Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Early Working Alliance Mediates the Association between Engagement and Outcomes among Users of an AI Mental Health Conversational Agent: A Secondary Hierarchical Clustering Analysis of a Randomized Controlled Trial

Valerie L. Forman-Hoffman

valerie@soniahealth.com

Sonia Health

Nico Roever[1,2]

Sonia Health

Jonathan Winter[3]

Tel Aviv University

Matan Rubin[4]

Hebrew University of Jerusalem

Dustin Klebe[1]

Sonia Health

Lukas Wolf[1]

Sonia Health

Jie-Ming Li[1]

Sonia Health

Patricia Arean[1]

Sonia Health

Article

Keywords: digital mental health, engagement, working alliance, AI conversational agents, cluster analysis, mediation, moderation

Posted Date: August 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10553887/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations:
Statement of Ethics Approval: This study was reviewed and approved by the Ethics Review Board of the Faculty of Management, Economics and Social Sciences, University of Cologne (approval reference: 240088NR; approved March 11, 2025).Statement on Participant Consent: All participants provided informed consent electronically prior to enrollment. Consent was collected in two parts during the baseline survey: a general consent to the processing of personal data for research purposes (including special categories of data under GDPR Art. 9(2)(a)), and a study-specific consent form informing participants that they would be asked about their mental health, that participation was voluntary and could be withdrawn at any time without consequence, and providing crisis-line resources. Participants assigned to the AI intervention arm additionally viewed and acknowledged an in-app consent screen before first use, which reiterated that the app is AI-based and not a substitute for professional care. No identifiable participant data are reported in this manuscript; all results reflect de-identified, aggregated data.Data Availability: The raw, de-identified data generated in this study are available at: https://www.google.com/url?q=https://doi.org/10.5281/zenodo.21979397&source=gmail&ust=1787083007964000&sa=E"

Competing interest reported. VFH, DK, LW, and JL are employees of Sonia Health. PA is a paid consultant of Sonia Health. NR, JW and MR have no competing interests to declare.

Early Working Alliance Mediates the Association between Engagement and Outcomes among Users of an AI Mental Health Conversational Agent: A Secondary Hierarchical Clustering Analysis of a Randomized Controlled Trial

Valerie L. Forman-Hoffman, PhD, MPH[1], Nico Roever[1,2], Jonathan Winter[3], Matan Rubin[4], Dustin Klebe[1], Lukas Wolf[1], Jie-Ming Li[1], Patricia Arean[1]

[1] Sonia Health, San Francisco, CA, USA
[2] University of Cologne
[3] Coller School of Management at Tel Aviv University
[4] Hebrew University of Jerusalem

Corresponding author: Valerie Hoffman, PhD, MPH | valerie@soniahealth.com

Target journal: npj Mental Health Research
Date: July 30, 2026
Word count: 4894

ABSTRACT

Word count: 301

Background: Users of digital mental health tools engage in heterogeneous ways. Working alliance, a significant correlate of symptom improvement, may operate as a mechanism, linking engagement to improvement, but this relationship has not been examined to date.

Objective: To characterize engagement patterns of users of an AI mental health conversational agent, *Sonia,* examine the associations between engagement cluster pattern and outcomes, and test whether working alliance mediates or moderates behavioral engagement-outcome associations.

Methods: Hierarchical clustering analysis was applied to data from participants randomized to *Sonia* in a four-week RCT (n=200), using 37 behavioral engagement variables and 3 Working Alliance Inventory (WAI) subscale scores (Goal, Bond, and Task) assessed weekly post-randomization examined as indicators of affective engagement. Outcomes included depressive and anxiety symptom minimum clinically important difference (MCID), user satisfaction, and product recommendation ratings. Mediation (Week 4 WAI change) and moderation (Week 1 WAI change) analyses examined WAI overall and its 3 subscales at Week 1 as mechanistic pathways.

Results: Four engagement clusters of the 40 engagement metrics were identified and validated. Cluster membership was associated with satisfaction and likelihood to recommend (~14-point differences, FDR $q < 0.05$) but not depressive or anxiety symptom MCIDs (all FDR $q > 0.08$). All four WAI metrics significantly mediated the engagement-outcome relationship (8/8 models; indirect effects −0.036 to −0.062, all 95% CIs excluding zero). WAI Composite and Goal were the most consistent mediators; WAI Task was the most robust across measurement timepoints. WAI level, but not WAI change over time, predicted outcomes. No WAI subscale moderated engagement effects.

Conclusions: Early working alliance established within the first week of app use mediates the pathway from behavioral engagement to symptom improvement in an AI conversational agent. These findings suggest that engagement only affects outcomes via the establishment of working alliance; onboarding experiences that promote alliance formation may be more consequential for clinical outcomes than interventions aimed at increasing sustained engagement volume.

Keywords: digital mental health, engagement, working alliance, AI conversational agents, cluster analysis, mediation, moderation

1. INTRODUCTION

1.1 The Engagement Paradox in Digital Mental Health

Engagement heterogeneity is pervasive in digital mental health: most users do not engage uniformly, with a substantial proportion disengaging early (Perski et al., 2017; Torous et al., 2018). This heterogeneity creates an interesting challenge for research investigating the association between engagement and outcomes, which, analytically, typically aggregates findings across all users. Examining findings in aggregate precludes the ability to identify meaningful subgroup differences in how and whether digital app users benefit differently according to their engagement level and patterns.

Prior research has repeatedly demonstrated that behavioral metrics of engagement such as session frequency, total app opens, or minutes of use do not consistently predict who improves (Linardon et al., 2025). Critically, the field has conflated behavioral persistence with therapeutic benefit, overlooking the possibility that the quality of engagement, rather than its quantity, may determine outcome. This paradox calls for mechanistic investigation: not whether engagement produces improvement, but how it might do so.

Dropout patterns further complicate interpretation. In traditional therapy, early termination encompasses two qualitatively distinct phenomena: patients who disengage due to poor alliance or poor fit, and those who disengage because they have achieved their goals (Brown et al., 2015). Aggregating these groups produces misleading estimates of dropout effects. Similar dynamics have been documented in technology-mediated psychotherapy platforms, including SMS-based services most analogous to AI conversational agents (or chatbots), where disengagement and improvement co-occur in a U-shaped pattern (Darnell et al., 2022). Even brief mobile app use has been associated with clinical benefit in some studies (Arean et al., 2016), though the "good-enough artifact" complicates dose-response interpretations because fast responders tend to terminate earlier (Robinson et al., 2020).

Taken together, these findings suggest that engagement must be understood through the variation within it by incorporating both the distinct patterns of use and the qualities that separate them that are often hidden by aggregate metrics. Understanding engagement in digital mental health may require moving beyond simple usage metrics to examine the mechanisms through which engagement may generate benefit.

1.2 Engagement Clustering as a Tool for Characterizing Heterogeneity

One potential solution to better understand heterogeneity and consider engagement metrics beyond behavioral ones is to use cluster analysis. This technique offers a data-driven approach to surfacing distinct engagement subgroups whose different outcome trajectories are obscured by examining only effects on average (Lee et al., 2014). Rather than treating engagement as a continuous variable, cluster-analytic approaches identify naturally occurring patterns of behavioral and experiential data that may correspond to different mechanisms of action,

different populations, or different needs for product adaptation.

Prior engagement clustering work in digital mental health has largely relied on behavioral metrics alone such as session frequency and number of logins and/or feature use, failing to incorporate the affective and cognitive dimensions of engagement that theory and evidence suggest are most critical to outcomes (Hoffman et al., 2023). This represents a gap: behavioral persistence is a necessary but likely insufficient condition for intervention benefit. Approaches that integrate relational and experiential engagement indicators may better characterize who benefits and why.

1.3 Working Alliance in Tools Delivered via AI

Working alliance is among the most consistent predictors of outcome across treatment modalities in human psychotherapy, with meta-analytic effect sizes in the moderate range that are robust across treatment type, outcome measure, and population (Flückiger et al., 2018; Baldwin et al., 2007; Zilcha-Mano, 2017). The Working Alliance Inventory (WAI) and WAI-Short Report (WAI-SR) operationalize alliance across three theoretically distinct subscales: Goal (agreement on the preferred outcomes), Task (agreement on the methods and activities used to achieve the Goal), and Bond (the quality of the emotional connection between client/user and therapist/AI; Horvath & Greenberg, 1989; Hatcher & Gillaspy, 2006).

Emerging evidence suggests that AI systems can elicit meaningful working alliance in studies that have measured alliance using the WAI Short-Revised (WAI-SR) (Darcy et al., 2021; Beatty et al., 2022). Typically examined as a dependent variable, no prior study has examined WAI subscales as mechanistic pathways in how AI tools lead to preferred outcomes. Thus, the question of which subscale, if any, drives outcome in AI contexts remains open.

The distinction matters theoretically. In human-delivered interventions, task and bond components are typically highly correlated, but their relative contributions to outcome vary across treatment types (Horvath & Bedi, 2002). In AI contexts, the emotional bond may be qualitatively different from a human relationship and potentially more salient to user experience than to symptom change. Task agreement, which reflects the user's endorsement of the specific methods the AI recommends, may be more proximally linked to whether users actually engage with the cognitive and behavioral solutions proposed by the AI. A recently published systematic review and meta-analysis found that task completion, rather than engagement volume, was associated with clinical outcomes in smartphone apps for depression and anxiety (Linardon et al., 2025), consistent with this hypothesis.

Importantly, prior work on alliance formation in both human and digital interventions suggests that alliance is established early and tends to remain stable thereafter, rather than developing linearly over time (Flückiger et al., 2018). This stability has implications for both measurement timing and mechanism: if alliance is established rapidly and then maintained, its effect on clinical outcomes reflects the initial quality of the relationship fit rather than an ongoing developmental process or result of engagement. The present study was designed to test this

interpretation directly.

1.4 The Current Study

This paper reports a secondary analysis of a four-week RCT of Sonia, a conversational AI emotional support application. Participants were randomized to Sonia or a control condition; this analysis focuses exclusively on the Sonia arm (n = 200). The study addresses two primary research questions:

What distinct engagement profiles, built of combinations of both behavioral and affective engagement metrics, exist among Sonia users? What engagement and clinical outcomes do these profiles predict? What baseline characteristics predict cluster membership?

Do WAI scores, either totaled or of specific subscales (Goal, Task, and/or Bond), assessed at one week post-randomization, mediate or moderate the relationship between behavioral engagement and four-week outcomes?

A key methodological contribution is the decomposition of WAI subscales as independent mediators in engagement-outcome analyses, an approach that required a two-stage clustering strategy to avoid circularity, and the use of early WAI measurement to establish temporal precedence for the mediation pathway. To our knowledge, this is the first study to examine early working alliance as a mechanistic pathway linking behavioral engagement to outcomes in an AI wellness tool.

