Reliability and Validity Testing of the Chinese Version of the Mobile Agnew Relationship Measure (mARM–C)

Authors: Moran, Mo Ran

Date: 2023-05-27T00:00:00+00:00

Abstract

To examine the reliability and validity of the Chinese version of the Mobile Agnew Relationship Measure (mARM-C). The mARM-C and criterion questionnaires were administered to 574 college students who had recently used meditation apps, with 102 participants selected for retesting two weeks later. Exploratory factor analysis and network analysis revealed that the questionnaire comprises five factors with a total of 19 items; confirmatory factor analysis also demonstrated good model fit, and criterion-related validity, convergent validity, discriminant validity, and internal consistency reliability all met psychometric standards. Therefore, the mARM-C exhibits good reliability and validity and can be utilized to measure the digital therapeutic alliance between users and programs in online self-help interventions.

Full Text

Reliability and Validity of the Chinese Version of the Mobile Agnew Relationship Measure (mARM-C)

Mo Ran
(1. Department of Psychology, Faculty of Education, Guangxi Normal University, Guilin 541006, China)

Abstract: This study examined the reliability and validity of the Chinese version of the Mobile Agnew Relationship Measure (mARM-C). A total of 574 university students who had recently used meditation apps completed the mARM-C and criterion measures, with 102 participants retested after two weeks. Exploratory factor analysis and network analysis revealed a five-factor structure comprising 19 items. Confirmatory factor analysis demonstrated good model fit, and the questionnaire exhibited satisfactory criterion-related validity, convergent validity, discriminant validity, and internal consistency reliability, meeting

established psychometric standards. These results indicate that the mARM-C is a reliable and valid instrument for measuring the digital therapeutic alliance between users and programs in internet-based self-help interventions.

Keywords: mobile Agnew relationship measure, reliability, validity, network analysis, digital therapeutic alliance

In recent years, Internet-based Self-help Interventions (ISIs) have garnered significant attention due to their cost-effectiveness, scalability, and flexibility, with their feasibility and effectiveness widely validated as a promising complement to face-to-face counseling and psychotherapy (Izzaty et al., 2021; Johansson et al., 2021). However, higher levels of automation also mean reduced human support and supervision, which can lead to low user engagement and limit the effectiveness of ISIs (Pratap et al., 2020). Consequently, addressing these challenges has become a key research trend in the field.

Therapeutic Alliance (TA) refers to the quality and strength of the collaborative relationship between client and counselor in pursuit of therapeutic goals (Zhu & Jiang, 2011) and represents one of the most robust predictors of treatment outcomes (Flückiger et al., 2018). Given its central role in traditional therapy, the concept has been adapted to digital mental health contexts as the Digital Therapeutic Alliance (DTA; Henson et al., 2019), defined as the quality and strength of the therapeutic relationship formed between users and digital programs. Since DTA can predict both user engagement and intervention effectiveness in ISIs (Goldberg et al., 2021), research in this area has gradually expanded. For example, Henson et al. (2019) adapted the Working Alliance Inventory-Short Revised (WAI-SR) to create a six-item unidimensional scale—the Digital Working Alliance Inventory (DWAI). Subsequently, Goldberg et al. (2021) validated the DWAI's reliability and validity across two studies. However, the DWAI was not developed through a rigorous process and contains few items, potentially limiting its ability to comprehensively assess DTA. In contrast, the Mobile Agnew Relationship Measure (mARM), systematically developed based on the Agnew Relationship Measure (ARM), comprises five factors with 25 items (Berry et al., 2018) and represents the most comprehensive DTA questionnaire currently available (D'Alfonso et al., 2020). Its psychometric properties have also been examined in recent empirical interventions (Wulffen et al., 2022).

Currently, research on DTA in China remains in its early stages (Mo et al., in press; Mo et al., 2023), and validated measurement tools are lacking. Moreover, most ISI studies have simply adapted traditional therapeutic alliance measures (e.g., WAI-SR) by superficially replacing "counselor" with "program" (Berry et al., 2018). Therefore, revising the mARM to provide a reliable measurement tool for future ISI research in China is warranted.

