

The Online Alliance Assessment

Sanne Peereboom, Esmée Roosenmaallen

May 24, 2022

Abstract

The recent COVID-19 pandemic has highlighted the necessity for online counseling. To determine the quality of counseling, the therapeutic alliance can be assessed. This consists of personal bond between client and counselor, agreement on tasks, and agreement on goals. For in-person counseling settings this concept has been widely researched. However, there is no tool to assess alliance in an online setting, suitable for youth, and reciprocal. That is why we have developed the Online Alliance Assessment. In this research we assess the construct validity of this measure; whether the components of alliance are measured congruently between clients and counselors in a student sample; and whether these groups differ in their ratings of alliance. A multiple group confirmatory factor analysis (MGCFA) confirmed that at least two components of alliance are present and construct validity is achieved. Additionally, the constructs are measured congruently and the questionnaire is reciprocal. Analysis of sum scores shows some disagreement within client-counselor pairs on task and goal agreement. We suggest further research with representative samples and to balance the questionnaire so that each factor is represented more equally and reliably.

Keywords: therapeutic alliance, online counseling, confirmatory factor analysis, MGCFA, sum scores

Introduction

Over recent years, online therapy has been discovered to be just as effective to treat patients as face-to-face treatment (Wagner et al., 2014). Thus, there has been an increase in development of online tools to assist therapists (Andersson et al., 2019). However, there was no immediate need for health care professionals to adopt these tools, until the recent outbreak of COVID-19 resulting in a worldwide lock-down (Feijt et al., 2020). As a result, the normal way of providing health care was no longer applicable and therapeutic sessions had to be moved online. Even after the pandemic, online health care is beneficial due to the high convenience, flexibility, and accessibility to mental health care it provides (Feijt et al., 2020). Therefore, it is important to be able to assess the quality of online counseling.

In any therapeutic relationship, also known as alliance, there are three key components: agreement on treatment goals, agreement on the tasks, and a positive personal bond between client and therapist (Bordin, 1979). These elements are required in an optimal alliance because the patient has to believe in the therapist's credibility and the therapist has to believe in the patient's willingness to change (Ardito & Rabellino, 2011). This allows the patient to believe and accept the required treatment. Unsurprisingly, the quality of this reciprocal bond has been found to be a moderate and robust predictor of a positive treatment outcome (Ardito & Rabellino, 2011). This makes alliance an interesting concept when it comes to providing feedback to the counselor. As the varying opinions on the reciprocal alliance can be used to improve the counseling quality. On a group level, there could be structural differences in how alliance is experienced between clients and counselors. This information could be used in organisations, where training's can be provided to counselors to improve a component of alliance where there is a difference in alliance rating. The (reciprocal) assessment of alliance using questionnaires has been abundantly explored in in-person therapy sessions. However,

as online therapy is a rather young field, proper tools for measuring alliance under these circumstances are still lacking.

At the moment, sources of online counseling, such as Kindertelefoon (a children's helpline) and ChatmetFIER offer anonymous chat-based support. In many of these organisations every contact moment is unique and cannot be resumed: there is usually no ongoing relationship between client and counselor. Since therapeutic alliance helps predict treatment outcome, it is worth assessing the quality of this alliance after a single online anonymous counseling session. As of yet, no tool exists to determine whether an alliance can be created in such a short session. Additionally, we were unable to find a single instrument measuring alliance which is suitable under the circumstances the Kindertelefoon and ChatmetFIER operate under. To assess alliance in these organisations we have to consider a questionnaire that handles both youth and family (i.e., short and easy to understand); is suitable for online counseling interactions; and is reciprocal. Thus, in this research we present and test the Online Alliance Assessment to provide feedback to the therapist after a short and anonymous chat session. This tool meets all our criteria and consists of items from several available alliance questionnaires (Horvath & Greenberg, 1989).

In the present research, we assess the validity of the Online Alliance Assessment and how the results can be used for feedback to organisations and counselors to improve the quality of alliance. The first research question is then as follows: "What is the construct validity of the Online Alliance Assessment?". The pre-specified structure should fit the data well: we expect to find that our alliance scale indeed consists of the three components of alliance as mentioned in Bordin (1979). Our second question is: "Are these constructs measured congruently between counselor and client?". The same structure is expected between clients and counselors as the Online Alliance Assessment asks clients and counselors

the same questions, except from differing perspectives. This would indicate that the Online Alliance Assessment can indeed be considered reciprocal. The final research question is: "Do clients and counselors differ in their rating of the therapeutic alliance?". Both individual client and counselor pair differences as well as client and counselor group differences will be investigated. Any potential significant differences in ratings between clients and counselors could be used to raise awareness amongst counselors about said differences and to provide feedback and improve the quality of counseling on both a group level and an individual level. Clients and counselors are expected to have some variation between their responses, but a discrepancy of 25% or higher would require a closer look at the counselor's performance.

Methods

Participants and design

This study has a cross-sectional design. The client and counselor samples consist of students from Avans Hogeschool. The students participated in the study as part of a course worth 3 ECTS. No general demographic information was collected. In total, 260 participants participated, corresponding to 130 client-counselor pairs

Measures

The measure used in this study is the Online Alliance Assessment. The questionnaire was administered through Qualtrics. It consists of 11 questions that should measure the client-counselor alliance (Appendix A). The questions were answered using a three-point Likert scale (insufficient - sufficient - good). For the components of alliance, there were eight questions associated with the personal bond between client and counselor. These were questions 1 to 9, excluding question 6. There were two questions about the agreement on treatment goals (questions 6 and 10), and one question about agreement on treatment tasks (question 11).

Procedure

Participants were assigned a client or counselor role in the context of a single anonymous online chat session that had a duration of 40 minutes. Client and counselor participants were seated in separate rooms to prevent them from communicating outside of the chat session. Clients were given a fictional case that they role played during this chat session. After completion, client and counselor pairs filled in their respective version of the Online Alliance Assessment in Qualtrics

(Appendix A). Immediately afterwards, clients and counselors swapped roles and repeated the above procedure with a new fictional scenario.

Statistical analyses

Initial data analysis

The analyses were all performed in R-studio version 1.4.1717 running on R version 4.1.2 (2022). The Qualtrics exported SPSS data set was read into R and all relevant columns for this analysis were saved in a client and counselor data set. We removed any entries that were test runs and measurements that were not chat medium. Clients and counselors were paired using an ID code representing one unique client-counselor pair. There were no negatively phrased questions in the questionnaire, so re-coding of variables was not necessary.