2. METHODS

2.1 Study Design and Participants

Data derive from a pre-registered, parallel-arm RCT (N = 400; 200 randomized to Sonia, 200 to psychoeducation control) conducted over four weeks. Full trial design, eligibility criteria, randomization procedures, and primary outcomes are described in the primary RCT paper (Roever et al., 2026). Briefly, eligible participants were U.S. adults aged 18 or older with self-reported anxiety symptoms (baseline GAD-7 score 13–20) who owned an iPhone. All participants completed assessments at screening and then, if eligible and enrolled, baseline and at the end of weeks 1, 2, and 4.

The present secondary analysis includes the Sonia arm only (n = 200; Figure 5). Of 10,490 assessed for eligibility, 423 passed phone screening and 400 were randomized; 200 were allocated to the Sonia condition. 195 had Week 1 data, 193 had Week 2 data, and 194 completed Week 4 assessments. Thus, complete cases for mediation analyses included 192 study participants.

2.2 The Sonia Intervention

Sonia is a conversational AI emotional support application available on iOS. The platform guides users through evidence-informed techniques that they self-apply, including cognitive reframing, behavioral activation, and mindfulness-based practices. Users self-direct their use of the application; optional structured features include a morning journaling module and an evening reflection feature. No humans are involved in service delivery; the application is available continuously on demand. All users in the Sonia arm received full access to the app for the four-week trial period.

2.3 Engagement Variables

Primary clustering solution (40 variables): 28 daily behavioral indicators from Week 1 (app opens, session duration, morning journal completion, evening reflection completion, each for days 1–7) plus 9 weekly behavioral aggregates across Weeks 2–4 (mean app opens, session duration, morning journal, each for weeks 2, 3, and 4) + 3 WAI-SR subscale scores (Goal, Task, Bond) assessed at Week 1 post-randomization.

Behavioral-only clustering solution (37 variables): The same 28 daily and 9 weekly behavioral indicators as the primary solution, with WAI subscales excluded to enable circularity-free mechanistic analyses (see Section 2.6).

All variables were standardized (z-scored) prior to clustering. Missing behavioral values were imputed to zero, corresponding to non-use, consistent with the assumption that absence of engagement data reflects non-engagement (Torous et al., 2018).

Note on WAI measurement: The WAI-SR was administered at two timepoints. The Week 1 wave used a general AI referent ("the AI"); the Week 2 wave used the user's selected coach name (e.g., Sonia, Bill, etc.). This minor wording difference is unlikely to affect subscale scores substantially but is noted as a caveat, particularly for the Bond subscale.

2.4 Cluster Analysis

Hierarchical agglomerative clustering with Ward's linkage was applied to the standardized 40-variable engagement dataset. Cluster number was selected based on silhouette scores across k = 2–6 and visual inspection of the dendrogram. K-means analysis was conducted as a sensitivity check to confirm the solution was not an artifact of the hierarchical linkage method. Bootstrap stability was assessed over 1,000 iterations with Adjusted Rand Index (ARI) quantifying agreement between bootstrap solutions and the primary solution.

WAI subscales included in the primary clustering solution cannot serve as independent mediators or moderators without inducing circularity, as the mediator would partially define the exposure cluster. To address this, a second clustering was performed using behavioral variables only (37 variables), and behavioral-only cluster labels were used as the categorical exposure in all Aim 2 mechanistic analyses. The ARI between primary and behavioral-only solutions (0.392) confirms substantial concordance, and WAI scores differentiate behavioral-

only clusters despite not entering the clustering algorithm and confirms engagement predicts alliance (path a).

2.5 Outcome Measures

Primary binary outcomes: (1) PHQ-8 MCID: ≥5-point decline from baseline (Lowe et al., 2004), assessed in the elevated-baseline subsample (PHQ-8 ≥10 at baseline; n = 172). (2) GAD-7 MCID: ≥4-point decline from baseline (Toussaint et al., 2020), assessed in the full sample (n = 194 with Week 4 data).

Secondary continuous outcomes: User satisfaction with Sonia (0–100 scale); likelihood to recommend Sonia (0–100 scale); both measured at Week 1.

Mechanistic continuous outcomes (Aim 2): Baseline to Week 4 change scores for PHQ-8 and GAD-7.

2.6 Statistical Analyses

Aim 1 (Characterization)
Cluster differences in binary outcomes were tested using logistic regression with FDR correction (Benjamini-Hochberg) across 12 tests. Continuous outcomes were compared using OLS regression with the same correction. Baseline predictors of cluster membership were examined using multinomial logistic regression with the cluster named Sporadic Users as the reference category; all predictors were standardized. Minimum effective dose was examined via individual ROC curves for all 40 engagement variables per outcome, with an AUC threshold of 0.60 as a pre-specified indicator of discriminative utility.

Aim 2 (Mechanism)
The primary mediation model used Week 1 app opens (standardized) as the engagement predictor, Week 1 WAI as the mediator, and Week 4 symptom change from baseline as the outcome, providing clean temporal ordering (predictor precedes mediator precedes distal outcome), analogous to dose-response mediation designs in the psychotherapy literature. Mediation was estimated using bootstrapped causal mediation analysis with 5,000 iterations, controlling for baseline symptom severity. Moderation was tested using OLS regression with an engagement × WAI interaction term and the same covariate.

Eight mediation models (4 WAI mediators × 2 outcomes) and eight moderation models were estimated. As a sensitivity check, the primary mediation model was rerun with total four-week engagement as the predictor (mediator and outcome held fixed; Table S3). A further sensitivity check entered Week 1 WAI level and the change in WAI score between Weeks 1 and 2 simultaneously as predictors of Week 4 outcomes, to test whether stable alliance level or alliance trajectory drives outcomes (Table S6).

Post-hoc power for the primary mediation (WAI Task, the a priori most specific hypothesized

mediator) was estimated by Monte Carlo simulation (1,000 replications × 5,000-iteration bootstrap each), using fitted path coefficients and residual variance as the assumed true effect. Post-hoc power for cluster-outcome contrasts was estimated using two-proportion z-tests (binary outcomes) and two-sample t-tests (continuous outcomes).

Missing Data

Missing outcome data were characterized by chi-square test across cluster membership and by comparison of baseline severity between participants with and without follow-up data to assess missing-at-random (MAR) patterns.

3. RESULTS

3.1 Sample Characteristics

The Sonia arm included 200 participants (see Table 1 and Figure 5 for CONSORT flow). The sample was predominantly female (76%) with a mean age of 39.0 years (SD = 11.0). Baseline PHQ-8 and GAD-7 did not differ significantly across the four identified clusters (PHQ-8: $p = 0.836$; GAD-7: $p = 0.144$), indicating that symptom severity at entry did not determine engagement pattern.

3.2 Cluster Profiles

Hierarchical clustering identified four clusters as the optimal solution based on dendrogram inspection and silhouette score pattern (k=4 silhouette = 0.068, representing a local maximum above k=3 and k=5 after the k=2 solution was discarded as clinically uninformative; Figure 1). K-means analysis confirmed the solution structure. The four clusters are characterized as follows (see Figure 2 for heatmap):

Sporadic Users (n = 58, 29%): Low and inconsistent engagement across all features and time points. Rapid initial decline in use, with approximately 64% returning briefly in weeks 2–4 before final disengagement (Figure S2). Lowest WAI scores across all subscales.

Moderate Session Users (n = 80, 40%): Moderate session-level engagement with limited multi-feature use. Use concentrated in weeks 1–2. The modal engagement pattern.

Engaged Multi-Feature Users (n = 49, 24.5%): Moderate-to-high app use across all features. Distinguished by consistent use of both morning journal and evening reflection in addition to conversational features. Strong alliance scores.

Power Users (n = 13, 6.5%): Highest use across all behavioral indicators and all time windows. Sustained engagement through week 4. Highest WAI scores across all subscales.

WAI scores showed a visualized systematic gradient from Sporadic Users to Power Users even in the behavioral-only clustering solution, confirming co-variation between behavioral

persistence and affective engagement (ARI between primary and behavioral-only solutions = 0.392). Bootstrap stability analysis across 1,000 iterations yielded mean ARI = 0.438; 0% of bootstrap solutions exceeded ARI > 0.80, confirming these are engagement profiles rather than discrete latent taxonomic classes (Supplementary Figure S3).

Significant baseline predictors of cluster membership (Table 3, Figure 3): Age was positively associated with Engaged Multi-Feature (RRR = 1.82, p = 0.015) and Power User (RRR = 2.73, p = 0.012) membership relative to Sporadic Users. College-level or higher education predicted Engaged Multi-Feature membership (RRR = 2.96, p = 0.037). Neither PHQ-8 nor GAD-7 baseline symptom severity was a significant predictor of any cluster comparison.

3.3 Cluster Differences in Outcomes

After FDR correction across 12 tests, engagement clusters did not significantly predict PHQ-8 or GAD-7 MCID (Table 2). PHQ-8 MCID rates were uniform across clusters (all FDR q > 0.70). GAD-7 MCID differences were not significant after correction (all FDR q > 0.14). Post-hoc power for PHQ-8 MCID comparisons ranged from 5.4% to 10.0%, indicating the study was substantially underpowered to detect small differences in clinical outcomes across clusters; null findings here should be interpreted cautiously.

In contrast, clusters robustly predicted user experience outcomes. Engaged Multi-Feature Users reported satisfaction scores 14.77 points higher than Sporadic Users (FDR q < 0.001) and likelihood to recommend 15.79 points higher (FDR q < 0.001). Moderate Session Users reported likelihood to recommend 9.44 points higher than Sporadic Users (FDR q = 0.048). Satisfaction and recommendation were measured at Week 1 (the same window used to define clusters), so these associations are correlational rather than prospective.

Engagement depth was therefore associated with the quality of user experience but not symptom outcomes (Table 2).

3.4 Minimum Effective Dose

ROC analysis across all 40 engagement variables identified distinct predictors of GAD-7 versus PHQ-8 improvement (Figure 4). Sustained behavioral engagement (top: app opens week 4, AUC = 0.640) was the top discriminant of GAD-7 MCID; 5 of 40 variables exceeded AUC = 0.60. PHQ-8 MCID was most discriminated by therapeutic alliance variables (WAI Bond AUC = 0.619, WAI Task AUC = 0.606); 2 of 40 variables exceeded AUC = 0.60. This differential pattern, with behavioral engagement predicting anxiety symptom improvement, alliance predicting depressive symptom improvement indicates users engage in different ways and foreshadows the mediation results below.

