UNIVERSITY OF VIRGINIA

MASTER OF SCIENCE THESIS

Understanding the Relationship Between Engagement Markers and Psychosocial Outcomes in a Digital Mental Health Intervention for Anxiety

---

Author:

Ángel F. VELA DE LA GARZA

EVIA

Advisor:

Dr. Laura BARNES

Committee Members:

Dr. Mehdi Boukhechba

Dr. Bethany Teachman

A thesis submitted in fulfillment of the requirements

for the degree of Master of Science

in the

Department of Engineering Systems and Environment

School of Engineering and Applied Science

August 2022

UNIVERSITY OF VIRGINIA

Abstract

Department of Engineering Systems and Environment

Master of Science

Understanding the Relationship Between Engagement Markers and Psychosocial Outcomes in a Digital Mental Health Intervention for Anxiety

by Ángel F. VELA DE LA GARZA EVIA

As the prevalence of anxiety and depression continues to grow worldwide, digital mental health interventions (DMHIs) have played a key role in scaling and expanding the reach of mental health services in a cost-effective manner. Although studies have shown that DMHIs reduce symptom severity, low user engagement and high attrition rates limit the significance of these conclusions. Consequently, it is important to develop a better understanding of how engagement patterns relate to intervention outcomes. This research aims to understand the relationship between participant engagement and the psychosocial outcomes of anxiety and interpretation bias in MindTrails, a free web-based DMHI. In our work, we defined engagement markers based on completion rate and time spent on training and assessment components. We then extracted engagement features related to these markers from 697 participants who enrolled in the MindTrails Calm Thinking study. These features were used in a clustering analysis to identify two engagement pattern groups characterized by the amount of time spent in the intervention. After defining engagement groups, we developed multilevel models to investigate between-group differences in outcomes throughout the intervention. Our results demonstrate that while there were no significant differences in anxiety outcomes, both

engagement groups significantly differed in their improvement of certain interpretation bias outcomes. Overall, the findings highlight the complexity of using time-related engagement markers while furthering the understanding of participant interaction with DMHIs.

Acknowledgements

I am deeply grateful to my advisor, Dr. Laura Barnes. Thank you for your thoughtful guidance, cheerful support, and assertive feedback on this research study.

I would also like to thank Dr. Mehdi Boukhechba and Dr. Bethany Teachman for giving their expertise and time as committee members and for all they have taught me while working at MindTrails.

I would especially like to thank Jeremy Eberle for his incredible patience, outstanding mentorship, and endless support. You taught me so much and I could not have done this work without you.

To my colleagues at the Sensing Systems for Health Lab, especially Anna Baglione and Sonia Baee, thank you so much for your mentorship and companionship. I am also thankful for everyone from the MindTrails team. It has been an extraordinary pleasure working together. I am going to miss seeing you all every Friday.

Finally, I would like to thank my wife, Valeria, for always believing in me and pushing me to work harder every day. This is for you.

Contents

List of Figures

List of Tables

Dedicated to my wife, Valeria.

1 Introduction

In the recent years, the prevalence of anxiety and depression has continued to grow worldwide (Santomauro et al. 2021) leading to an increase in demand for mental health services (Bethune 2020). However, there are multiple barriers preventing individuals from receiving adequate mental health services, such as limited availability, inaccessibility to resources, and high cost of mental health care (Andrade et al. 2014). Digital mental health interventions (DMHIs) have the potential to reduce these barriers by scaling and expanding the reach of mental health services in a cost-effective manner (Newby et al. 2021). In addition, through the use of digital technology, DMHIs support behavioral change by encouraging healthy habits, helping individuals cope with long-term mental health conditions, and enabling online access to treatments (Murray et al. 2016; Michie et al. 2017). Although studies have shown that DMHIs for anxiety and depression have been successful in producing significant small-to-moderate effects on symptom reduction (Fu et al. 2020; Lehtimaki et al. 2021), high rates of attrition and low user engagement limit the true efficacy of these interventions and challenge the statistical significance of their conclusions (Gan et al. 2021; Linardon and Fuller-Tyszkiewicz 2020). This is emphasized in Eysenbach 2005's law of attrition that describes two types of attrition commonly present in digital health interventions. The first type, labeled as dropout attrition, relates to the proportion of participants who do not come back to complete follow-up assessments and are considered to have dropped out of the study. The second type, labeled as nonusage attrition, describes participants who continued working on the intervention, but disengaged with it over time. Nonetheless, low levels

