-0.04</td>
<td>-0.22*</td>
<td>-0.21*</td>
<td>-0.19</td>
<td>-0.29</td>
<td>0.49*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAIP_t3</td>
<td>-0.03</td>
<td>-0.14</td>
<td>-0.13</td>
<td>-0.22</td>
<td>-0.31</td>
<td>0.46*</td>
<td>0.64*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAIP_t4</td>
<td>0.12</td>
<td>0.13</td>
<td>0.09</td>
<td>-0.28*</td>
<td>-0.26</td>
<td>0.46*</td>
<td>0.59*</td>
<td>0.81*</td>
<td></td>
</tr>
<tr>
<td>SAIP_t5</td>
<td>0.18</td>
<td>0.16</td>
<td>-0.02</td>
<td>-0.42*</td>
<td>-0.45*</td>
<td>0.36*</td>
<td>0.50*</td>
<td>0.60*</td>
<td>0.79*</td>
</tr>
</tbody>
</table>

Note. Estimates reported in the article are not reproducible from this table, since these are based on latent disaggregation of within-from between-patient sources of variance.
Bayesian estimation

Estimation of Bayesian models is usually done using simulation-based methods, called Markov Chain Monte Carlo (MCMC) estimation. MCMC simulates values of parameters from the posterior distribution, given the model, the prior distributions, 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 should converge on the most likely parameter estimates. Usually more than one chain are 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 present study was not done using a fully Bayesian approach, which would have involved incorporating prior information in the estimation so that results represent a combination of prior knowledge and information in the data. Instead, a frequentist use of Bayesian estimation was used, and uninformative model priors were used to ensure that estimates were based on the data only. For instance, priors for variances were specified to follow an inverse Gamma distribution with alpha = -1 and Beta = 0, which means that the means and variances of the prior are infinite. This is a so-called improper prior, which is often used as an uninformative prior (Asparouhov & Muthén, 2010). Importantly, although this prior allows any positive variance estimate with equal probability, negative variance estimates cannot occur.

Model fit for 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 predicted and observed covariance
matrices is smaller than the discrepancy between predicted and simulated covariance matrices for future samples (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 generated data) indicates good fit.

Convergence of the Markov Chains needs to be carefully assessed. In this study we used the Gelman-Rubin convergence criterion, which compares the estimated between- and within-chain variances for each model parameter. The Potential Scale Reduction (PSR) factor is defined as the ratio of the pooled between and within chain variances and the within-chain variance. The PSR should approach 1.00 when convergence has been achieved. We used the criterion of PSR < 1.05 (Asparouhov & Muthén, 2010), and the Kolmogorov-Smirnoff test of between-chain parameter differences (Zyphur & Oswald, 2013). The latter should ideally not be statistically significant for any parameter. After initial convergence was established, the estimation was re-run with at least twice the number of chains to ensure that PSR values remained below the convergence criterion.
Figures of models testing assumptions

1) Measurement errors (Figure 2), 2) missing data (Figure 3), and 3) correlated trends (Figure 4).

Note. Circles represent latent variables while rectangles represent observed variables. Variables wc1-xc5 and ws1-ws5 are latent within-person deviation variables (i.e. Level-1 variables), while Between_CORE and Between_SAI are the Level-2 random intercepts. The variables labeled ‘d’ represent so-called dynamic errors, or innovations, while the variables labeled ‘e’ represent measurement errors. The arced, double-headed arrow between the ‘e’ variables at the same occasion show the estimated correlation between measurement errors of SAI and CORE-OM.

Figure S1. Random Intercept Cross-Lagged Panel model of CORE-OM and SAI over the first five sessions of therapy, including the separation of measurement error.
Note. Circles represent latent variables while rectangles represent observed variables. Variables wc1-xc5 and ws1-ws5 are latent within-person deviation variables (i.e. Level-1 variables), while Between_CORE and Between_SAI are the Level-2 random intercepts. The variables labeled ‘Missing_2’ – ‘Missing_5’ are missing data indicators, which are coded ‘0’ for non-missing, ‘1’ the first time data is missing, and as missing the second time data is missing (Enders, 2011). Paths from SAI_t1 – SAI_t5 to Missing_2 – Missing_5 were included but are not shown in the diagram to avoid cluttering.

Figure S2. Diggle-Kenward selection model for the Random Intercept Cross-Lagged Panel model of CORE-OM and SAI over the first five sessions of therapy.
Note. Circles represent latent variables while rectangles represent observed variables. Variables wc1-xc5 and ws1-ws5 are latent within-person deviation variables (i.e. Level-1 variables), ‘Intercept_CORE’ and ‘Intercept_SAI’ are level-2 random intercepts representing varying initial status among patients, and ‘Slope_CORE’ and ‘Slope_SAI’ are random linear time-slopes representing variations in linear change over time.

Figure S3. Structured Residuals Model (Curran, Howard, Bainter, Lane, & McGinley, 2014) of CORE-OM and SAI over the first five sessions of therapy.
Table S1. Zero-order correlation matrix.

<table>
<thead>
<tr>
<th></th>
<th>CORE_t1</th>
<th>CORE_t2</th>
<th>CORE_t3</th>
<th>CORE_t4</th>
<th>CORE_t5</th>
<th>SAIP_t1</th>
<th>SAIP_t2</th>
<th>SAIP_t3</th>
<th>SAIP_t4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CORE_t2</td>
<td>0.50*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CORE_t3</td>
<td>0.37*</td>
<td>0.58*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CORE_t4</td>
<td>0.35*</td>
<td>0.48*</td>
<td>0.54*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CORE_t5</td>
<td>0.34*</td>
<td>0.27</td>
<td>0.50*</td>
<td>0.61*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAIP_t1</td>
<td>-0.11</td>
<td>-0.15*</td>
<td>-0.15</td>
<td>-0.15</td>
<td>-0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAIP_t2</td>
<td>-0.04</td>
<td>-0.22*</td>
<td>-0.21*</td>
<td>-0.19</td>
<td>-0.29</td>
<td>0.49*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAIP_t3</td>
<td>-0.03</td>
<td>-0.14</td>
<td>-0.13</td>
<td>-0.22</td>
<td>-0.31</td>
<td>0.46*</td>
<td>0.64*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAIP_t4</td>
<td>0.12</td>
<td>0.13</td>
<td>0.09</td>
<td>-0.28*</td>
<td>-0.26</td>
<td>0.46*</td>
<td>0.59*</td>
<td>0.81*</td>
<td></td>
</tr>
<tr>
<td>SAIP_t5</td>
<td>0.18</td>
<td>0.16</td>
<td>-0.02</td>
<td>-0.42*</td>
<td>-0.45*</td>
<td>0.36*</td>
<td>0.50*</td>
<td>0.60*</td>
<td>0.79*</td>
</tr>
</tbody>
</table>

Note. Estimates reported in the article are not reproducible from this table, since these are based on latent disaggregation of within-from between-patient sources of variance.