3.5 Mediation and Moderation Analyses

Primary mediation (Table 4). Using week-1 app opens as the predictor, week-1 WAI as the

mediator, and week-4 symptom change as the outcome (N=192), all eight mediation models yielded significant bootstrapped indirect effects (all 95% CIs excluding zero). The direct effect of engagement (path c') was non-significant in all models once WAI was included in the model, consistent with full mediation. Thus, engagement improves outcomes primarily through alliance formation rather than through raw usage volume.

WAI Composite and WAI Goal were the most consistent mediators, showing the largest proportions mediated (WAI Composite: PHQ-8 51.4%, GAD-7 46.7%; WAI Goal: PHQ-8 39.7%, GAD-7 39.7%). WAI Bond was also significant for both outcomes (PHQ-8 50.5%, GAD-7 44.0%). WAI Task was significant for both outcomes (PHQ-8 38.7%, GAD-7 33.4%) and showed the most robust pattern across timepoint sensitivity analyses (see below and Table S7).

Sensitivity: broader engagement window (Table S3). Rerunning with total four-week engagement as the predictor (holding week-1 WAI mediator and week-4 outcome fixed) yielded significant indirect effects in 7 of 8 models, confirming the primary finding is not an artifact of the narrow week-1 engagement window.

Sensitivity: WAI level vs. trajectory (Table S6). When week-1 WAI level (absolute value) and WAI change between Weeks 1 and 2 were entered simultaneously as predictors of week-4 outcomes, WAI level was consistently significant across all models (p range 0.002–0.056) while WAI trajectory (change) was not significant in any model (p range 0.141–0.844). This indicates that the initial level of WAI, which remained stable, rather than any change in alliance over time, drives the effect on outcomes. This finding is consistent with evidence of the early establishment of a working alliance that remains relatively stable over time (Flückiger et al., 2018).

Sensitivity: WAI timing robustness (Table S7). A fully timing-aligned week-2 specification of engagement during Weeks 1 and 2 leading to WAI at Week 2 leading to Week 2 outcomes showed that WAI Task was the most temporally robust subscale, as the relationships were: consistent or even strengthened at Week 2. The associations with WAI Composite and Goal, in contrast, attenuated at Week 2, with WAI Bond being non-significant at either Time in this specification.

Moderation: No WAI subscale significantly moderated the engagement-outcome relationship (all engagement × WAI interaction terms non-significant, p range 0.232–0.745, Table S2). WAI subscales showed significant main effects on outcomes across models, confirming they predict symptom change independently and that the benefit of engagement does not vary as a function of alliance level. Instead, WAI operates as a pathway through which engagement produces benefit, not as an amplifier of engagement effects.

3.6 Power Analysis

Post-hoc power for the primary mediation (WAI Task, the most specific a priori hypothesis) was estimated at 62.1% for the PHQ-8 outcome and 35.1% for the GAD-7 outcome (both below the conventional 80% benchmark). These estimates indicate the study was underpowered for

confirmatory mediation inference, particularly for GAD-7. As a result, all mediation results should be interpreted as exploratory and hypothesis-generating rather than confirmatory.

3.7 Missing Data

Overall missing outcome data were 3% (6/200 participants). Missing data rates were higher in Sporadic Users (8.8%) than in other clusters (chi-square $p = 0.015$). No significant differences in baseline severity were observed between participants with and without follow-up data, supporting a missing-at-random (MAR) assumption.

4. DISCUSSION

4.1 Engagement Is Multi-Dimensional and Behaviorally-Affectively Aligned

This study characterized common engagement patterns in an AI emotional support app and examined how those patterns relate to early working alliance, four-week outcomes, and the mechanisms of symptom change. Four distinct engagement profiles emerged from clusters of behavioral engagement and WAI ratings assessed one week post-randomization. The principal finding is that early working alliance, established in the first week of app use and stable thereafter, fully mediates the pathway from behavioral engagement to symptom improvement. Said differently, behavioral engagement requires a meaningful relational element to lead to significant improvements.

A key structural finding is that WAI subscales co-vary with behavioral engagement patterns even when excluded from the clustering algorithm. The ARI of 0.392 between the primary (behavioral and affective) and behavioral-only solutions confirms that affective engagement and behavioral persistence are mutually reinforcing: users who engage more also tend to form stronger early alliance and those who do not stop engaging. This convergence validates the conceptual coherence of the cluster profiles and suggests that behavioral and affective engagement may reflect a common underlying disposition toward the intervention.

4.2 Engagement Depth Predicts Experience, Not Symptom Outcomes

PHQ-8 and GAD-7 MCID rates were uniform across clusters (56–67%), suggesting that brief contact with Sonia may be sufficient to achieve clinically meaningful symptom benefit. This pattern is consistent with a brief-contact model of action (Schleider et al., 2025) and with prior findings from mobile mental health RCTs in which even short-duration engagement was associated with benefit (Arean et al., 2016). Notably, power to detect cluster differences in clinical outcomes was very low (5–10% for PHQ-8 MCID contrasts), so these null findings should not be interpreted as evidence of true equivalence.

In contrast, engagement depth strongly predicted user experience outcomes: the approximately 14-point satisfaction and recommendation differences between higher- and lower-engagement clusters are commercially meaningful and likely to influence retention and word-of-mouth. This

suggests a decoupling of clinical and experiential pathways: optimizing sustained engagement intensity is unlikely to improve population-level clinical outcomes but could meaningfully improve retention and user satisfaction. Product development efforts aimed solely at maximizing behavioral engagement metrics may therefore not translate to better symptom outcomes.

4.3 Demographic Rather Than Symptom Severity Predictors of Engagement

Engagement profile membership was predicted by age and education rather than baseline symptom severity. Older age predicted membership in both higher-engagement clusters; college-level or higher education predicted Engaged Multi-Feature membership, in line with prior studies of digital mental health solutions (Alfaro et al., 2024; Linardon et al., 2025). Neither PHQ-8 nor GAD-7 predicted any cluster comparison, which argues against severity-based triage as a driver of engagement support decisions. Although the findings are underpowered, the findings suggest that perhaps dropout risk can be identified at intake from demographic profile alone, a finding, if confirmed, that has practical implications for proactive engagement support targeting.

4.4 Early Working Alliance as the Mechanism: Theoretical Interpretation

The primary mechanistic finding is that early working alliance, measured at one week post-randomization and stable thereafter, fully mediates the engagement-outcome relationship across all WAI subscales and both clinical outcomes. WAI Composite and Goal were the most consistent mediators; WAI Task was the most temporally robust across measurement windows; WAI Bond was significant in the primary model but did not survive a fully timing-aligned week-2 specification.

This pattern suggests that the mechanism is not uniquely attributable to one WAI component: users who form early, broad alliance, i.e., agreeing that Sonia's goals are their goals, that Sonia's methods are worth trying, and that the relationship with the AI has quality, tend to improve more. The most parsimonious interpretation is that early alliance formation reflects a general endorsement of the intervention that subsequently drives deeper and more effective engagement with its therapeutic content.

The temporal stability finding is theoretically important. When Week 1 WAI level and WAI change between Weeks 1 and 2 were tested simultaneously against Week 4 outcomes, alliance was consistently significant and change in alliance was never significant. This indicates that alliance is not a process that deepens progressively over a four-week course; rather, it is established in the first week and its quality at that point determines subsequent outcomes. This is consistent with the broader psychological literature documenting early alliance as a robust predictor precisely because it captures a relatively stable relational quality rather than a trajectory (Flückiger et al., 2018). It also clarifies the critical product design window: the first week of use, not ongoing engagement deepening, is when alliance-relevant experiences are most consequential.

WAI Task showed a distinctive pattern: it was significant in the primary model predictive of Week 4 outcomes as well as in the Week 2 timing-aligned specification, making it the most cross-timepoint robust subscale. In the timing-aligned comparison, WAI Task was the only subscale that was strengthened (for GAD-7) or was consistent (for PHQ-8) at Week 2. This suggests that task agreement, the user's endorsement of Sonia's specific methods as credible and worth doing, may have particular durability as a mechanism even as the broader alliance signal attenuates across time windows. This interpretation converges with Linardon et al.'s (2025) meta-analytic finding that task completion, rather than engagement volume, predicts outcomes in smartphone apps for symptoms of depression and anxiety.

The non-significance of WAI Bond as a mechanism in the timing-aligned specification (though significant in the primary model) is theoretically noteworthy. Emotional bond may develop as a consequence of improvement rather than a cause, or may be more proximal to user experience outcomes (satisfaction, recommendation) than to symptom change. This interpretation is consistent with the present data: engagement depth robustly predicted satisfaction but not symptom outcomes, while WAI Task showed cross-outcome robustness.

4.5 WAI as a Mediator, Not a Moderator: Implications for Personalization

The consistent absence of moderation effects indicates that the benefit of behavioral engagement does not vary as a function of alliance level. This means that pre-identifying users with high task agreement as more likely to benefit from engagement nudges is not warranted based on present data. However, the mediation findings point to a different and actionable intervention target: since early working alliance transmits the engagement-outcome relationship, experiences that build alliance formation within the first week are the highest-leverage targets for improving clinical outcomes.

This points toward specific design elements: psychoeducation about why Sonia's techniques work (transparent rationale), explicit framing of activities as evidence-based, early personalization that connects app content to the user's stated goals, and onboarding sequences designed to create perceived credibility and method endorsement during the first week of engaging with the therapeutic content itself. The implication is that the alliance formed during the first sessions may matter more for clinical outcomes than anything that happens in Weeks 2–4.

4.6 Limitations

Several limitations warrant consideration. First, this is a secondary analysis of an RCT not powered for within-cluster comparisons; post-hoc power for the primary mediation was 62.1% (PHQ-8) and 35.1% (GAD-7), below the conventional 80% benchmark. All mediation results should be interpreted as exploratory and hypothesis-generating rather than confirmatory. Second, cluster stability was modest (bootstrap mean ARI = 0.438), consistent with engagement profiles representing naturally occurring phenotypic variation rather than discrete latent classes; findings should be interpreted accordingly. Third, WAI measurement used a

slightly different referent wording at week 1 (general AI) versus week 2 (coach name), introducing a minor measurement inconsistency that is most relevant for the Bond subscale. Fourth, outcomes were assessed at four weeks, limiting conclusions about longer-term effects. Fifth, satisfaction and recommendation outcomes were measured at week 1, the same window used to define clusters, so cluster-outcome associations for these measures are correlational. Sixth, the sample had clinically elevated anxiety symptoms by design and may not generalize to wellness populations seeking support without elevated symptoms. Seventh, all mediation analyses assume no unmeasured confounding of the WAI-outcome relationship, an assumption that cannot be verified with observational secondary data.