2.1 Participants and Procedure

The study proceeded through four distinct phases. First, for pre-test questionnaire development, we obtained authorization from Professor Berry, the

developer of the mARM, and conducted forward-backward translation considering Chinese cultural context and linguistic conventions. Under the guidance of an associate professor of psychology, two psychology graduate students independently translated the questionnaire. Two English graduate students then back-translated the mARM-C into English. Finally, items with ambiguous or unclear wording were revised, yielding the pre-test questionnaire.

Second, in the pilot phase, we distributed 180 questionnaires through the Credamo platform (https://www.credamo.com/#/), obtaining 144 valid responses (80.00% response rate) as Sample 1. This sample comprised 53 males (36.8%) and 91 females (63.2%) with a mean age of 28.84 years (SD = 8.37). Sample 1 data were then subjected to item analysis and exploratory factor analysis, with relevant items modified or deleted based on results to form the formal questionnaire.

Third, during formal testing, we used convenience sampling to recruit 756 university students from five universities in Guangxi to experience a mindfulness app for two weeks and complete an online survey, yielding 574 valid responses (75.60% response rate) as Sample 2. This sample included 114 males (19.9%) and 460 females (80.1%) with a mean age of 19.23 years (SD = 1.81). Sample 2 was then split into two databases by odd/even numbers, with one database (Sample 3, n = 287) randomly selected for exploratory factor analysis and the other (Sample 4, n = 287) for confirmatory factor analysis.

Finally, in the retest phase, 120 participants from Sample 2 were randomly selected to complete the questionnaire again after a two-week interval, producing 102 valid responses (85.00% response rate) as Sample 5.

2.2 Instruments

Mobile Agnew Relationship Measure (mARM). Developed by Berry et al., this questionnaire includes five factors: Partnership (PRS), Openness (OPN), Bond (BD), Confidence (CONF), and Client Initiative (CI), with 25 items total (Berry, Salter, Morris, James, & Bucci, 2018). It uses a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree), with higher scores indicating higher quality DTA between users and programs.

Digital Working Alliance Inventory (DWAI). Developed by Henson et al. for measuring DTA (Henson, Wisniewski, Hollis, Keshavan, & Torous, 2019), the Chinese version of the DWAI has demonstrated good reliability and validity (Zhao, 2022). This six-item scale uses a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree). In the present study, its internal consistency coefficient was 0.92.

Client Satisfaction Questionnaire (CSQ-8). Developed by Larsen et al. (1979) to evaluate user satisfaction with digital mental health services, the Chinese version of the CSQ-8 has good psychometric properties (Zhao, 2022). This eight-item questionnaire uses a 4-point Likert scale with scores ranging

from 8 to 32, where higher scores indicate greater satisfaction. In this study, its internal consistency coefficient was 0.89.

Trust of Counseling Scale (TCS). Developed by Zhao, Jiang, and Gu (2011), this scale comprises five dimensions with 36 items using a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree). For research purposes, we selected the "Practitioner Ability" subscale (TCS-A) as a criterion measure. In this study, this subscale's internal consistency coefficient was 0.94.

2.3 Statistical Methods

We used SPSS 26.0 for item analysis, exploratory factor analysis, criterion-related validity analysis, and reliability analysis; JASP 17.2 for network analysis; and Amos 26.0 for confirmatory factor analysis.

3.1 Pre-Test Questionnaire Analysis

Item analysis of Sample 1 (N = 144) revealed that all item-total correlations were statistically significant ($p < 0.001$), ranging from 0.47 to 0.76 and exceeding the 0.30 threshold. Participants were then ranked by total score, with the top 27% designated as the high-scoring group and the bottom 27% as the low-scoring group. Independent samples t-tests showed significant differences between groups for 24 items ($p < 0.001$), while Item 20 showed no significant difference ($p > 0.05$) and was consequently deleted.