of engagement leading to nonusage attrition do not necessarily imply that the intervention was not beneficial (Christensen and Mackinnon 2006). Usage and engagement could be influenced by the specific intervention dose that the participant needs in order to benefit, which is described by the dose-response relationship (McVay et al. 2019). Thus, researchers are increasingly interested in analyzing the association between dose-response effects and complex patterns of engagement to help understand the extent to which digital interventions are effective at achieving their intended goal (Nahum-Shani et al. 2022).

The dose of a digital health intervention refers to the amount of intervention sent to and received by the user at the given time. It has been conceptualized into the dose that the creators of the interventions want to provide, known as the intended dose, and the dose that the user receives and the actions that the user provides to the intervention, known as the enacted dose. Although the creators of the intervention can adjust the former, the latter is entirely dependent on the user (McVay et al. 2019). The enacted dose can be explained by engagement. Perski et al. 2017 conceptualized engagement in digital behavior change interventions as a two-part construct. The first part describes how the individual uses the intervention over time. In the literature, this has been quantified through the analysis of behavioral usage data obtained from self-report questionnaires, ecological momentary assessments, sensors, or system log data (Yardley et al. 2016). This quantitative conceptualization of engagement accounts for the intervention's amount, duration, breadth, and depth. Amount describes user interaction frequency; duration refers to the amount of time that the user is exposed to the intervention; breadth captures the number of features and pages accessed; depth accounts for the number of measures and modules completed by the user (e.g., doing self-report questionnaires) (Pham et al. 2019). The second part of the engagement construct focuses on the user's subjective experience while completing the intervention and accounts for the levels of affect, attention, and interest. Similarly, Nahum-Shani

et al. 2022 defined engagement as "a state of energy investment involving physical, affective, and cognitive energies directed toward a focal stimulus or task." The physical energy component refers to the action of doing a task like finishing a training session or answering an assessment questionnaire. The affective energy component encompasses the positive affective response of an individual while doing a task. Finally, the cognitive energy component describes the level of attention placed on the task. These definitions describe engagement as a multifaceted and dynamic construct, which makes it more difficult to comprehend and evaluate.

Donkin, Christensen, et al. 2011 conducted a systematic review on usage metrics and outcomes and reported that two of the metrics that have been extensively analyzed in digital health interventions are number of logins and number of modules completed. The latter was shown to have a greater relationship to improved outcomes in DMHIs; however, metrics like number of logins, measures completed, and time spent were not related to outcomes in DMHIs. Mixed findings and differences in engagement metrics across studies make it harder to determine how engagement metrics are associated with outcomes. Furthermore, confounding factors like user variability in enrollment motivation, self-regulation skills, and symptom severity may influence this association (Christensen and Mackinnon 2006; Yardley et al. 2016). Despite this fact, researchers agree that engagement is a crucial part of understanding effectiveness of DMHIs. For DMHIs to be adopted in real-world healthcare settings, further analysis is needed to comprehend the impact that engagement patterns have on outcomes.

This work aims to analyze the relationship between engagement markers and the psychosocial outcomes of anxiety and interpretation bias in MindTrails, a DMHI. The main objective of this study is to investigate whether distinct engagement patterns, extracted from markers that are characteristic of completion rate and time spent on intervention components, lead to differences in outcomes over time. If so, we aim to determine if a particular pattern is associated with better outcomes. We hypothesize

that patterns indicative of higher engagement will result in lower anxiety levels and greater improvements in interpretation bias. The paper is structured as follows. First, we review and describe the multiple approaches that have been taken to conceptualize engagement and assess its relationship to outcomes. Then, we explain the steps to cluster participants into engagement groups and develop multilevel models on each of the target outcomes. Finally, we present the results of the models and discuss the implications that our findings have for understanding participant engagement with DMHIs.