5. CONCLUSIONS

This secondary analysis identified four engagement profiles among users of an AI emotional support app. Engagement depth predicted user experience quality but not clinical symptom outcomes, suggesting that brief contact with the app may be sufficient for clinically meaningful improvement. Engagement profiles were predicted by demographic characteristics, notably age and education, rather than symptom severity, indicating that dropout risk can be identified at intake without clinical information.

The principal mechanistic finding is that early working alliance, established within the first week of app use, fully mediates the pathway from behavioral engagement to four-week symptom improvement, with alliance level rather than alliance change driving outcomes. WAI Composite and Goal showed the largest and most consistent mediation effects; WAI Task was the most robust across measurement timepoints. No WAI subscale moderated engagement effects. These findings point toward onboarding as the critical clinical lever: personalized experiences that build early alliance such as transparent rationale, method credibility, and goal connection are more likely to improve clinical outcomes than efforts to increase sustained engagement volume. In plain terms, affective components matter more than behavioral, empowering engagement and leading to better outcomes.

Future work should examine whether real-time monitoring of early alliance can identify users at risk for non-response, test whether targeted onboarding interventions designed to strengthen early alliance improve outcomes, and evaluate these mechanisms in larger samples adequately powered for confirmatory mediation inference.

FUNDING DECLARATION
The study was funded by Sonia Health.

REFERENCES

Alfaro, A.J., Wielgosz, J., Kuhn, E., Carlson, C., Gould, C.E. (2024). Determinants and outcome correlates of engagement with a mobile mental health intervention for depression and anxiety in middle-aged and older adults. Journal of Clinical Psychology, 80(3):509-521. doi: 10.1002/jclp.23636. PMID: 38157399.

Arean, P. A., Hallgren, K. A., Jordan, J. T., Gazzaley, A., Atkins, D. C., Heagerty, P. J., & Anguera, J. A. (2016). The use and effectiveness of mobile apps for depression: Results from a fully remote clinical trial. Journal of Medical Internet Research, 18(12), e330. https://doi.org/10.2196/jmir.6482

Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the alliance-outcome correlation: Exploring the relative importance of therapist and patient variability in the alliance. Journal of Consulting and Clinical Psychology, 75(6), 842–852. https://doi.org/10.1037/0022-006X.75.6.842

Beatty C, Malik T, Meheli S, Sinha C. Evaluating the Therapeutic Alliance With a Free-Text CBT Conversational Agent (Wysa): A Mixed-Methods Study. *Frontiers in Digital Health*. 2022;4:847991. PMID: 35480752.

Brown, H. J., Andreason, H., Melling, A. K., Imel, Z. E., & Simon, G. E. (2015). Problems with using patient retention in the evaluation of mental health providers: Differences in type of dropout. Psychiatric Services, 66(8), 879–882. https://doi.org/10.1176/appi.ps.201400059

Darcy, A., Daniels, J., Salinger, D., Wicks, P., & Robinson, A. (2021). Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. *JMIR Formative Research*, *5*(5), e27868.

Darnell, D., Pullmann, M. D., Hull, T. D., Chen, S., & Areán, P. (2022). Predictors of disengagement and symptom improvement among adults with depression enrolled in Talkspace, a technology-mediated psychotherapy platform: Naturalistic observational study. JMIR Formative Research, 6(6), e36521. https://doi.org/10.2196/36521

Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. Psychotherapy, 55(4), 316–340. https://doi.org/10.1037/pst0000172

Hatcher, R. L., & Gillaspy, J. A. (2006). Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy research*, *16*(1), 12-25.

Hoffman, V., Flom, M., Mariano, T. Y., Chiauzzi, E., Williams, A., Kirvin-Quamme, A., ... & Perski, O. (2023). User engagement clusters of an 8-week digital mental health intervention guided by a relational agent (Woebot): exploratory study. *Journal of medical Internet research*,

25, e47198.

Horvath, A. O., & Bedi, R. P. (2002). The alliance. In J. C. Norcross (Ed.), Psychotherapy relationships that work: Therapist contributions and responsiveness to patients (pp. 37–69). Oxford University Press.

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

Lee, B. S., Sen, P. K., Park, N. S., Boothroyd, R. A., Peters, R. H., & Chiriboga, D. A. (2014). A clustering method to identify who benefits most from the treatment group in clinical trials. *Health Psychology and Behavioral Medicine*, 2(1), 1030–1040.

Linardon, J., Torous, J., Messer, M., Liu, C., Bell, I. H., Nicholas, J., & Goldberg, S. B. (2025). Association between user engagement and clinical outcomes in smartphone apps for depression and anxiety: A systematic review and meta-analysis. Psychiatry Research, 116864. https://doi.org/10.1016/j.psychres.2025.116864.

Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the Patient Health Questionnaire-9. *Medical Care*. 2004;42(12):1194-1201. PMID: 15550799.

Perski, O., Blandford, A., West, R., & Michie, S. (2017). Conceptualising engagement with digital behaviour change interventions: A systematic review using principles from critical interpretive synthesis. Translational Behavioral Medicine, 7(2), 254–267. https://doi.org/10.1007/s13142-016-0453-1.

Robinson, L., Delgadillo, J., & Kellett, S. (2020). The dose-response effect in routinely delivered psychological therapies: A systematic review. Psychotherapy Research, 30(1), 79–96. https://doi.org/10.1080/10503307.2019.1566676

Roever, N., Feld, B., Forman-Hoffman, V., Arean, P., Wolfe, L. (2026). Short-term Evaluation of a Voice-Based Generative AI Emotional Support Companion: A Randomized Controlled Trial. https://osf.io/preprints/psyarxiv/9m8aw_v1.

Schleider JL, Zapata JP, Rapoport A, Wescott A, Ghosh A, Kaveladze B, Szkody E, Ahuvia IL. Single-Session Interventions for Mental Health Problems and Service Engagement: Umbrella Review of Systematic Reviews and Meta-Analyses. *Annual Review of Clinical Psychology*. 2025;21:279-303. https://doi.org/10.1146/annurev-clinpsy-081423-025033

Torous, J., Lipschitz, J., Ng, M., & Firth, J. (2018). Dropout rates in clinical trials of smartphone apps for depressive symptoms: A systematic review and meta-analysis. Journal of Affective Disorders, 263, 413–419.

Toussaint A, Hüsing P, Gumz A, Wingenfeld K, Härter M, Schramm E, Löwe B. (2020). Sensitivity to change and minimal clinically important difference of the 7-item Generalized Anxiety Disorder Questionnaire (GAD-7). Journal of Affective Disorders, 265, 395-401. PMID: 32090765.

Zilcha-Mano, S. (2017). Is the alliance really therapeutic? Revisiting this question in light of recent methodological advances. American Psychologist, 72(4), 311–325. https://doi.org/10.1037/a0040435

Hoffman et al. | Engagement Clusters Study

Tables and Figures

Table 1. Baseline Characteristics by Engagement Cluster (N = 200)

Variable	Sporadic Users (n=58, 29%)	Moderate Users (n=80, 40%)	Engaged Multi-Feature (n=49, 24.5%)	Power Users (n=13, 6.5%)	p
Demographics					
Age, M (SD)	36.8 (10.3)	36.7 (9.6)	43.6 (11.1)	45.7 (14.7)	<0.001 *
Annual income, M (SD)	$35,905 (29,086)	$39,494 (31,786)	$49,740 (39,113)	$31,923 (20,004)	0.113
Gender, n (%)					0.599
Female	40 (69%)	63 (79%)	38 (78%)	12 (92%)	
Male	17 (29%)	16 (20%)	11 (22%)	1 (8%)	
Non-binary	1 (2%)	1 (1%)	0 (0%)	0 (0%)	
Race/Ethnicity, n (%)					0.870
White or Caucasian	40 (69%)	48 (60%)	34 (71%)	9 (69%)	
Black or African American	14 (24%)	24 (30%)	11 (23%)	3 (23%)	
Asian	3 (5%)	2 (2%)	2 (4%)	0 (0%)	
Other	1 (2%)	5 (6%)	1 (2%)	1 (8%)	
Education, n (%)					0.208
HS diploma or GED	21 (38%)	23 (29%)	6 (12%)	6 (50%)	
Some college, no degree	16 (29%)	26 (33%)	24 (50%)	4 (33%)	
Associates degree	7 (12%)	6 (8%)	6 (12%)	1 (8%)	
Bachelor's degree	4 (7%)	14 (18%)	6 (12%)	0 (0%)	
Graduate/professional degree	3 (5%)	3 (4%)	3 (6%)	1 (8%)	
Some HS or less	5 (9%)	7 (9%)	3 (6%)	0 (0%)	
Employment, n (%)					0.446
Working full-time	17 (29%)	27 (34%)	8 (16%)	4 (31%)	
Working part-time	8 (14%)	10 (12%)	11 (22%)	1 (8%)	
Homemaker/stay-at-home parent	4 (7%)	16 (20%)	9 (18%)	2 (15%)	
Unemployed	15 (26%)	12 (15%)	8 (16%)	1 (8%)	
Student	6 (10%)	7 (9%)	5 (10%)	1 (8%)	
Retired	1 (2%)	1 (1%)	2 (4%)	1 (8%)	
Other	7 (12%)	7 (9%)	6 (12%)	3 (23%)	

Variable	Sporadic Users (n=58, 29%)	Moderate Users (n=80, 40%)	Engaged Multi-Feature (n=49, 24.5%)	Power Users (n=13, 6.5%)	p
Baseline Clinical and Psychological					
PHQ-8, M (SD)	14.4 (4.6)	14.2 (4.2)	13.7 (4.3)	14.2 (2.6)	0.836
GAD-7, M (SD)	16.1 (2.2)	16.1 (2.3)	15.2 (2.1)	15.8 (2.2)	0.144
AI Efficacy Estimate, M (SD)	75.2 (18.9)	76.7 (19.5)	77.7 (18.9)	70.9 (21.7)	0.694
AI Emotional Comfort, M (SD)	81.9 (19.2)	80.5 (18.9)	85.5 (17.9)	81.8 (16.1)	0.529
AI Privacy Concern, M (SD)	2.1 (0.8)	2.1 (0.8)	2.1 (0.8)	1.9 (0.8)	0.856
UCLA Loneliness, M (SD)	7.8 (1.6)	7.4 (1.5)	7.7 (1.6)	7.8 (1.5)	0.488
ATQ (Automatic Thoughts), M (SD)	25.3 (7.2)	25.5 (7.3)	23.9 (6.7)	25.0 (7.1)	0.618
Prior therapy (before this year), n (%)	44 (76%)	62 (78%)	36 (73%)	11 (85%)	0.853
Current medication, n (%)	13 (23%)	22 (28%)	12 (24%)	3 (23%)	0.740
Working Alliance — Week 1 Post-Randomization (WAI-SR, w3 wave)					
WAI Goal, M (SD)	3.3 (1.1)	3.6 (1.0)	4.1 (0.8)	4.0(1.0)	<0.001
WAI Task, M (SD)	3.1 (1.0)	3.3 (0.9)	3.9 (0.8)	3.8(1.0)	<0.001
WAI Bond, M (SD)	3.3 (1.1)	3.5 (1.1)	4.1 (0.9)	3.9(0.7)	<0.001

Note. M = mean; SD = standard deviation. Continuous variables compared by one-way ANOVA; categorical variables by chi-square. Reference = Sporadic Users. WAI subscales measured at week 1 post-randomization (w3 wave).. *p < 0.05. ATQ = Automatic Thoughts Questionnaire. Cluster sizes reflect the primary 40-variable (week-1 WAI) clustering solution.