Exploratory factor analysis using principal component analysis with orthogonal rotation was conducted on the 25 items. Results indicated a KMO value of 0.92 and a Bartlett's test ^{2} value of 1577.65 ($p < 0.001$), confirming suitability for factor analysis. The scree plot suggested four factors before the inflection point, with cumulative variance explained at 39.93%, which did not match the original five-factor structure. Given the small pilot sample size and considering the theoretical importance of items and the integrity of the questionnaire structure, we did not delete items directly but instead revised items 16, 18, 21, 23, and 25 to form the formal questionnaire.

3.2 Formal Questionnaire Analysis

Item analysis of Sample 3 (n = 287) showed item-total correlations ranging from 0.60 to 0.81, all significant at the 0.001 level. After ranking participants by total score and designating the top and bottom 27% as high and low groups respectively, independent samples t-tests revealed that 24 items demonstrated significant differences at the 0.001 level, indicating good discriminability (see Table 1).

Exploratory factor analysis using principal component analysis with orthogonal rotation was then performed on the remaining items from Sample 3 (n = 287). Results showed a KMO value of 0.94 and a Bartlett's test ^{2} value of 4988.02 (p

< 0.001), confirming suitability for factor analysis. Following the original questionnaire structure, we conducted a five-factor analysis, revealing factor loadings ranging from 0.595 to 0.819 and total variance explained at 49.49%. However, items 12, 15, and 17 did not align with their original factor assignments. We therefore employed the novel network analysis technique to create a network diagram (Hevey, 2018) to reassess item-factor associations. Exploratory Graph Analysis (EGA) generates network diagrams that visually display which items cluster together and their connection strength (Golino et al., 2020). In the item-factor network diagram, blue lines indicate positive correlations (red lines indicate negative correlations), with thicker lines representing stronger connections between nodes (Figure 1 [Figure 1: see original paper]). Results showed that Bond items 12 (BD1), 15 (BD2), and 17 (BD3) overlapped with Partnership items, making them difficult to distinguish. We subsequently deleted items that did not match the original factor structure or showed cross-loadings (items 6, 12, 15, 17, 18). After re-running exploratory factor analysis and network analysis on the remaining 19 items, results aligned with the original questionnaire structure, with factor loadings ranging from 0.491 to 0.844 and total variance explained at 51.19% (see Table 2). Network analysis also revealed an improved item-factor network structure (Figure 2 [Figure 2: see original paper]). Item centrality analysis showed that BD5, PRS3, OPN3, and CONF4 had high centrality, while factor centrality analysis revealed that the Bond (BD) node had the highest centrality (Closeness = 1.275) (Figure 3 [Figure 3: see original paper]). Nodes with more numerous and stronger connections are considered central, and intervening on these nodes can directly influence others (Mullarkey et al., 2019). Psychopathological network theory also suggests that targeting high-centrality nodes is more efficient and beneficial for optimizing the entire psychometric network (Borsboom & Cramer, 2013; Chen et al., 2021). Based on these findings, we established the final mARM-C comprising five factors and 19 items.

[Figure 1: see original paper]
[Figure 2: see original paper]
[Figure 3: see original paper]

Confirmatory factor analysis using Sample 4 (n = 287) demonstrated good model fit across all indices (^{2} = 385.035, df = 142, ^{2}/df = 2.712, $p < 0.001$, SRMR = 0.046, RMR = 0.040, RMSEA = 0.077, IFI = 0.940, TLI = 0.927, CFI = 0.940), indicating stable structure for the revised questionnaire.

Criterion-related validity analysis using Sample 4 (n = 287) examined correlations between mARM-C total and subscale scores with the DWAI, CSQ-8, and TCS-A. Results showed significant positive correlations between mARM-C total and subscale scores with all criterion measures. Additionally, the five mARM-C dimensions correlated highly with the total score (r = 0.784–0.889, $p < 0.01$), while inter-dimensional correlations were lower (r = 0.507–0.692, $p < 0.01$) (see Table 3).