Due to input errors and time constraints during the procedure, some participant data was either unable to be paired to their corresponding client or counselor, or missing altogether. This data can be considered missing completely at random (MCAR), that is, there is no relationship between the missingness and the outcome or measured variables. For this reason, we opted for pairwise deletion in the paired client-counselor data set, and listwise deletion for the separate client and counselor groups (Schreiber et al., 2010). So, paired client and counselor observations were removed if either client data or counselor data was missing. For unpaired client and counselor groups, we only removed the individuals whose data was missing and not their corresponding client or counselor. Two of the unique ID codes occurred twice within the client or counselor group, presumably due to input errors. These samples cannot be used to create unique client-counselor pairs and are treated as missing in the paired client-counselor analyses. They are still used in the analyses to assess construct validity and group differences in alliance rating, as no client-counselor pairing is necessary there. After removing the

missing data, there were 106 client responses, 105 counselor responses, and 100 unique client-counselor pairs.

Main analysis

To answer the first two research questions, a multiple group confirmatory factor analysis (MGCFA) comparing client and counselor groups was used. The MGCFA tests whether a predetermined grouping of items that measures one or more unobserved constructs fits the data, in our case the components of therapeutic alliance. Before a CFA is run, we look at the Spearman correlations of the 11 questions. High inter-item correlations imply that certain pairs of questions could be considered duplicates: if they are very strongly related, they might be an indicator for the same thing. In that case, we could consider omitting or replacing one of these items.

To run a CFA, we must specify which questions belong to what component of alliance. In this analysis, this is called a factor. We have defined the factors as explained in the Measures section, with a small adjustment. For a CFA, each component needs at least two questions, optimally three, to work. As there is only one question related to treatment goals, we grouped this question together with agreement on tasks into one factor representing agreement on tasks and goals. The two factors "personal bond" and "agreement on tasks and goals" will be used for all MGCFA models (Appendix B). These are run from a least restrictive to a most restrictive model.

For the final research question, we consider the client-counselor pairs. For each pair, we calculate the differences within client-counselor pairs in their ratings of the alliance factors and the overall alliance. Pairs that differ more than 25% on at least one of the components, or on their overall rating, indicate an unexpectedly large discrepancy between client and counselor. The responses of these pairs can

then be examined more closely in order to give a counselor feedback.

Assumption checks

The maximum likelihood (ML) loss function in the CFA assumes that the items are continuous. It is possible to specify that variables are categorical by using a diagonally weighted least squares (DWLS) loss function, but the interpretation of the results is not as straightforward. Therefore, we first assess whether we can assume that our questionnaire responses are continuous without substantially compromising the model fit. To do this, we will run two series of analyses: one series assuming continuous data by setting the loss function to ML, and one series assuming categorical data by setting the loss function to DWLS. For both series of analyses, we will use the unit variance (UVI) constraint. The variance of the factors are set to 1, so that all factor loadings can be freely estimated. This does not change the results or interpretation of the results, but it is preferable because the factor loadings are not directly restricted, making it easier to set them equal between groups when restricting the MGCFA models. We will compare the fit of each invariance model and allow some difference in fit between the categorical and the continuous CFA models, as long as it does not change the conclusion. Should the continuous CFA models fit the data substantially worse, we will proceed with the categorical CFA models. For the MGCFA with DWLS loss there needs to be an observation in every answer category in both groups. This was not the case for the rating of 1 on question 3 in the client group, so we had to remove two counselor observations that answered question 3 with a score of 1. So, for the MGCFA with DWLS loss, there were 103 counselor responses instead of 105.

The least restrictive configural invariance model must have a logical solution to continue: all variances must be positive; all R^2 values and correlations must be between 0 and 1; and standard errors (SEs) should not be unexpectedly large or negative. This is to check whether the model is defined correctly. If the model is

logical, we check whether the CFA is a good fit for the data. So, the specified factors and their items accurately describe the data. The model fit can be assessed by different fit measures, where some fit measures describe "goodness-of-fit", and other fit measures describe "badness-of-fit". We consider one goodness-of-fit measure: the Comparative Fit Index (CFI), which must be above 0.90 for the model fit to be acceptable (Xia & Yang, 2018). Additionally, we look at three different badness-of-fit measures. The first is the Root Mean Square Error of Approximation (RMSEA): when this value is smaller than 0.05 we have an excellent fit, between 0.05 and 0.08 is acceptable, between 0.08 and 0.1 is marginal, and any higher than this is a poor fit (Kim et al., 2016). We also consider the Standardized Root Mean Square Residual (SRMR), which must be smaller than 0.08 for the fit to be acceptable. The last badness-of-fit measure we consider is the chi-squared (X^2) statistic. X^2 can assess a single model fit or compare two different model fits, where non-significance indicates a good fit. If these fit measures indicate a properly fitting model, we can continue to restrict the configural invariance model.

Restricting the model means to force certain parameters to be equal across groups. If the model still fits well after this restriction, the groups are indeed equal on the restricted parameter. This must be done in a sequential manner. After the configural invariance model, we restrict the factor loadings to be equal for both groups, resulting in a metric invariance model. This means that each question is equally important to the construct for the client and counselor groups. If the model fits equally well, we can restrict the model further to a scalar invariance model. This model sets the mean response per question equal between groups and shows what questions may have had significantly different average answers between groups. If the model still fits well, the model is further restricted to a strict factorial invariance model. This model sets the ranges of the question scores to be equal. This model could show, for example, that clients were more variable in how well they trusted the counselor (as compared to the counselors' ratings).

When the difference in fit between the models is significant, there is at least one item for which the groups differ in the restricted aspect. In that case, we cannot run a subsequent model with additional restrictions. However, a partial invariance model can be used in which some items are set to be equal, while the items on which the groups differ are allowed to be freely estimated. In this way, we can determine if there are individual questions that cannot be considered to measure the same thing (metric invariance); have the same mean score (scalar invariance); or have the same variance (strict factorial invariance) between groups.

Using the final model we will calculate sum scores to analyze the differences within individual client-counselor pairs. We can use a raw sum score or a summed factor score. A raw sum score is the sum of the responses where all questions are weighted equally, while a summed factor score considers the weight of each item. A t-test will assess whether the raw sum scores significantly differ from summed factor scores. If that is not the case, we can use raw sum scores. Otherwise, the summed factor scores are more representative of the data.

All analyses can be replicated by use of the R-code uploaded to Github (Boolytical, 2022). This code includes some features that might be desirable for future research, such as the possibility to filter clients and counselors by organisation. In the current data set the organisations were not used as they were not representative or recorded in the same style. For an explanation of the code, refer to Appendix D. An explanation of the output can be found in Appendix E.