Table 2. FDR-Corrected Outcome Differences by Engagement Cluster

Outcome	Compar-ison	N	Estimate (95% CI)	p (raw)	q (FDR)	FDR sig.
Clinical Outcomes (Week 4)						
PHQ-8 MCID	Engaged Multi-feature Users vs Sporadic Users	167	OR = 0.99 [0.41, 2.42]	0.985	0.985	No
PHQ-8 MCID	Moderate Users vs Sporadic Users	167	OR = 0.81 [0.37, 1.79]	0.608	0.768	No
PHQ-8 MCID	Power Users vs Sporadic Users	167	OR = 0.72 [0.19, 2.79]	0.640	0.768	No
GAD-7 MCID	Engaged Multi-feature Users vs Sporadic Users	194	OR = 2.41 [0.76, 7.66]	0.136	0.204	No
GAD-7 MCID	Moderate Users vs Sporadic Users	194	OR = 2.38 [0.89, 6.39]	0.086	0.147	No
GAD-7 MCID	Power Users vs Sporadic Uswers	194	OR = 0.81 [0.19, 3.53]	0.780	0.851	No
User Experience Outcomes (Week 1 — correlational)						
Satisfaction	Engaged Multi-feature Users vs Sporadic Users	184	+14.77 [7.17, 22.37]	<0.001	<0.001	**YES***
Satisfaction	Moderate Users vs Sporadic Users	184	+7.07 [0.17, 13.97]	0.045	0.090	No
Satisfaction	Power Users vs Sporadic Uswers	184	+12.87 [1.17, 24.57]	0.031	0.074	No

Outcome	Compar-ison	N	Estimate (95% CI)	p (raw)	q (FDR)	FDR sig.
Recommend	Engaged Multi-feature Users vs Sporadic Users	186	+15.79 [7.65, 23.92]	<0.001	<0.001	**YES***
Recommend	Moderate Users vs Sporadic Users	186	+9.44 [2.11, 16.76]	0.012	0.048	**YES***
Recommend	Power Users vs Sporadic Uswers	186	+14.97 [2.50, 27.45]	0.019	0.057	No

Note. PHQ−8 MCID and GAD−7 MCID = logistic regressions (odds ratios); Satisfaction and Recommend = OLS regressions (mean differences, 0–100 scale). p−values 3 dp; *p < .05 raw. q (FDR) = Benjamini−Hochberg correction across 12 tests. Shaded rows = FDR q < 0.05. PHQ−8/GAD−7 MCID assessed at week 4 (w5) vs. baseline; elevated−baseline subsample (PHQ−8 ≥ 10, n=172) for PHQ−8 MCID. Satisfaction and Recommend measured at week 1 (w3) — correlational associations only. Post−hoc power for PHQ−8 MCID contrasts: 5.4%–10.0%; null results should be interpreted cautiously.

Table 3. Multinomial Logistic Regression: Baseline Predictors of Cluster Membership (n = 190, Reference = Sporadic Users)

Predictor (standardized)	Engaged Multi-Feature vs. Sporadic Users		Moderate Users vs. Sporadic Users		Power Users vs. Sporadic Users	
	RRR (95% CI)	p	RRR (95% CI)	p	RRR (95% CI)	p
Age (+1 SD)	**1.82 [1.12, 2.94]**	**0.015 ***	1.01 [0.66, 1.54]	0.968	**2.73 [1.25, 5.97]**	**0.012 ***
Income	1.19 [0.72, 1.98]	0.497	1.05 [0.66, 1.67]	0.834	0.78 [0.30, 2.04]	0.616
GAD-7 (baseline)	0.73 [0.45, 1.19]	0.209	0.93 [0.62, 1.39]	0.724	0.88 [0.39, 2.00]	0.764
PHQ-8 (baseline)	1.00 [0.55, 1.80]	0.991	0.98 [0.59, 1.62]	0.940	0.78 [0.28, 2.17]	0.636
Automatic Thoughts	1.26 [0.65, 2.43]	0.497	1.50 [0.85, 2.65]	0.163	1.16 [0.40, 3.39]	0.785
Behavioral Activation	1.38 [0.80, 2.37]	0.252	1.31 [0.82, 2.08]	0.254	1.01 [0.39, 2.63]	0.984
Cognitive Distortions	0.85 [0.50, 1.45]	0.560	1.21 [0.77, 1.89]	0.411	1.85 [0.78, 4.35]	0.160
Loneliness	1.11 [0.65, 1.89]	0.703	0.69 [0.44, 1.09]	0.110	1.15 [0.46, 2.85]	0.760
AI Privacy Concern	1.09 [0.67, 1.76]	0.736	1.00 [0.66, 1.51]	0.998	0.70 [0.32, 1.54]	0.370
AI Efficacy Estimate	1.13 [0.70, 1.84]	0.612	1.12 [0.75, 1.67]	0.576	0.75 [0.38, 1.49]	0.412
AI Emotional Comfort	1.12 [0.68, 1.84]	0.661	0.86 [0.58, 1.29]	0.470	1.02 [0.46, 2.30]	0.956
Male (vs. Female)	0.57 [0.18, 1.79]	0.340	0.54 [0.20, 1.42]	0.211 †	0.19 [0.02, 2.09]	0.175
Non-White (vs. White)	1.58 [0.52, 4.79]	0.420	2.03 [0.82, 5.03]	0.126	2.04 [0.34, 12.11]	0.431

Predictor (standardized)	Engaged Multi-Feature vs. Sporadic Users		Moderate Users vs. Sporadic Users		Power Users vs. Sporadic Users	
	RRR (95% CI)	**p**	**RRR (95% CI)**	**p**	**RRR (95% CI)**	**p**
College+ Education	**2.96 [1.07, 8.22]**	**0.037 ***	1.15 [0.53, 2.52]	0.720	0.79 [0.18, 3.50]	0.760 ‡
Employed	0.65 [0.25, 1.66]	0.364	1.12 [0.51, 2.48]	0.779	1.21 [0.26, 5.56]	0.808
Prior Therapy	0.70 [0.24, 2.07]	0.524	0.79 [0.31, 2.04]	0.631	0.64 [0.10, 4.10]	0.634
Current Medication	1.12 [0.39, 3.21]	0.829	1.81 [0.74, 4.44]	0.193	1.25 [0.24, 6.57]	0.790

Note. RRR = relative risk ratio vs. Sporadic Users; 95% CI; *p < .05. All predictors standardized. Significant findings: Age predicts Engaged Multi-Feature (RRR=1.82, p=0.015) and Power Users (RRR=2.73, p=0.012); College+ education predicts Engaged Multi-Feature (RRR=2.96, p=0.037). PHQ-8 and GAD-7 non-significant in all comparisons. †Male gender for Moderate Users was p=0.007 under the prior week-2 WAI clustering solution; no longer significant under week-1 WAI solution. ‡College+ education for Power Users was p=0.016 under the prior solution; no longer significant.

Table 4. Mediation Results: Week-1 WAI as Mediator of Engagement-Outcome Relationships

Week-1 engagement → Week-1 WAI → Week-4 outcomes. N = 192. All 8 models significant.

Mediator	Outcome	Path a (Eng→ WAI)	Path b (WAI→ Outcome)	Path c (Total effect)	Indirect effect [95% CI]	% Med.
WAI Composite	PHQ-8	0.248 (<0.001)*	−0.249 (<0.001)*	−0.120 (0.080)	-0.062[-0.122, -0.016]	51.4
WAI Composite	GAD-7	0.248 (<0.001)*	−0.206 (0.005)*	−0.109 (0.133)	-0.051[-0.114, -0.008]	46.7
WAI Goal	PHQ-8	0.239 (0.001)*	−0.199 (0.003)*	−0.120 (0.080)	-0.048[-0.103, -0.008]	39.7
WAI Goal	GAD-7	0.239 (0.001)*	−0.182 (0.012)*	−0.109 (0.133)	-0.043[-0.098, -0.005]	39.7
WAI Task	PHQ-8	0.197 (0.007)*	−0.235 (<0.001)*	−0.120 (0.080)	-0.046[-0.103, -0.008]	38.7
WAI Task	GAD-7	0.197 (0.007)*	−0.185 (0.010)*	−0.109 (0.133)	-0.036[-0.090, -0.003]	33.4
WAI Bond	PHQ-8	0.246 (<0.001)*	−0.246 (<0.001)*	−0.120 (0.080)	-0.060[-0.118, -0.018]	50.5
WAI Bond	GAD-7	0.246 (<0.001)*	−0.195 (0.007)*	−0.109 (0.133)	-0.048[-0.105, -0.007]	44.0

Note. Outcome = week-4 (w5) symptom change relative to baseline. Path a = week-1 app opens → week-1 WAI; path b = week-1 WAI → week-4 outcome; path c = total effect; path c' (not shown) non-significant in all models (range p=0.275–0.412), consistent with full mediation. Standardized OLS betas; *p < .05. Indirect effects from 5,000-iteration percentile bootstrap (*bold = 95% CI excludes 0). Engagement = week-1 app opens only. See Table S3 for total 4-week engagement sensitivity (7/8 significant); Table S7 for fully timing-aligned week-2 comparison.