Convergent and discriminant validity were assessed using Sample 4 (n = 287)

data through Average Variance Extracted (AVE) and Composite Reliability (CR) values. Results showed all AVE values exceeded 0.5 and all CR values exceeded 0.7, indicating good convergent validity (see Table 4). Discriminant validity was verified using the HTMT method, with a maximum value of 0.76 (< 0.85), indicating good discriminant validity.

Reliability analysis using Sample 4 ($n = 287$) revealed an internal consistency coefficient of 0.95 for the mARM-C total score, with subscale coefficients ranging from 0.87 to 0.94. Split-half reliability was 0.87 for the total score and 0.78–0.88 for subscales. Test-retest reliability using Sample 5 ($N = 102$) showed a total score reliability of 0.95 and subscale reliabilities of 0.79–0.88.

This study revised the mARM to create the mARM-C and examined its psychometric properties. We integrated item analysis, exploratory factor analysis, network analysis, and confirmatory factor analysis to demonstrate that the mARM-C meets psychometric standards.

In the pilot test, item analysis revealed that Item 20 did not show significant differences between high and low groups ($p > 0.05$) and was deleted. Exploratory factor analysis yielded a four-factor structure that diverged from the original five-factor model, and some items had ambiguous semantics. However, based on the theoretical foundation of the original questionnaire, we retained corresponding items and revised items 16, 18, 21, 23, and 25 to better fit the Chinese cultural context, forming the formal questionnaire.

In the formal testing, the mARM-C demonstrated strong performance in item analysis, with all high-low group comparisons reaching significance. Item-total correlations for the 24 items ranged from 0.60 to 0.81, indicating good discriminability. Notably, this study innovatively employed network analysis as a complement to exploratory factor analysis, clearly visualizing item and factor associations and using centrality indices to evaluate item importance within factors (Mullarkey et al., 2019), providing a valuable approach for future research. Combined results from exploratory factor analysis and network analysis revealed that items 12, 15, and 17 did not align with their original factor assignments and their content did not fit the intended dimensions. After deleting items 6, 12, 15, 17, and 18, both exploratory factor analysis and network analysis showed good fit, supporting a five-factor structure with 19 items. Moreover, the Bond dimension occupied a central position in the factor network, suggesting it has greater potential to influence DTA development, which aligns with previous research emphasizing the critical role of emotional bonds in unguided ISIs (Mo et al., 2023). Future interventions could therefore strengthen relational cues to enhance Bond levels and more efficiently develop DTA. Confirmatory factor analysis further supported the cross-cultural stability of the mARM-C with good model fit. Additionally, significant positive correlations with criterion measures, including a strong correlation of 0.84 with the DWAI (> 0.7), demonstrated good criterion-related validity. Finally, internal consistency, split-half reliability, and test-retest reliability all exceeded 0.85, indicating excellent consistency and stability.

In summary, the revised mARM-C comprises five factors and 19 items, demonstrating good reliability and validity for measuring DTA between users and programs in ISIs.

References

Chen, C., Wang, L., Cao, C., & Li, G. (2021). Psychopathological network theory: Methods and challenges. *Advances in Psychological Science*, 29(10).

Mo, R., Fang, J., & Chang, B. (in press). From "anthropomorphic attribution" to "alliance establishment": The influence of human-chatbot relationships on engagement. *Advances in Psychological Science*.

Mo, R., Fang, Z., & Fang, J. (2023). How to establish a digital therapeutic alliance between chatbots and users: The role of relational cues. *Advances in Psychological Science*, 31(4), 669–683.

Zhao, C. (2022). *Mobile network-based ACT intervention for PTSD: Development, effectiveness, mechanisms, and matching* (Doctoral dissertation). Central China Normal University, Wuhan.

Zhao, L., Jiang, G., & Gu, Q. (2011). Development and validation of the Trust of Counseling Scale for university students. *Chinese Mental Health Journal*, 25(3), 175–179.