Results

The Spearman correlation matrix between the 11 questionnaire items can be found in Figure 1. No inter-item correlation above 0.6 is present, thus, the correlations are relatively low. However, the correlation between Q10 and Q11 (which both belong to the construct "agreement on tasks and goals") is relatively high compared to the other inter-item correlations ($\rho = .55$).

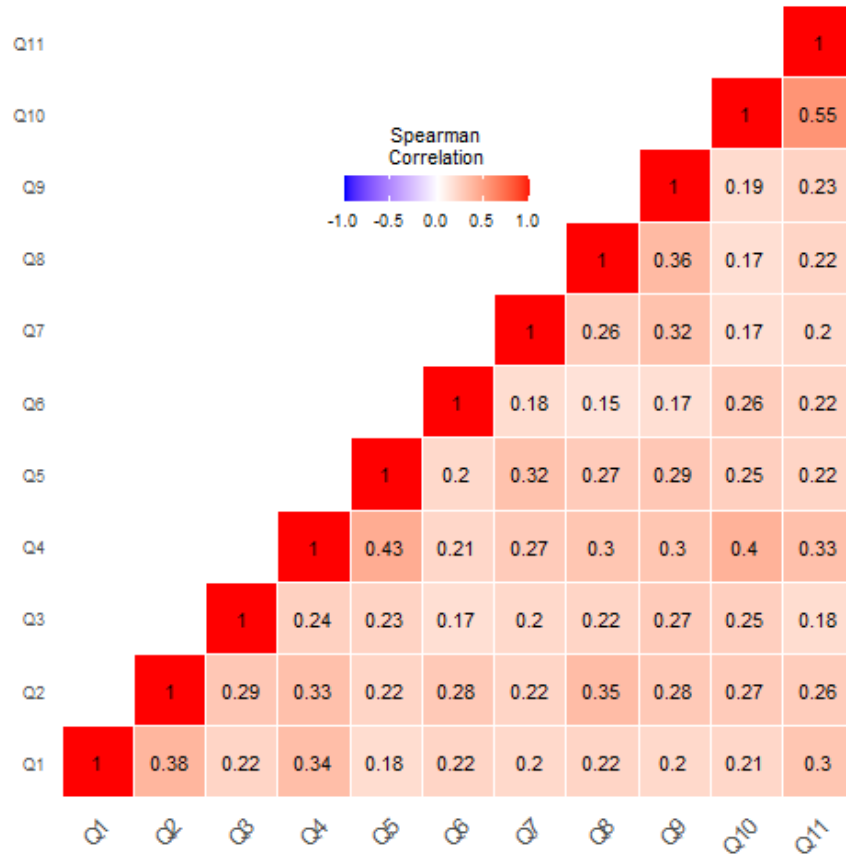


Figure 1: Spearman correlation matrix of the 11 closed questions of the Online Alliance Assessment. There is no distinguish between client or counselor questionnaire.

The results of the MGCFA are summarized in Table 1. In both sets of analyses, all X^2 values were non-significant, so the models fit well. One p-value is also missing for the categorical MGCFA model. This is because setting the residual variance to be equal did not lead to changes in the model compared to the equal items means model. The other fit measures are good. In short, the factor loadings, item means, and residual variances are all approximately equal. The fit measures of the continuous and categorical MGCFA both lead to the same conclusions. As such, we choose the continuous MGCFA (with ML loss function) with equal residual variance as the final model.

Table 1: Fit indices for the Multi-Group Confirmatory Factor Analysis with two factors (Appendix B) on Online Alliance Assessment. There are four specified models using the maximum likelihood (ML) loss function (which assume data are continuous) and four models using the Diagonally Weighted Least Squares (DWLS) loss function (which assume data are categorical).

	Chi-square (X^2)	df	CFI	RMSEA	SRMR	p-value
MGCFA with ML loss						
Two group model	108.13	86	0.948	0.049	0.066	0.054
Equal factor loadings	117.27	97	0.952	0.045	0.075	0.609
Equal item means	128.36	106	0.947	0.045	0.079	0.269
Equal residual variance	142.25	117	0.940	0.045	0.086	0.239
MGCFA with DWLS loss						
Two group model	91.73	86	0.996	0.025	0.091	0.316
Equal factor loadings	117.30	97	0.986	0.045	0.105	0.294
Equal item means	126.93	105	0.985	0.045	0.093	0.160
Equal residual variance	126.93	105	0.985	0.045	0.093	NA

The factor loadings of the chosen model for both the client and counselor group can be found in Figure 2. "Personal" represents the personal bond between client and counselor, and "Goals" represents agreement on tasks and treatment goals. The arrows from the factors towards Q1-Q11 represent the factor loadings. All factor loadings exceed 0.30, indicating a strong enough association with their respective factors. The arrow between the two factors shows the correlation between the factors, which is equal to 0.70 for the client group and equal to 0.59 for the counselor group. This implies that there is an overlap in what the constructs measure. Thus, a single factor model might be a better choice. However, creating a one-factor model, where all items measure the same construct, causes the two group model to be a poor fit ($p < 0.001$) to the data (Appendix C). It also reduces the fit measures compared to the two factor models (Table 1). Thus, we stick with a two factor model.