2 Related Work

Previous research has focused on understanding the relationship between engagement and outcomes in DMHI by assessing how individual metrics or patterns are associated with stronger outcomes. Donkin, Hickie, et al. 2013 analyzed individual engagement metrics and their impact on depressive symptoms in a DMHI for participants with depression and cardiovascular disease. In this study, number of modules and activities completed, time spent online, number of logins, and their composite metrics (e.g., average number of minutes per login) were measured; only number of activities completed per login turned out to be statistically related with a significant improvement in depression symptoms. Similarly, Zeng et al. 2020 studied how completion rate, frequency of items completed, and time spent on the intervention were related to depressive symptoms in a 3-month DMHI for participants with depression and HIV. Both higher completion rate and frequency of items completed were associated with significant symptom reduction at the end of the intervention; time spent had no significant relationship with the target outcome. Hanano et al. 2022 used number of practice logs completed and word count of weekly questionnaire responses to assess participants' levels of behavioral and attitudinal engagement, respectively. The number of weeks with at least five log entries and the number of weeks with a response that had a word count greater than average were significantly associated with decreased depression and anxiety symptoms. However, when conducting the same analysis on participants who completed at least one activity, none of the features were significantly related to symptom reduction.

Another approach to analyzing the relationship between engagement and outcomes has been to group users by engagement metrics and see if particular patterns lead to a significant change in symptom reduction. For example, Geramita et al. 2018 labeled users in an internet support group for anxiety and depression by number of logins and posting frequency. The group that contributed the most throughout the six-month intervention reported significant reductions in anxiety symptoms compared to the group that contributed the least. Enrique et al. 2019 studied differences in interaction behaviors in a DMHI for depression between users who obtained a reliable change in depressive symptoms versus those who did not. Higher exposure to the intervention was associated with users who obtained a reliable change in symptom reduction. This group of users significantly had higher levels of engagement in terms of time spent in the intervention, number of logins, features accessed, and program completion compared with the group whose symptoms did not improve. In addition, between-group differences in engagement decreased halfway through the intervention, highlighting the dynamic nature of engagement. Chien et al. 2020 explored engagement variability and patterns in a 14-week cognitive behavioral DMHI for anxiety and depression with over 50,000 participants. The log usage data, which stores participants' interactions with the different components of the intervention, was analyzed to calculate engagement features that included the number of modules completed, tools and sessions used, and weekly time spent. From these features, five distinct patterns of sustained engagement were found describing users who either had low engagement, varying levels of disengagement after initial engagement, or sustained high engagement throughout the course of the program. All patterns had some level of improvement, with a greater improvement associated with higher engagement. Sanatkar et al. 2019 identified three distinct engagement patterns by running a two-step clustering analysis on a DMHI designed for individuals with depression, anxiety, and stress. Five engagement metrics, which included number of logins, reminders received, and activities started and completed,

were used in the clustering algorithm. There were no significant differences between engagement patterns and outcomes. Finally, Li et al. 2022 ran a secondary analysis on the same dataset as Zeng et al. 2020, but focused on the relationship that completion rate and frequency of items completed had with depression, stress, and quality of life. Based on these two metrics, they clustered their participants into low and high engagement groups. Symptom reduction for depression and stress and the improvement of quality of life were greater for the high engagement group throughout the study; nonetheless, both groups reported having fewer symptoms. Between-group differences in symptom reduction widened over time, suggesting that the high engagement group may have benefited more from the intervention.

Although there is overlap in engagement metrics throughout studies, mixed findings and differences in study design and methodology make it difficult to validate the relationship between engagement and outcomes. Likewise, it is still unclear whether a specific set of engagement metrics or patterns lead to better results (Enrique et al. 2019), considering that some of the studies reported improvements in outcomes regardless of engagement behavior. It is still important to develop an objective knowledge of engagement with DMHI to understand what actions lead to behavioral change (Pham et al. 2019). By doing so, DMHIs can be improved to maximize the health benefits for the users.