Zhu, X., & Jiang, G. (2011). The concept of working alliance. *Chinese Journal of Clinical Psychology*, 19(2), 275–280.

Berry, K., Salter, A., Morris, R., James, S., & Bucci, S. (2018). Assessing therapeutic alliance in the context of mHealth interventions for mental health problems: Development of the mobile agnew relationship measure (mARM) questionnaire. *Journal of Medical Internet Research*, 20(4), 1–8.

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, 9(1), 91–121.

D'Alfonso, S., Lederman, R., Bucci, S., & Berry, K. (2020). The digital therapeutic alliance and human-computer interaction. *JMIR Mental Health*, 7(12), e21895. https://doi.org/10.2196/21895

Flückiger, C., Del, A. C., Wampold, B. E., & Horvath, A. O. (2018). The Alliance in Adult Psychotherapy: A Meta-Analytic Synthesis. *Psychotherapy*, 55(4), 316–340.

Goldberg, S. B., Baldwin, S. A., Riordan, K. M., Torous, J., Dahl, C. J., Davidson, R. J., & Hirshberg, M. J. (2021). Alliance with an unguided smartphone app: Validation of the digital working alliance inventory. *Assessment*, 29(6), 1331–1345. https://doi.org/10.1177/10731911211015310

Golino, H., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., … Martinez-Molina, A. (2020). Investigating the performance of exploratory

graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychological Methods*, 25(3), 292–320.

Henson, P., Wisniewski, H., Hollis, C., Keshavan, M., & Torous, J. (2019). Digital mental health apps and the therapeutic alliance: Initial review. *BJPsych Open*, 5(1), 1–5.

Izzaty, R. E., Astuti, B., & Cholimah, N. (2021). Evaluation of an online cognitive behavioural therapy, mindfulness meditation, and yoga (CBT-MY) intervention for posttraumatic stress disorder: A single arm clinical trial with psychometric and psychophysiological outcomes. *Angewandte Chemie International Edition*, 6(11), 951–952.

Johansson, M., Berman, A. H., Sinadinovic, K., Lindner, P., Hermansson, U., & Andréasson, S. (2021). Effects of internet-based cognitive behavioral therapy for harmful alcohol use and alcohol dependence as self-help or with therapist guidance: Three-armed randomized trial. *Journal of Medical Internet Research*, 23(11), e29666. https://doi.org/10.2196/29666

Larsen, D. L., Attkisson, C. C., Hargreaves, W. A., & Nguyen, T. D. (1979). Assessment of client/patient satisfaction: Development of a general scale. *Evaluation and Program Planning*, 2(3), 197–207.

Lederman, R., & D'Alfonso, S. (2021). The digital therapeutic alliance: Prospects and considerations. *JMIR Mental Health*, 8, 1–4.

Mullarkey, M. C., Marchetti, I., & Beevers, C. G. (2019). Using network analysis to identify central symptoms of adolescent depression. *Journal of Clinical Child & Adolescent Psychology*, 48(4), 656-668.

Pratap, A., Neto, E. C., Snyder, P., Stepnowsky, C., Elhadad, N., Grant, D., … Omberg, L. (2020). Indicators of retention in remote digital health studies: A cross-study evaluation of 100,000 participants. *Npj Digital Medicine*, 3(1), 1–10.

Sun, S., Lin, D., Goldberg, S., Shen, Z., Chen, P., Qiao, S., … Operario, D. (2021). A mindfulness-based mobile health (mHealth) intervention among psychologically distressed university students in quarantine during the COVID-19 pandemic: A randomized controlled trial. *Journal of Counseling Psychology*, 69(2), 157–171.

Wulffen, C. von, Marciniak, M. A., Rohde, J., Kalisch, R., Binder, H., Tuescher, O., & Kleim, B. (2022). German version of the mobile Agnew Relationship Measure (mARM-G): Translation validation study. *PsyArXiv*, Advance online publication. https://doi.org/10.31234/osf.io/76gey

Note: Figure translations are in progress. See original paper for figures.