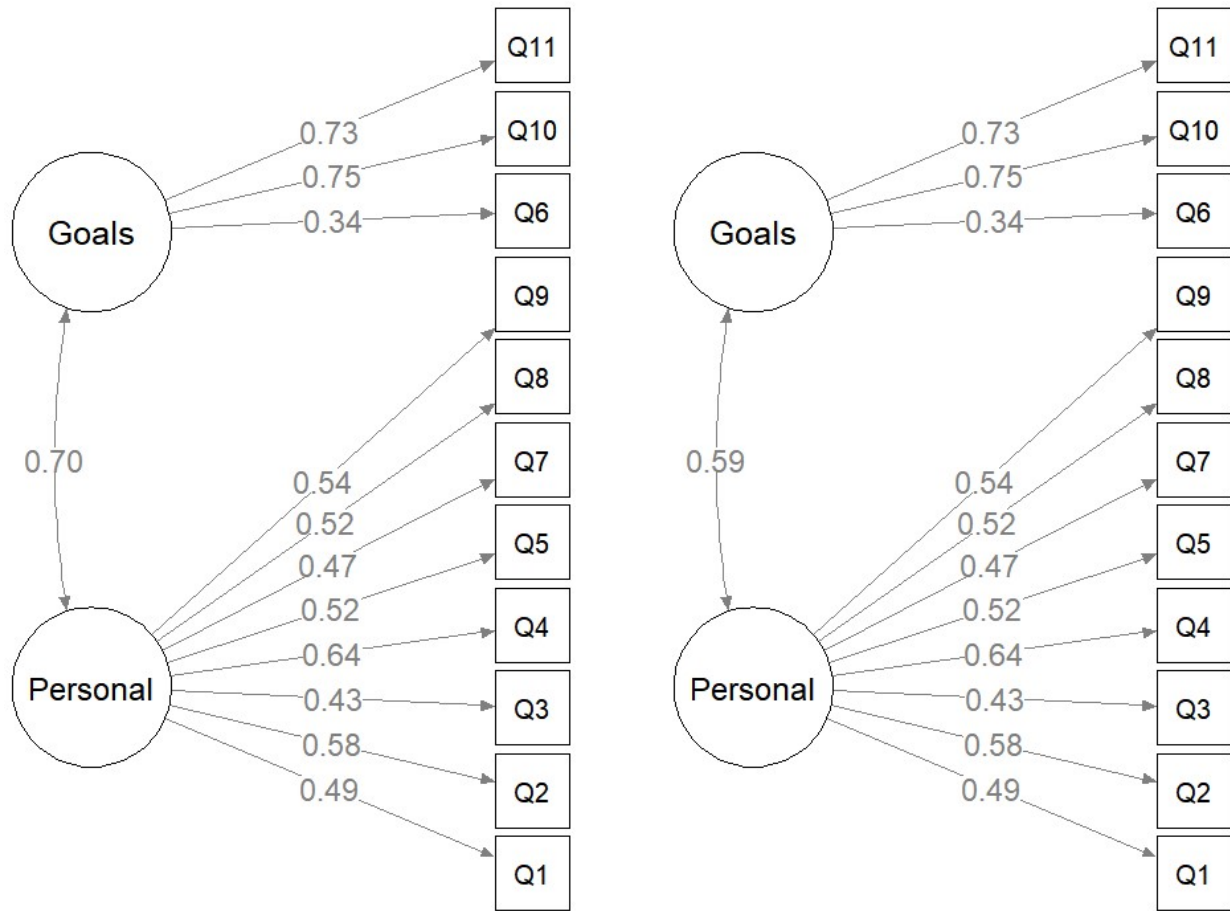


Figure 2: Factor loadings for the equal residual variance MGCFA model using maximum likelihood as the loss function. The left plot depicts the model for the client group, the right plot depicts the model for the counselor group.

An independent samples t-test on the factor scores between client and counselor showed no difference between raw sum scores and summed factor scores for both personal bond ($p = 0.999$) and agreement on tasks and goals ($p = 0.999$; further called "agreement"), so the raw sum scores were used for further computations. In the current sample, 27% of client-counselor pairs differed more than 25% on their ratings of either personal bond, agreement, or on the overall sum score. Most pairs only differed 0-5% in their ratings of personal bond, but the proportion of incongruent ratings within pairs was higher in agreement (Figure 3). The total rating of alliance was again mostly similar within client-counselor pairs, with most pairs only disagreeing 0 to 10% of the time.

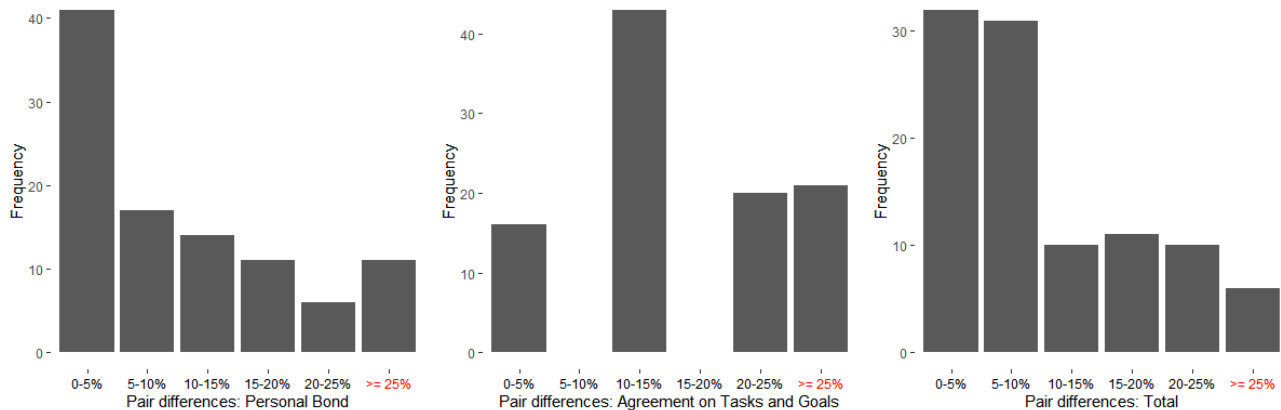


Figure 3: Frequency distribution of the percentages of disagreement within client and counselor pairs on the ratings of personal bond, agreement of tasks and goals, and the overall rating.

Discussion

The first research question pertained to the construct validity of the Online Alliance Assessment, where the specified factor structure was expected to fit the data well. This hypothesis was partly confirmed. Initially, we intended to test whether specified items related to the three components of alliance. However, because one component only consisted of one question, we had to collapse the components "agreement on tasks" and "agreement on goals" into one component, "agreement on tasks and goals". After running, the model fits well, so it is still powerful. The Online Alliance Assessment is a valid measure of alliance, but for the two constructs personal bond and agreement on tasks and goals, not the original three components of alliance we previously intended to test. Even though there was a correlation between the two constructs, it was determined that the two factor model was preferable over a one-factor model. Thus, we validate the two constructs in the questionnaire. Additionally, we find that the two constructs are measured congruently between clients and counselors, confirming our second hypothesis. The two group model has confirmed the presence of the constructs, and the equal factor loadings model confirmed the factors to be equally strong in both groups. In practice, this means that we can see the client and counselor versions of the questionnaire as equivalent: the Online Alliance Assessment is reciprocal.

The final research question was whether clients and counselors differ in their ratings of therapeutic alliance. The equality restrictions in the MGCFA models showed no evidence of differences on a group level. In our current data set, the differences within organisations could not be analysed. However, we did analyse the differences in individual pairs. The majority of pairs only had a small disagreement, between the 0 - 10%, on their overall rating of alliance and on their rating of personal bond. However, we did find larger differences in agreement on tasks and goals. Which peaked at 10 - 15% disagreement and had higher frequencies

between 20 - 25% and above. This agreement is quite important as the counselor risks their client having unrealistic expectations of what needs to be done, which may affect the counseling outcome. This is especially the case in a setting like a single anonymous chat session. If this happens often, or if the MGCFA finds a structural difference in the mean and variance of the responses between groups, the nature of these differences should be investigated. Then, additional employee training could be implemented to improve alliance in future relationships. This way, high quality online counseling is more viable and offers proper support to those in need who cannot get in-person counseling. This can be especially important in a preventative way, where early intervention can prevent the development of more severe problems.