3 Methods

3.1 MindTrails Calm Thinking Study

The MindTrails Project is a free web-based DMHI that provides cognitive bias modification for interpretation (CBM-I) training. This training is designed to help individuals with anxiety reduce their levels of negative interpretation bias, which refers to the tendency to assign a negative or threatening meaning to ambiguous situations. Negative interpretation bias has been associated with anxiety (Beard 2011). Individuals with anxiety are more likely to interpret ambiguous situations in a negative way (Mathews 2012). CBM-I training provides the opportunity to practice resolving ambiguous scenarios positively (MacLeod and Mathews 2012). By doing so, individuals start developing flexible and positive thinking patterns towards everyday situations.

The Calm Thinking study is a sequential multiple-assignment randomized controlled trial, which is part of the MindTrails program. It officially launched on March 19, 2019; enrollment for the study closed on April 1, 2020, and data collection concluded on November 27, 2020. The primary goal of this study was to assess the effectiveness of this DMHI in decreasing interpretation bias associated with anxious thinking. The study consisted of a pretreatment assessment, five training sessions, and a 2-month Post Follow-Up assessment. Training sessions were designed to take about fifteen minutes and participants were asked to complete one training session per week. At least five days had to pass before participants were able to begin the next training session and

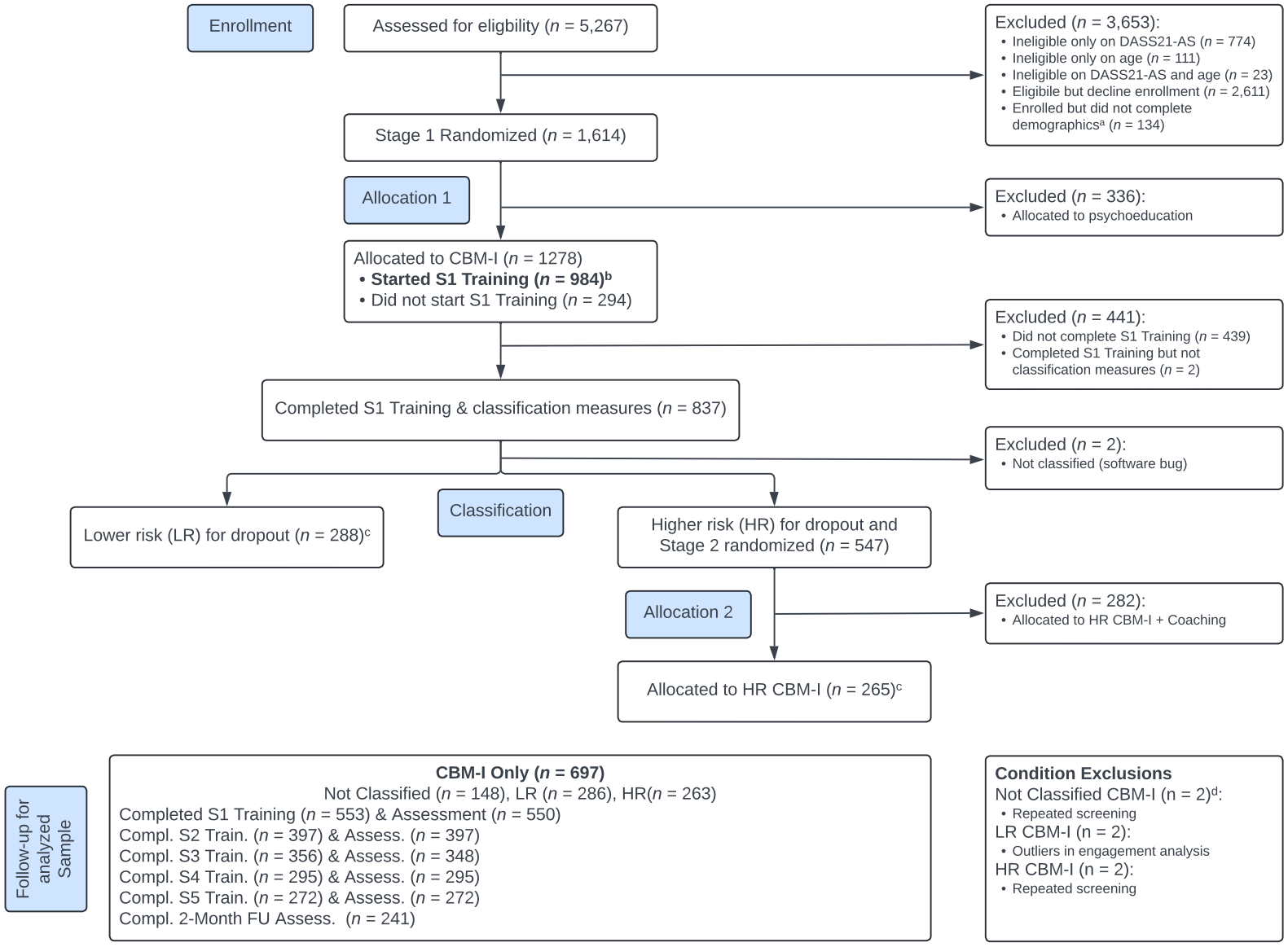
at least sixty days after completing all training sessions to start the Post Follow-Up assessment. The participants received $5 in gift cards after completing the pre-treatment, session 3, and session 5 assessments, and $10 after completing the Post Follow-Up assessment, for a total compensation of up to $25. More information about this study is described in the main outcomes paper (Eberle, Daniel, et al. 2022).