Figure 1. Engagement Trajectory by Cluster (Four Weeks, Three Panels)

Outcome (Mediator)	Path a	Path b	Path c (Total)	Path c' (Direct)	Indirect (95% CI)	% Med.
PHQ-8 (WAI Comp.)	**0.248 (p=<0.001)**	**-0.249 (p=<0.001)**	-0.120 (p=0.080)	-0.058 (p=0.391)	**-0.062 [-0.122, -0.016]**	51.4
GAD-7 (WAI Comp.)	**0.248 (p=<0.001)**	**-0.206 (p=0.005)**	-0.109 (p=0.133)	-0.060 (p=0.412)	**-0.051 [-0.114, -0.008]**	46.7
PHQ-8 (WAI Goal)	**0.239 (p=0.001)**	**-0.199 (p=0.003)**	-0.120 (p=0.080)	-0.072 (p=0.293)	**-0.048 [-0.103, -0.008]**	39.7
GAD-7 (WAI Goal)	**0.239 (p=0.001)**	**-0.182 (p=0.012)**	-0.109 (p=0.133)	-0.067 (p=0.360)	**-0.043 [-0.098, -0.005]**	39.7
PHQ-8 (WAI Task)	**0.197 (p=0.007)**	**-0.235 (p=<0.001)**	-0.120 (p=0.080)	-0.074 (p=0.275)	**-0.046 [-0.103, -0.008]**	38.7
GAD-7 (WAI Task)	**0.197 (p=0.007)**	**-0.185 (p=0.010)**	-0.109 (p=0.133)	-0.074 (p=0.306)	**-0.036 [-0.090, -0.003]**	33.4
PHQ-8 (WAI Bond)	**0.246 (p=<0.001)**	**-0.246 (p=<0.001)**	-0.120 (p=0.080)	-0.059 (p=0.382)	**-0.060 [-0.118, -0.018]**	50.5
GAD-7 (WAI Bond)	**0.246 (p=<0.001)**	**-0.195 (p=0.007)**	-0.109 (p=0.133)	-0.063 (p=0.392)	**-0.048 [-0.105, -0.007]**	44

Note: N=192 for all models. Outcome = week-4 (w5) symptom change relative to baseline, 4 weeks post-randomization. Path a = engagement→WAI, path b = WAI→outcome, path c = total effect, path c' = direct effect. Standardized OLS betas; p-values 3 dp, bold = p<.05. Indirect effect significance from 5,000-iteration percentile bootstrap (bold = 95% CI excludes 0), not a p-value. Engagement = week-1 app opens only (predates WAI measurement): week-1 engagement → week-1 WAI → week-4 outcome, a clean temporal ordering (see Table S3 for a total-4-week-engagement sensitivity and Table S7 for a fully timing-aligned week-2 comparison).

Main Figures

Figure 1. Engagement trajectory (4 weeks, 3 panels)

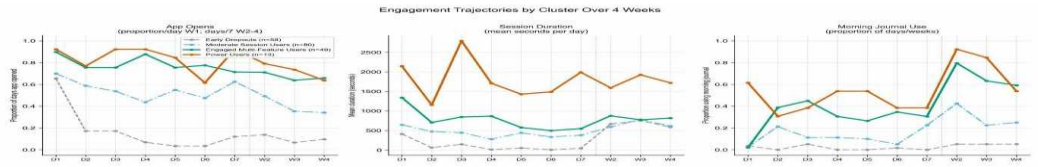


Figure 1: Figure 1. Engagement trajectory

Mean engagement trajectories by cluster. Panel A: App opens (proportion of days active, W1 days 1–7; days per week, W2–4). Panel B: Mean session duration (seconds/day). Panel C: Morning journal use (proportion of days/weeks). Cluster sizes: Sporadic Users n=58; Moderate Users n=80; Engaged Multi-Feature n=49; Power Users n=13.

Figure 2. Cluster Heatmap — Primary 40-Variable Solution (Week-1 WAI)

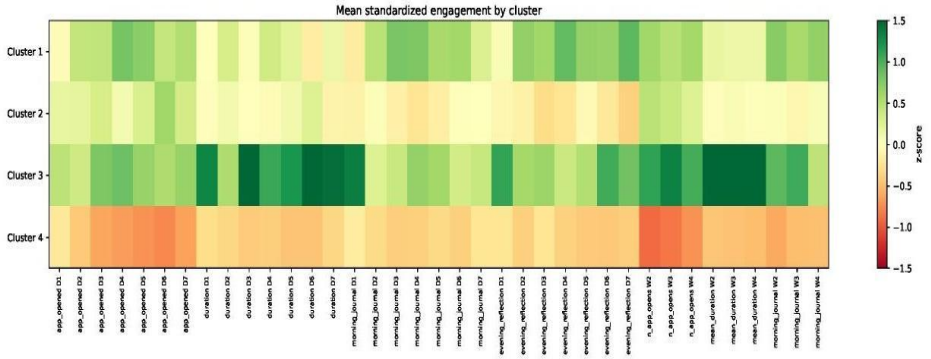


Figure 2: Figure 2. Cluster heatmap (40-variable solution)

Mean standardized engagement by cluster (N=200). Color scale: green = above sample mean; red = below. Rows: daily behavioral variables (W1), weekly aggregates (W2–4), WAI-SR subscales (Goal, Task, Bond; week-1 wave). ARI between primary (40-variable) and behavioral-only (37-variable) solutions = 0.392.

Figure 3. Forest Plot: Baseline Predictors of Cluster Membership

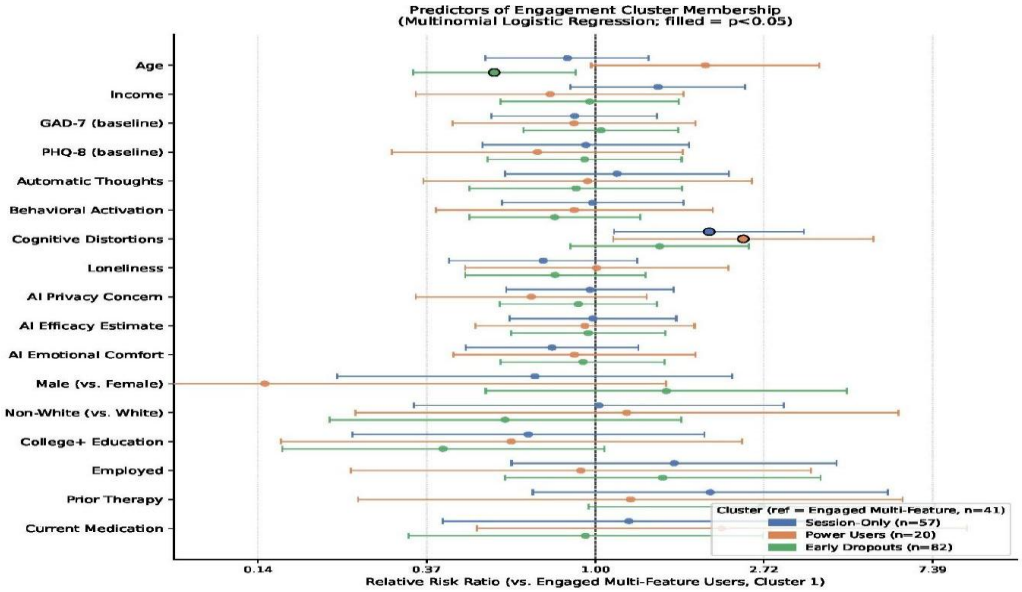
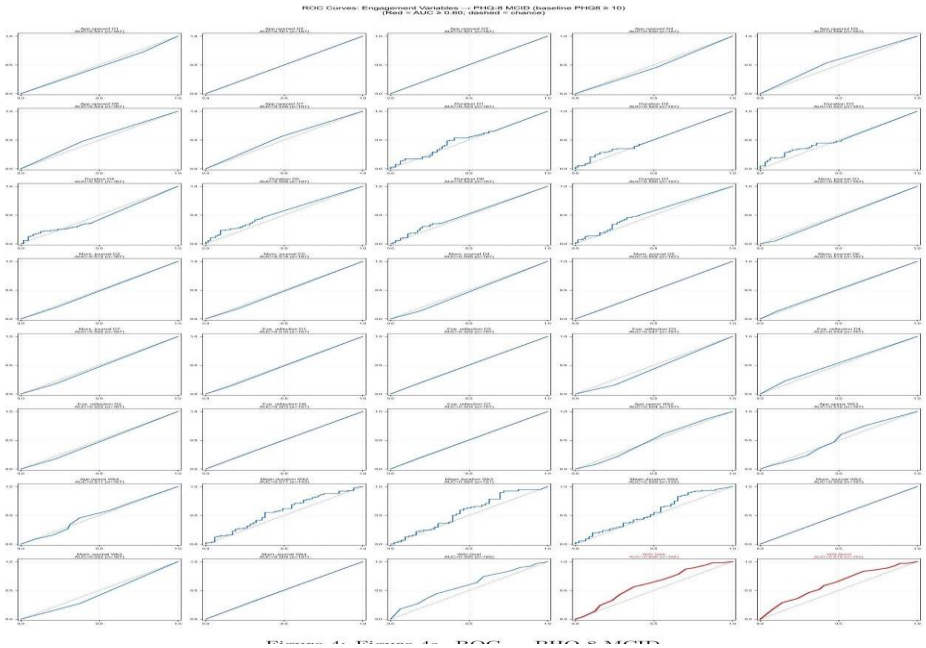


Figure 3: Figure 3. Predictors of cluster membership

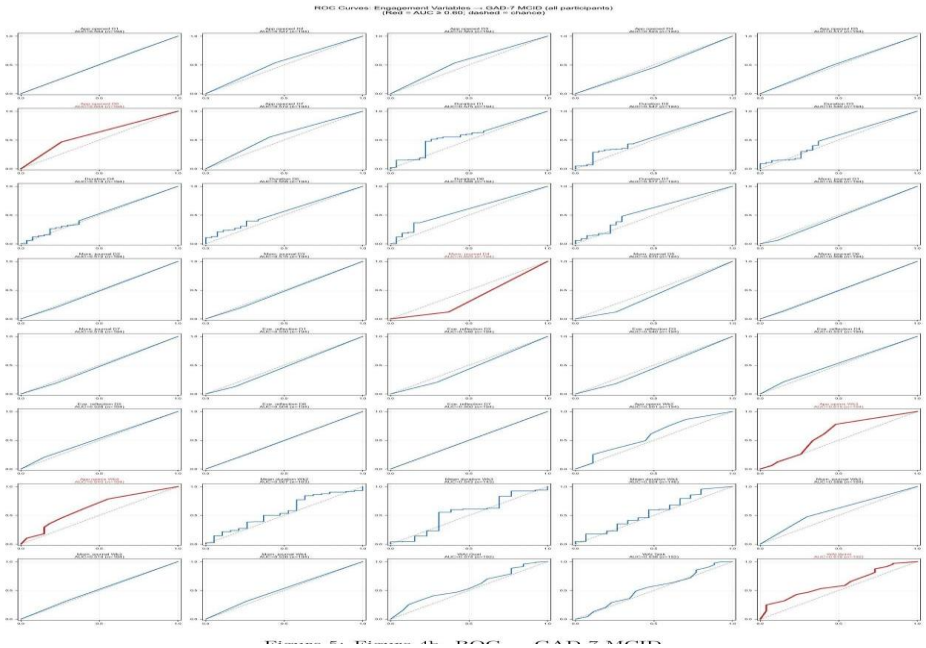
RRRs and 95% CI from multinomial logistic regression (n=190). Reference = Sporadic Users. Filled points = p < 0.05. Log scale. Significant: Age → Engaged Multi-Feature (RRR=1.82, p=0.015) and Power Users (RRR=2.73, p=0.012); College+ → Engaged Multi-Feature (RRR=2.96, p=0.037). PHQ-8 and GAD-7 non-significant throughout.