This study has a few limitations. Firstly, the sample consisted of students in a role-playing setting where each student has played both counselors and clients. Thus, the sample is not representative for actual clients and counselors as there is a dependence structure in the data. We cannot account for this dependence as the data have been anonymized, so there is bias in the sample. The analysis can be performed on new data with real counselors and clients, but this does not guarantee similar results. It is possible, that not all factor loadings, item means, and residual variances can be set equal for all questions. In that case, a partial invariance model is required. This allows for equality in most questions while some are allowed to be free. For a guide on how to do this, refer to Buchanan, 2019. Additionally, a current limitation is that the questions of the Online Alliance Assessment are not equally distributed among the factors. The "personal bond" factor contains 8 questions, while "agreement on tasks and goals" contains 3 questions. Thus, we advise to adjust the questionnaire in the future. Preferably with three constructs with an equal amount of questions with a minimum of three questions. This would make the results of the (MG)CFA more reliable when it comes to the factors.

References

- Andersson, G., Titov, N., Dear, B. F., Rozental, A., & Carlbring, P. (2019). Internet-delivered psychological treatments: From innovation to implementation. *World Psychiatry, 18*(1), 20–28. <https://doi.org/10.1002/wps.20610>
- Ardito, R. B., & Rabellino, D. (2011). Therapeutic alliance and outcome of psychotherapy: Historical excursus, measurements, and prospects for research. *Frontiers in Psychology, 2*. <https://doi.org/10.3389/fpsyg.2011.00270>
- Boolytical. (2022). Boolytical - overview. https://github.com/Boolytical/Online_Alliance_Assessment
- Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice, 16*(3), 252–260. <https://doi.org/10.1037/h0085885>
- Buchanan, E. M. (2019). R - multigroup cfa with lavaan example. <https://www.youtube.com/watch?v=t6uHROGalBo>
- Feijt, M., de Kort, Y., Bongers, I., Bierbooms, J., Westerink, J., & IJsselstein, W. (2020). Mental health care goes online: Practitioners' experiences of providing mental health care during the COVID-19 pandemic. *Cyberpsychology, Behavior, and Social Networking, 23*(12), 860–864. <https://doi.org/10.1089/cyber.2020.0370>
- Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the working alliance inventory. *Journal of Counseling Psychology, 36*(2), 223–233. <https://doi.org/10.1037/0022-0167.36.2.223>
- Kim, H., Ku, B., Kim, J. Y., Park, Y.-J., & Park, Y.-B. (2016). Confirmatory and exploratory factor analysis for validating the phlegm pattern questionnaire for healthy subjects. *Evidence-Based Complementary and Alternative Medicine, 2016*, 1–8. <https://doi.org/10.1155/2016/2696019>

- Rosseel, Y. (2022). The lavaan Project. Retrieved May 24, 2022, from <https://lavaan.ugent.be/tutorial/groups.html>
- Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2010). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, 99(6), 323–338. <https://doi.org/10.3200/JOER.99.6.323-338>
- Wagner, B., Horn, A. B., & Maercker, A. (2014). Internet-based versus face-to-face cognitive-behavioral intervention for depression: A randomized controlled non-inferiority trial. *Journal of Affective Disorders*, 152-154, 113–121. <https://doi.org/10.1016/j.jad.2013.06.032>
- Xia, Y., & Yang, Y. (2018). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, 51(1), 409–428. <https://doi.org/10.3758/s13428-018-1055-2>

Appendix A

Table 2: 11 questions of the Online Alliance Assessment for counselors.

Counselor	
1.	Ik begreep de situatie
2.	Ik heb echt geluisterd naar de zorgen
3.	Ik behandelde de hulpclient met respect
4.	Ik speelde goed in op zijn/haar/hun (minderheids-)positie
5.	Ik werkte met de hulpclient samen op basis van vertrouwen
6.	Ik herkende en versterkte wat de hulpclient al goed doet
7.	Ik denk dat de hulpclient mij aardig vond
8.	Ik was echt betrokken op hoe het met de hulpclient gaat
9.	Ik merkte dat de hulpclient mij waardeerde
10.	Ik ondersteunde de hulpclient in wat hij/zij wil bereiken
11.	Ik hielp om actief gebruik te maken van allerlei mogelijkheden om verder te komen

Table 3: 11 questions of the Online Alliance Assessment for clients.

Client	
1.	Degene die mij de online hulp gaf begreep mijn situatie
2.	Degene die mij de online hulp gaf luisterde echt naar mijn zorgen
3.	Degene die mij de online hulp gaf behandelde mij met respect
4.	Degene die mij de online hulp gaf speelde goed in op mijn (minderheids-)positie
5.	Degene die mij de online hulp gaf werkte met mij samen op basis van vertrouwen
6.	Degene die mij de online hulp gaf herkende en versterkte wat ik al goed doe
7.	Degene die mij de online hulp gaf vond mij aardig, geloof ik
8.	Degene die mij de online hulp gaf was echt betrokken op hoe het met me gaat
9.	Degene die mij de online hulp gaf waardeerde mij, merkte ik
10.	Degene die mij de online hulp gaf ondersteunde mij in wat ik wil bereiken
11.	Degene die mij de online hulp gaf hielp me om actief gebruik te maken van allerlei mogelijkheden om verder te komen

Appendix B

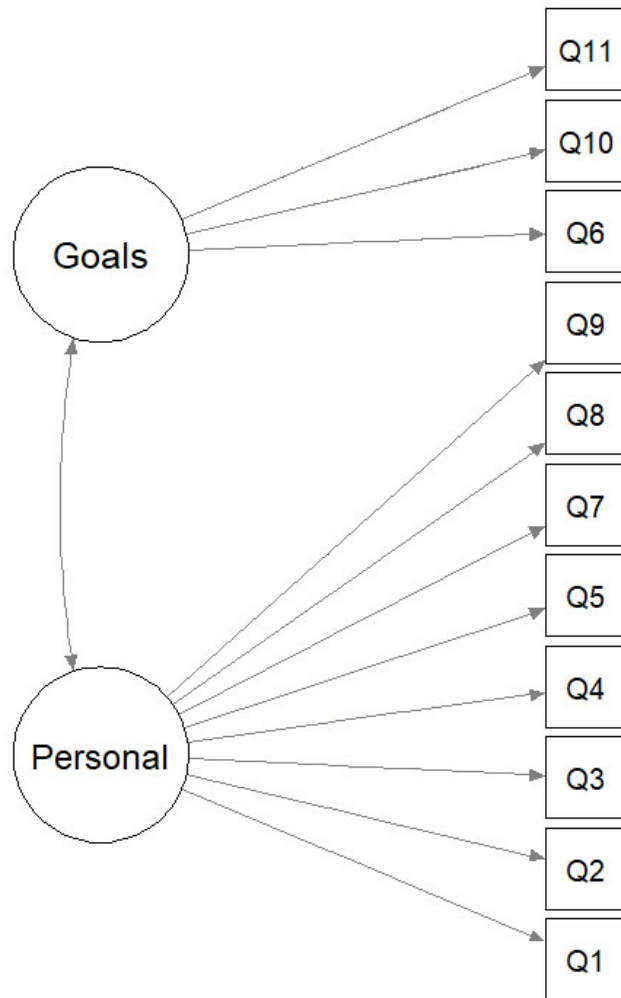


Figure 4: Depiction of the personal bond and agreement on tasks and goals factors used in the MGCFA with their corresponding question numbers.

Appendix C

Table 4: Fit indices for the Multi-Group Confirmatory Factor Analysis with one factor on Online Alliance Assessment. There are four specified models using the Maximum Likelihood (ML) loss function.

	Chi-square (X^2)	df	CFI	RMSEA	SRMR	P-value
MGCFA with ML loss						
Two group model	141.09	88	0.874	0.076	0.068	<0.001
Equal factor loadings	148.62	99	0.882	0.069	0.079	0.755
Equal item means	165.73	109	0.865	0.070	0.084	0.072
Equal residual variance	179.54	120	0.859	0.069	0.091	0.244

Appendix D

Please store all of these files in the same folder.

initial_data_analysis.R file. On line 17 and 18 the SPSS data frame is read into R. When you have a new data set exported from Qualtrics, the information between quotation marks ("Online+hulp...") has to be adjusted to the folder and file name of the new Qualtrics export, starting from the location where the R file was opened. If you keep the Qualtrics exported folder (which contains the data set) in the same folder as the R files, you only need to copy the Qualtrics export folder name and SPSS file name, similar to the provided file.

main_data_analysis.R file. This file can be used to create a correlation matrix of the 11 questions as shown in Figure 1. If you do not want to save this plot as an image, a # should be added in front of line 18 and 20. This changes the lines from code to comments, so they will not be executed. If this is not done, a .png file of the correlation matrix can be found in the folder of the R script.

To run the CFA analysis on the client and counselor group, the parameters on lines 23-27 have to be specified. All possible values for these parameters can be seen in the comment section in R. For the CFA between groups on all data, the selection variable has to be set to "all". In this situation, the ID variable should be specified as NaN. Then, the type variable can be set to either "ML" when you want to use the ML Loss function (continuous data assumed); or "DWLS" when you want to use the the DWLS Loss function (categorical data assumed). For the analysis variable, you can use "configural" when you want the function to return the output of the configural invariance model; "metric" to return the output of the metric invariance model (equal factor loadings); "scalar" when you want to see the results of the scalar invariance model (equal item means); "strict" for the strict

error invariance model output (equal residual variance; or "table_all" when you want to see a table with the output of Table ?? and ??). Lastly, the plot variable can be set to TRUE or FALSE. Where, TRUE will result in a plot of the factor loadings being saved as a .png file while FALSE will not create a plot.

To run a CFA analysis on separate organizations, the selection variable has to be set as "organization". To determine what organization is being selected, the ID variable has to be set as the name of the organization between quotation marks, so that the program recognizes it as a word. For this, first make sure that the organization is spelled in the exact same way for every client and counselor. R will differentiate between "Avans" and "avans". The usage of the type variable and the analysis variable are the same as described above.

For an overview of the paired client-counselor data, the selection variable should be set to "individual", and the ID variable should be set to "all". The other variables can be specified as described above. This will output the proportion of client-counselor pairs that differ 25% or more on either of the components of alliance, or on their overall sum score. R will automatically check if it should use raw sum scores or summed factor scores. You will also get an overview of the ID numbers in which these differences were found, and a plot similar to Figure 3 if you set the plot variable to TRUE. If you keep selection as "individual" but change ID to one of these unique ID numbers, you can obtain a description of the scores per question and sum score for a specific client-counselor pair. For this, no quotation marks should be used around the number, otherwise R will see it as a word instead of a number. As an example, 1643 would work but "1643" would not work. Again, the usage of the type variable and the analysis variable are the same as described above.

When the type variable is set to "DWLS" the following error can be given;

”lavaan ERROR: some categories of variable ‘Q3’ are empty in group 1; frequencies are [19 87 0]”. The variable, group number, or the frequencies between square brackets may vary. This error happens if some answers do not occur at all in a particular question within a group. If this is the case, a line has to be adjusted in the CFA.R. On line xx the text ”##### Alteration on Lavaan error #####” can be found. Below this, the following line of code has to be adjusted; `data = data[data$Q3! = 1,]` in which ”Q3” can be replaced with the variable mentioned in the error. Additionally, the ”1” has to be adjusted for the location of the 0 in the square brackets in the error. Where the left number indicates the number of 3s, the middle number indicates the number of 2s, and the right number is the number of 1s. However, be careful with doing this as this line of code will remove all the data in these rows. If the number of rows is substantial, it is better not to use this method. You could instead opt to collapse answer categories. The number of removed rows, together with the selected method, is printed in the console before the output.