Figure 4a. ROC Grid: PHQ-8 MCID



AUC for all 40 engagement/alliance variables (elevated-baseline subsample, n=172). Red = AUC ≥ 0.60. Top: WAI Bond (AUC=0.619), WAI Task (AUC=0.606). 2 of 40 variables exceed threshold.

Figure 4b. ROC Grid: GAD-7 MCID



All participants, n=194. Top: App opens Week 4 (AUC=0.640). 5 of 40 variables exceed AUC=0.60.

Figure 5. CONSORT Flow Diagram — Treatment (AI App) Arm

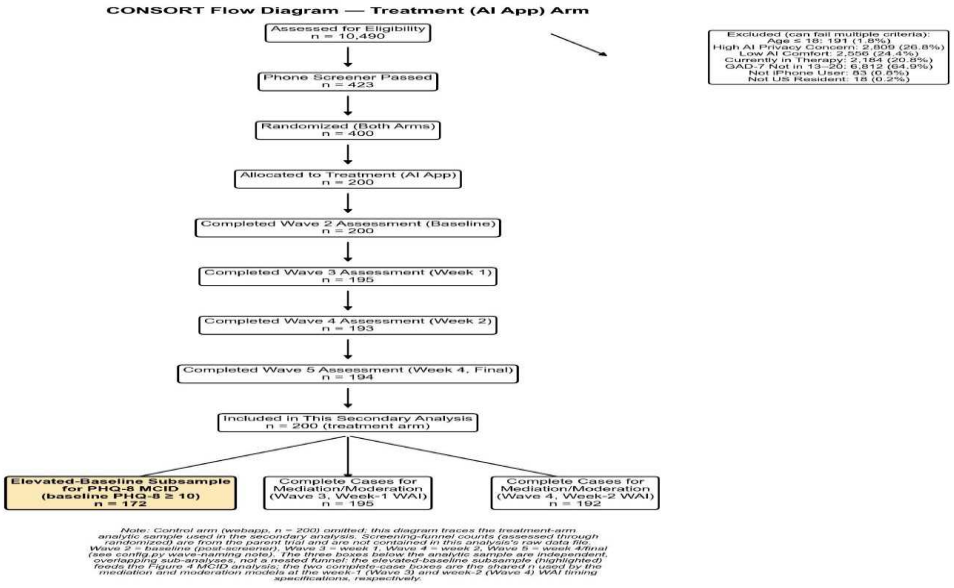


Figure 6: Figure 5. CONSORT flow diagram — treatment (AI App) arm

Participant flow through the RCT treatment arm. Assessed for eligibility n=10,490; phone screener passed n=423; randomized n=400; allocated to treatment n=200. Wave completers: w3 (week 1) n=195; w4 (week 2) n=193; w5 (week 4) n=194. Secondary analysis N=200. Sub-samples: PHQ-8 MCID elevated-baseline n=172; complete cases mediation (week-1 WAI) n=192.

Supplementary Tables

Supplementary Table S1. ARI Comparison: Primary vs. Behavioral-Only Clustering Solutions

Comparison	ARI	Interpretation
Primary (40-var, week-1 WAI) vs. Behavioral-only (37-var)	0.392	Substantial concordance; confirms WAI co-varies with behavioral clusters even when excluded
Bootstrap stability — primary solution (1,000 iterations)	Mean = 0.438	Profiles with fuzzy boundaries; not discrete latent classes
Bootstrap solutions exceeding ARI > 0.80	0%	Consistent with engagement profile rather than latent taxon interpretation
Primary (week-1 WAI) vs. Primary (week-2 WAI)	0.480	Moderate-to-high cluster stability across WAI measurement timepoints

Note. ARI = Adjusted Rand Index. Bootstrap stability over 1,000 resampling iterations.

Supplementary Table S2. Moderation Results: Week-1 WAI as Moderator (N=195)

Week-1 engagement × Week-1 WAI; Week-1 symptom-change outcome. All 8 interaction terms non-significant.

Outcome (Moderator)	N	Engage-ment b (p)	WAI b (p)	Interact-ion b (p)	p (int.)	R^2
PHQ-8 (WAI Composite)	195	−0.085 (0.226)	−0.185 (0.008)*	0.053	0.423	0.157
GAD-7 (WAI Composite)	195	−0.067 (0.372)	−0.158 (0.033)*	0.032	0.657	0.044
PHQ-8 (WAI Goal)	195	−0.085 (0.226)	−0.207 (0.003)*	0.030	0.654	0.163
GAD-7 (WAI Goal)	195	−0.067 (0.370)	−0.169 (0.022)*	0.023	0.745	0.047
PHQ-8 (WAI Task)	195	−0.093 (0.178)	−0.192 (0.005)*	0.057	0.396	0.160
GAD-7 (WAI Task)	195	−0.077 (0.300)	−0.136 (0.063)	0.057	0.430	0.040
PHQ-8 (WAI Bond)	195	−0.100 (0.157)	−0.114 (0.103)	0.082	0.232	0.140
GAD-7 (WAI Bond)	195	−0.075 (0.316)	−0.130 (0.080)	0.025	0.733	0.035

Note. Standardized OLS betas; *p < .05. All 8 interaction terms non-significant (p range 0.232–0.745). WAI subscales show significant main effects in several models (path b) but do not modify the strength of engagement's effect. See Table S5 for fully timing-aligned week-2 robustness (0/8 significant, N=192).

Supplementary Table S3. Sensitivity: Total 4-Week Engagement → Week-1 WAI → Week-4 Outcomes (N=192)

Mediator (week-1 WAI) and outcome (week-4) held fixed; engagement broadened to total 4-week composite. 7/8 models significant.

Outcome (Mediator)	N	Path a (p)	Path b (p)	Path c (p)	Indirect [95% CI]	% Med.
PHQ-8 (WAI Composite)	192	0.200 (0.008)*	−0.235 (<0.001)*	−0.195 (0.005)*	−0.047 [−0.101, −0.008]*	24.1%
GAD-7 (WAI Composite)	192	0.200 (0.008)*	−0.201 (0.005)*	−0.146 (0.050)*	−0.040 [−0.095, −0.004]*	27.6%
PHQ-8 (WAI Goal)	192	0.197 (0.009)*	−0.187 (0.005)*	−0.195 (0.005)*	−0.037 [−0.085, −0.004]*	18.9%
GAD-7 (WAI Goal)	192	0.197 (0.009)*	−0.177 (0.013)*	−0.146 (0.050)*	−0.035 [−0.084, −0.002]*	24.0%
PHQ-8 (WAI Task)	192	0.177 (0.018)*	−0.224 (<0.001)*	−0.195 (0.005)*	−0.040 [−0.093, −0.002]*	20.3%
GAD-7 (WAI Task)	192	0.177 (0.018)*	−0.181 (0.011)*	−0.146 (0.050)*	−0.032 [−0.084, 0.000]	22.0%
PHQ-8 (WAI Bond)	192	0.178 (0.018)*	−0.235 (<0.001)*	−0.195 (0.005)*	−0.042 [−0.089, −0.007]*	21.4%
GAD-7 (WAI Bond)	192	0.178 (0.018)*	−0.192 (0.007)*	−0.146 (0.050)*	−0.034 [−0.084, −0.002]*	23.5%

Note. Standardized OLS betas; *p < .05. Indirect effects from 5,000-iteration percentile bootstrap (*bold = CI excludes 0). WAI Task × GAD-7 borderline (upper CI = 0.000). Confirms primary finding is not an artifact of the narrow week-1 engagement window.

Supplementary Table S4. Cluster Stability: Week-1 vs. Week-2 WAI in Primary Clustering (N=200)

	Week 2 Engaged Multi-feature Users	Week 2 Moderate Users	Week 2 Sporadic Users	Week 2 Power Users
Week 1	9	15	0	25

	Week 2 Engaged Multi-feature Users	Week 2 Moderate Users	Week 2 Sporadic Users	Week 2 Power Users
Engaged Multi-feature Users				
Week 1 Moderate Users	0	66	7	7
Week 1 Sporadic Users	0	8	50	0
Week 1 Power Users	12	0	0	1

Note. ARI between Week 1 WAI and Week 2 WAI solutions = 0.480. Moderate-to-high stability consistent with WAI stability over time.

Supplementary Table S5. Moderation Robustness: Fully Timing-Aligned Week-2 Specification

Week-2 WAI moderator; week-1+2 engagement; week-2 symptom-change outcome. N=192. 0/8 interaction terms significant.

Outcome (Moderator)	N	Engagement b (p)	WAI b (p)	Interaction b (p)	p (int.)	R^2
PHQ-8 (WAI Composite)	192	−0.249 (<0.001)*	−0.101 (0.148)	−0.042	0.510	0.211
GAD-7 (WAI Composite)	192	−0.168 (0.027)*	−0.119 (0.108)	−0.070	0.314	0.089
PHQ-8 (WAI Goal)	192	−0.251 (<0.001)*	−0.097 (0.164)	−0.036	0.567	0.209
GAD-7 (WAI Goal)	192	−0.160 (0.033)*	−0.148 (0.047)*	−0.072	0.294	0.096
PHQ-8 (WAI Task)	192	−0.235 (<0.001)*	−0.165 (0.015)*	−0.026	0.679	0.226
GAD-7 (WAI Task)	192	−0.155 (0.036)*	−0.176 (0.016)*	−0.041	0.557	0.103
PHQ-8 (WAI Bond)	192	−0.275 (<0.001)*	−0.017 (0.803)	−0.057	0.410	0.203
GAD-7 (WAI Bond)	192	−0.202 (0.008)*	−0.009 (0.904)	−0.089	0.234	0.080

Note. 0 of 8 interaction terms significant (p range 0.234–0.679). Confirms moderation null result is robust across WAI timing specifications. *p < .05.

Supplementary Table S6. WAI Level vs. Trajectory as Simultaneous Predictors of Week-4 Outcomes (N=190)

Week-1 WAI level and week2-week1 WAI change entered simultaneously → week-4 outcomes. WAI level consistently significant; WAI trajectory never significant.