Appendix E

Figure 5 shows what the output of the full MGCFA looks like if you want to compare the fit measures of the continuous MGCFA with the categorical MGCFA. This information was used to create Table 1. The DWLS model (categorical MGCFA) shows how many rows it removes in the analysis. The warning message means that that two models are exactly the same after restricting, in this case the scalar and strict error models for the categorical MGCFA. They are similar, so we proceed with type = "ML" for the continuous MGCFA.

```

R 4.0.2 - ~/GitHub/StatCon/ <>
> selection = "all" # "all" or "organization" or "individual"
> ID = NaN # NaN or organization name or pairing code or "all"
> type = "factor_ML" # "factor_ML" or "variance_DWLS"
> analysis = "table_all" # "configural" or "metric" or "scalar" or "strict" or "table_all"
> plot = FALSE # TRUE or FALSE
>
> # Run the MGCFA and save the plots
> MGCFA(selection, ID, type, analysis, plot)
table_all invariance analysis using factor_ML settings on data; all

|Model|      |X2|      |df|      |CFI|      |RMSEA|      |SRMR|      |p-values|
|-----|-----|-----|-----|-----|-----|-----|-----|
|Configural model| 108.129 | 86 | 0.948 | 0.049 | 0.066 | 0.054 |
|Metric model| 117.268 | 97 | 0.952 | 0.045 | 0.075 | 0.609 |
|Scalar model| 128.36 | 106 | 0.947 | 0.045 | 0.079 | 0.269 |
|Strict Error model| 142.246 | 117 | 0.94 | 0.045 | 0.086 | 0.239 |
>

> type = "variance_DWLS" # "factor_ML" or "variance_DWLS"
> MGCFA(selection, ID, type, analysis, plot)
table_all invariance analysis using variance_DWLS settings on data; all

[1] "Number of rows that will be removed: 2"

|Model|      |X2|      |df|      |CFI|      |RMSEA|      |SRMR|      |p-value|
|-----|-----|-----|-----|-----|-----|-----|-----|
|Configural model| 91.732 | 86 | 0.996 | 0.025 | 0.091 | 0.316 |
|Metric model| 117.3 | 97 | 0.986 | 0.045 | 0.105 | 0.294 |
|Scalar model| 126.925 | 105 | 0.985 | 0.045 | 0.093 | 0.16 |
|Strict Error model| 126.925 | 105 | 0.985 | 0.045 | 0.093 | NA |
Warning message:
In lavTestLRT(object = object, ..., model.names = NAMES) :
lavaan WARNING: some models have the same degrees of freedom
>

```

Figure 5: Output for overviews of continuous MGCFA and categorical MGCFA

If one of the more restrictive models would have fit significantly worse than the previous model (meaning that "p-value" would be <0.05), you could run a partial invariance test. Essentially, you perform the restriction on only part of the questions (for example the mean score per question). You can look in the "Estimate" columns to see if any question is very different from the rest, possibly causing this invariance. If clients and counselors only significantly differ on their average answer to question 1, for example, you could leave this question "free" while still setting the other questions as equal (Buchanan, 2019; Rosseel, 2022). This was not necessary in the current project, but might be necessary in new samples. Running each separate MGCFA model will give a lot of output. R will give separate output for both client and counselor, but we have only included a screenshot of the client group here (Figure 6).

The first "block" of information tells you about the factor loadings. Here, you see the factors "Personal" and "Goals", followed by the items we assigned to them. The column "Std.all" indicates the factor loading of each item onto its factor (these are the straight arrows from the factors to the items in Figure 2. In the second block you see "Covariances: Personal ~ Goals". We have standardized this so that the "Estimate" actually indicates the correlation between the factors, in this case .675. This is the arrow between the factors in Figure 2. Then, you get information about intercepts. The intercepts are calculations of the average score per item. There is also "Variances", indicating the spread of the scores around the average answer score. Finally, the "R-Square" indicates how much of the differences between individuals' responses is explained by each individual question.

Because the configural invariance model does not restrict anything to be equal between groups, the numbers for the counselor group will be different than these numbers, but they are interpreted the same.

Group 1 [client]:						
Latent Variables:						
	Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
Personal =~						
Q1	0.294	0.052	5.623	0.000	0.294	0.558
Q2	0.426	0.062	6.858	0.000	0.426	0.657
Q3	0.220	0.038	5.821	0.000	0.220	0.574
Q4	0.323	0.053	6.040	0.000	0.323	0.592
Q5	0.276	0.059	4.711	0.000	0.276	0.479
Q7	0.215	0.054	3.960	0.000	0.215	0.409
Q8	0.304	0.059	5.111	0.000	0.304	0.514
Q9	0.314	0.053	5.983	0.000	0.314	0.588
Goals =~						
Q6	0.232	0.072	3.211	0.001	0.232	0.342
Q10	0.483	0.061	7.959	0.000	0.483	0.802
Q11	0.465	0.064	7.291	0.000	0.465	0.733
Covariances:						
	Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
Personal ~ Goals	0.675	0.086	7.837	0.000	0.675	0.675
Intercepts:						
	Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
.Q1	2.208	0.051	43.131	0.000	2.208	4.189
.Q2	2.491	0.063	39.579	0.000	2.491	3.844
.Q3	2.821	0.037	75.716	0.000	2.821	7.354
.Q4	2.208	0.053	41.737	0.000	2.208	4.054
.Q5	2.396	0.056	42.716	0.000	2.396	4.149
.Q7	2.387	0.051	46.867	0.000	2.387	4.552
.Q8	2.274	0.057	39.585	0.000	2.274	3.845
.Q9	2.302	0.052	44.298	0.000	2.302	4.303
.Q6	2.311	0.066	35.098	0.000	2.311	3.409
.Q10	2.472	0.058	42.272	0.000	2.472	4.106
.Q11	2.368	0.062	38.433	0.000	2.368	3.733
Personal	0.000				0.000	0.000
Goals	0.000				0.000	0.000
Variances:						
	Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
Personal	1.000				1.000	1.000
Goals	1.000				1.000	1.000
.Q1	0.191	0.030	6.482	0.000	0.191	0.689
.Q2	0.239	0.040	5.934	0.000	0.239	0.568
.Q3	0.099	0.015	6.410	0.000	0.099	0.670
.Q4	0.192	0.030	6.324	0.000	0.192	0.649
.Q5	0.257	0.038	6.757	0.000	0.257	0.771
.Q7	0.229	0.033	6.926	0.000	0.229	0.832
.Q8	0.257	0.039	6.647	0.000	0.257	0.736
.Q9	0.187	0.030	6.347	0.000	0.187	0.655
.Q6	0.406	0.058	7.012	0.000	0.406	0.883
.Q10	0.129	0.040	3.237	0.001	0.129	0.356
.Q11	0.186	0.042	4.426	0.000	0.186	0.463
R-Square:						
	Estimate					
Q1	0.311					
Q2	0.432					
Q3	0.330					
Q4	0.351					
Q5	0.229					
Q7	0.168					
Q8	0.264					
Q9	0.345					
Q6	0.117					
Q10	0.644					
Q11	0.537					

Figure 6: Output for configural invariance model.