Outcome (WAI)	N	Engagement b (p)	WAI Level—Wk1 b (p)	WAI Trajectory—Wk2−Wk1 b (p)	R^2
PHQ-8 (WAI Composite)	190	−0.075 (0.285)	−0.238 (0.002)*	0.043 (0.552)	0.214
GAD-7 (WAI Composite)	190	−0.069 (0.362)	−0.194 (0.016)*	0.031 (0.695)	0.077
PHQ-8 (WAI Goal)	190	−0.082 (0.245)	−0.199 (0.012)*	0.016 (0.839)	0.192
GAD-7 (WAI Goal)	190	−0.066 (0.383)	−0.190 (0.023)*	−0.016 (0.844)	0.068
PHQ-8 (WAI Task)	190	−0.069 (0.317)	−0.264 (<0.001)*	−0.056 (0.446)	0.209
GAD-7 (WAI Task)	190	−0.064 (0.393)	−0.207 (0.010)*	−0.055 (0.484)	0.072
PHQ-8 (WAI Bond)	190	−0.085 (0.219)	−0.207 (0.005)*	0.106 (0.141)	0.220
GAD-7 (WAI Bond)	190	−0.083 (0.266)	−0.153 (0.056)	0.104 (0.178)	0.082

Note. Level = week-1 WAI absolute value (standardized); Trajectory = week2-week1 WAI change (standardized). Both entered simultaneously with week-1 engagement and baseline severity as controls. Outcome = week-4 (w5) symptom change. *p < .05.

Supplementary Table S7. Mediation Robustness: Fully Timing-Aligned Week-1 vs. Week-2 Specifications

Wk1 = week-1-only engagement + week-1 WAI + week-1 outcome (w3). Wk2 = week-1+2 engagement + week-2 WAI + week-2 outcome (w4). Note: Wk1 column here uses week-1 outcome and differs from Table 4 (week-4 outcome).

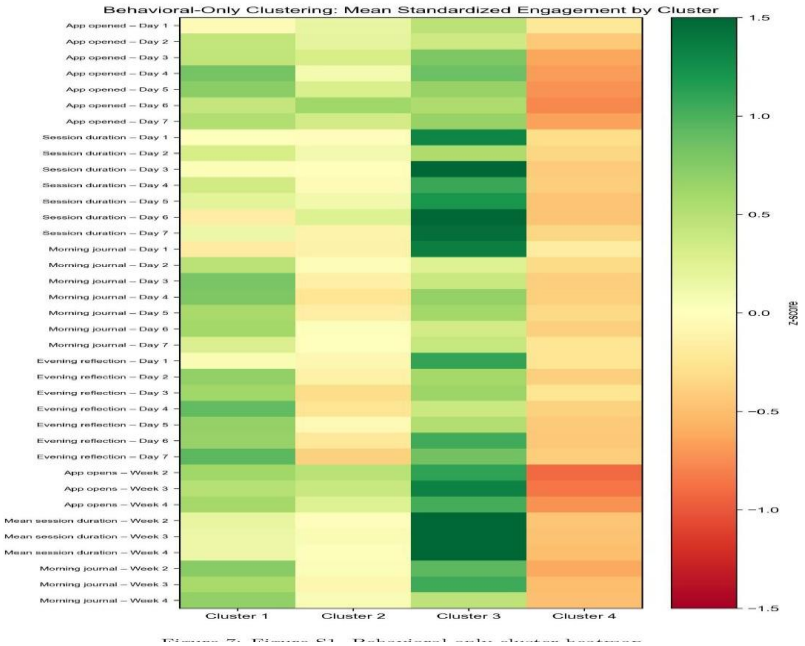
Outcome (Mediator)	Indirect Wk1 [95% CI]	% Med Wk 1	Indirect Wk2 [95% CI]	% Med Wk 2	Cross-timing robustness
PHQ-8 (WAI Composite)	−0.047 [−0.106, −0.008]*	34.2 %	−0.032 [−0.090, 0.013]	11.4 %	Attenuated at Wk2

Outcome (Mediator)	Indirect Wk1 [95% CI]	% Med Wk 1	Indirect Wk2 [95% CI]	% Med Wk 2	Cross-timing robustness
GAD-7 (WAI Composite)	**−0.040 [−0.097, −0.002]***	36.6 %	−0.037 [−0.100, 0.014]	18.5 %	Attenuated at Wk2
PHQ-8 (WAI Goal)	**−0.050 [−0.108, −0.011]***	36.2 %	−0.029 [−0.084, 0.014]	10.5 %	Attenuated at Wk2
GAD-7 (WAI Goal)	**−0.041 [−0.094, −0.004]***	37.1 %	−0.044 [−0.104, 0.005]	22.0 %	Attenuated at Wk2
PHQ-8 (WAI Task)	**−0.039 [−0.088, −0.006]***	28.2 %	**−0.045 [−0.102, −0.005]***	16.3 %	Consistent across waves
GAD-7 (WAI Task)	−0.028 [−0.073, 0.002]	25.1 %	**−0.048 [−0.106, −0.005]***	23.9 %	Strengthened at Wk2
PHQ-8 (WAI Bond)	−0.029 [−0.084, 0.007]	20.8 %	−0.005 [−0.051, 0.036]	1.9 %	NS at either wave
GAD-7 (WAI Bond)	−0.032 [−0.089, 0.004]	29.6 %	−0.003 [−0.051, 0.044]	1.4 %	NS at either wave

Note. Indirect effects = standardized OLS betas [95% CI], 5,000-iteration percentile bootstrap (*bold = CI excludes 0). WAI Task is the most temporally robust subscale: consistent (PHQ-8) or strengthened (GAD-7) at week 2. WAI Composite and Goal attenuate at week 2; WAI Bond non-significant at either wave in this specification.

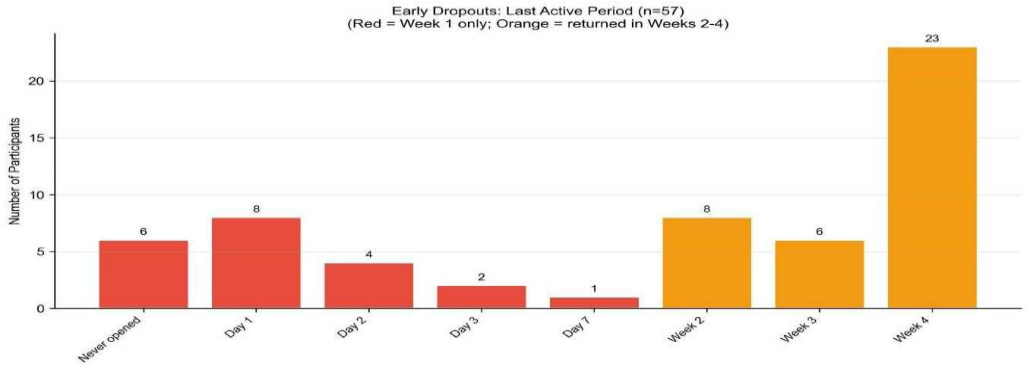
Supplementary Figures

Supplementary Figure S1. Behavioral-Only Cluster Heatmap (37-Variable Solution)



Mean standardized engagement by behavioral-only cluster (WAI excluded; N=200). ARI between primary and behavioral-only solutions = 0.392. Behavioral-only cluster labels are the exposure variable in all Aim 2 analyses (Tables 4, S2–S7) to avoid circularity.

Supplementary Figure S2. Dropout Timing: Sporadic Users Only (n=58)



Last active period for Sporadic Users. Red = active week 1 only; orange = returned briefly weeks 2–4. Of 58 Sporadic Users: 6 never opened the app; 21 active days 1–7 only; 37 (64%) returned briefly weeks 2–4. Largest group (n=23) last active at week 4, consistent with a "good-enough" disengagement pattern.

Supplementary Figure S3. Bootstrap Cluster Stability Distribution (1,000 iterations)

Figure 8: Figure S2. Dropout timing

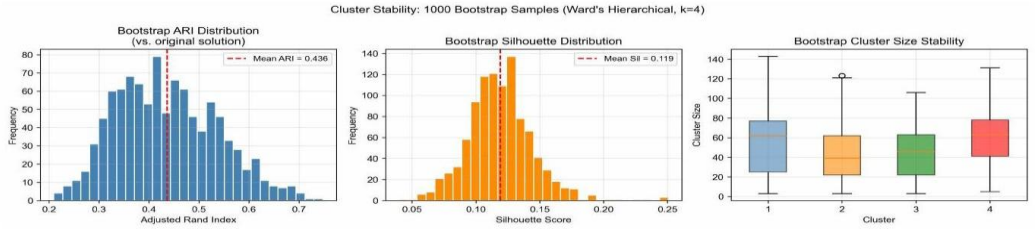


Figure 9: Figure S3. Bootstrap stability

ARI distribution across 1,000 bootstrap resampling iterations (primary 40-variable, week-1 WAI, k=4). Three panels: ARI distribution (mean = 0.438); silhouette distribution (mean = 0.119); cluster size stability boxplots. 0% exceeded ARI > 0.80. Supports engagement profile rather than discrete latent class interpretation.

Supplementary Figure S4. WAI Subscale Scores by Behavioral-Only Cluster (n=190)

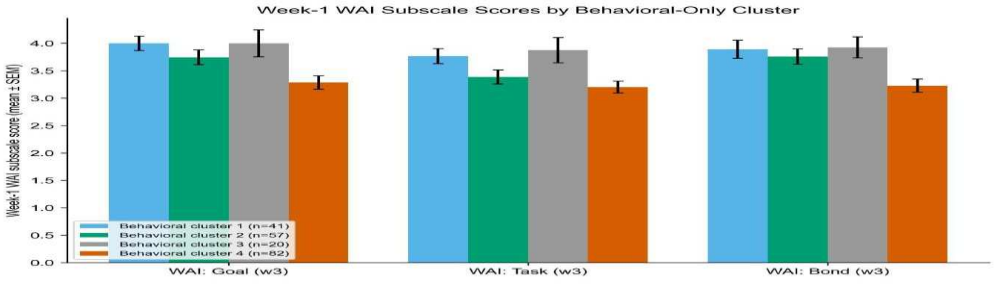


Figure 10: Figure S4. WAI subscales by behavioral-only cluster

Mean WAI subscale scores (Goal, Task, Bond; week-1 wave) by behavioral-only cluster. Error bars = ±1 SEM. All three subscales differentiate behavioral-only clusters despite WAI exclusion from the clustering algorithm, confirming co-variation between behavioral persistence and early alliance formation.