Group 1 [client]:							
Latent Variables:							
		Estimate	Std.Err	z-value	P(> z)	Std.lv	Std.all
Personal	≈						
Q1	(.p1.)	0.267	0.038	7.116	0.000	0.267	0.516
Q2	(.p2.)	0.348	0.042	8.382	0.000	0.348	0.565
Q3	(.p3.)	0.195	0.030	6.494	0.000	0.195	0.522
Q4	(.p4.)	0.351	0.039	9.072	0.000	0.351	0.631
Q5	(.p5.)	0.298	0.043	6.908	0.000	0.298	0.510
Q7	(.p6.)	0.257	0.040	6.428	0.000	0.257	0.475
Q8	(.p7.)	0.311	0.043	7.149	0.000	0.311	0.522
Q9	(.p8.)	0.307	0.040	7.583	0.000	0.307	0.578
Goals	≈						
Q6	(.p9.)	0.248	0.053	4.715	0.000	0.248	0.364
Q10	(.10.)	0.454	0.045	10.173	0.000	0.454	0.776
Q11	(.11.)	0.433	0.045	9.588	0.000	0.433	0.703

Figure 7: The metric invariance model sets the factor loadings as equal between groups.

If you run the separate MGCFA models, each model with additional restrictions will have very similar output, with a small difference. The difference between the output of the configural invariance model and the metric invariance model is shown in Figure 7.

As explained in the report, analysis = "metric" sets the factor loadings of the groups as equal. Now, these values will be exactly the same in the client group and the counselor group. This is shown by the row of indicators "(.p1.)" (and so on) in the block with the factor loadings. The rest of the output will look similar to the output in Figure 6, although the numbers will be slightly different as a result of the restriction.

Analysis = "scalar" also sets the intercepts (item means) to be equal. Figure 8 shows how the additional restriction is indicated. All intercept estimates are now exactly the same in the client and counselor groups. Setting analysis = "strict" additionally constrains the variances to be equal across groups (Figure 9).

The strict invariance model (Figure 9) is the most restricted model. If this model (and all of the other models) does not fit significantly worse than its pre-

Intercepts:						
		Estimate	Std.Err	z-value	P(> z)	Std.lv Std.all
.Q1	(.26.)	2.262	0.041	54.933	0.000	2.262 4.374
.Q2	(.27.)	2.536	0.049	51.646	0.000	2.536 4.118
.Q3	(.28.)	2.826	0.032	88.981	0.000	2.826 7.581
.Q4	(.29.)	2.182	0.048	45.716	0.000	2.182 3.904
.Q5	(.30.)	2.349	0.048	48.954	0.000	2.349 3.986
.Q7	(.31.)	2.404	0.043	56.249	0.000	2.404 4.460
.Q8	(.32.)	2.257	0.049	46.540	0.000	2.257 3.779
.Q9	(.33.)	2.284	0.046	49.991	0.000	2.284 4.280
.Q6	(.34.)	2.360	0.052	45.580	0.000	2.360 3.485
.Q10	(.35.)	2.482	0.054	45.676	0.000	2.482 4.262
.Q11	(.36.)	2.343	0.056	41.488	0.000	2.343 3.771
Person1		0.000				0.000 0.000
Goals		0.000				0.000 0.000

Figure 8: The scalar invariance model sets the item means as equal between groups.

vious model, it means that the groups have answered the questionnaire equally (see Figure 5). Their average scores and answer ranges on each question are the same, and the importance of each question to its factor is also equal between the groups. This set of analyses can for example be used when comparing client and counselor ratings on organization level. Then, analysis = "table.all" gives a quick overview of whether the sequentially restricted models all fit well (and if they do, client and counselor groups within the organization rate alliance the same), and the separate MGCFA analyses give more detailed information.

Variances:						
		Estimate	Std.Err	z-value	P(> z)	Std.lv Std.all
Person1		1.000				1.000 1.000
Goals		1.000				1.000 1.000
.Q1	(.14.)	0.206	0.022	9.399	0.000	0.206 0.765
.Q2	(.15.)	0.233	0.026	8.826	0.000	0.233 0.663
.Q3	(.16.)	0.144	0.015	9.615	0.000	0.144 0.812
.Q4	(.17.)	0.183	0.022	8.284	0.000	0.183 0.589
.Q5	(.18.)	0.262	0.028	9.199	0.000	0.262 0.726
.Q7	(.19.)	0.231	0.024	9.456	0.000	0.231 0.777
.Q8	(.20.)	0.267	0.029	9.238	0.000	0.267 0.733
.Q9	(.21.)	0.233	0.026	9.092	0.000	0.233 0.707
.Q6	(.22.)	0.409	0.042	9.822	0.000	0.409 0.882
.Q10	(.23.)	0.152	0.029	5.322	0.000	0.152 0.431
.Q11	(.24.)	0.178	0.030	5.996	0.000	0.178 0.474

Figure 9: The strict invariance model sets the item variances as equal between groups.

After checking for invariance, you can keep the most restrictive model that fits well selected. In our case, we keep analysis set to "strict". If you then set the selection to "individual" and ID to "all" you get the output as shown in Figure 10.

```
strict invariance analysis using factor_ML settings on data; individual
Client and counselor ratings per question for counselor ID: all
Set "plot = TRUE" for a graphical overview of pairwise percentages of rating differences.
27 % of the client-counselor pairs differ at least 25% in (one or more of) the ratings of Personal Bond,
Agreement on tasks and goals, or the total ratings.
These differences were found in the following pairs:
```

ID
1099
1457
2117
2585
3284
3519
3804
3886
4680
5558
5619
5978
6003
6115
6339
6478
6629
6994
7077
7391
7395
7618
7783
8081
8154
8317
8904

Figure 10: Output of the overview client-counselor pair differences.

If plot is set to TRUE, this will additionally give you a plot similar to Figure 3 inside the RStudio plot window. All of the pairs whose ID show a large difference in the summed responses of either of the components, or the overall sum score. If you keep selection as "individual" and change ID to, for example, 1099, you get an overview of the original answers on each question, the summed scores per component and in total, and the proportion of "satisfaction" per component and in total (Figure 11).

strict invariance analysis using factor_ML settings on data; individual
 Client and counselor ratings per question for counselor ID: 1099

	Client	Counselor	
Q1	1	3	
Q2	1	3	
Q3	2	3	
Q4	1	2	
Q5	2	1	
Q6	3	3	
Q7	2	2	
Q8	2	2	
Q9	1	2	
Q10	1	2	
Q11	1	1	
Sum Personal	12	18	
Sum Goals	5	6	
Total	17	24	
% rating Personal	50%	75%	25% difference
% rating Goals	55.56%	66.67%	11.11% difference
% rating Total	51.52%	72.73%	21.21% difference

Figure 11: Output of client-counselor pair differences for an individual pair.

In short, it is easy to get a quick overview of the pairs that highly disagree on their rating of the therapeutic alliance and on which components and statements they disagree. This information can then be used to provide feedback to individual counselors on aspects they could work to improve.