3.2 Participants

5,267 community participants were assessed for eligibility on the MindTrails project website. As per the pre-registered study on ClinicalTrials.gov (ID: NCT03498651), inclusion criteria consisted of being 18 years or older, endorsing having at least moderate anxiety levels by scoring ten or higher on the Depression, Anxiety, Stress Scales-Short Form: Anxiety Subscale (DASS-21 AS), and having Internet access on a mobile device or computer. A total of 1,614 participants met the eligibility criteria, gave their informed consent, made an account, provided their baseline demographic information, and were randomized into CBM-I ($n = 1,278$) and psychoeducation ($n = 336$) conditions. Out of those participants in the CBM-I condition, 984 started the first training session, from which 837 completed the training and classification measures. These participants were then assigned to low ($n = 288$) and high ($n = 547$) risk for dropout conditions. Participants in the high risk for dropout group were then randomized into high-risk CBM-I ($n = 265$) and high-risk CBM-I plus coaching ($n = 282$) conditions. For the engagement analysis, we focused on participants who had been assigned to a CBM-I condition and had started the first training session. We excluded participants in the psychoeducation and high-risk CBM-I plus coaching conditions because engagement was not comparable to those who only received CBM-I (e.g., receiving coaching could be a confounding variable in helping participants engage less or more with the intervention). We also excluded participants with repeated eligibility screenings ($n = 2$), who were not classified

despite having completed the first training session and classification measures ($n = 2$), and who were identified as outliers in most of the engagement markers ($n = 2$). With these exclusions, the engagement analyzed sample consisted of 697 participants. For a detailed breakdown, see CONSORT diagram in Figure 3.1 adapted from the Calm Thinking main outcomes paper (Eberle, Daniel, et al. 2022).

FIGURE 3.1: CONSORT Diagram



Note. Adapted from the Calm Thinking main outcomes paper CONSORT diagram. Analysis exclusions are included in flow but not analyzed. S1-5 = Session 1-5; FU = Follow-Up.[a] needed for stratification. [b] analyzed sample before exclusions. [c] condition classification count before exclusions. [d] 1 did not start S1 training; 1 started S1 training but did not complete classif. measures.

3.3 Cognitive Bias Modification for Interpretation (CBM-I) Task

CBM-I training consists of repeated practice in solving ambiguous situations to reinforce a selective pattern of interpretation, which for the most part is positive (Mathews and Mackintosh 2000; MacLeod and Mathews 2012). The Calm Thinking study consisted of five CBM-I training sessions. Each of the five sessions consisted of forty unique scenarios resolved positively 90% of the time and negatively 10% of the time. Most of these training scenarios were adapted from past research and work of Mathews and Mackintosh 2000, Steinman and Teachman 2010, and Ji et al. 2021, and portrayed day-to-day situations that could cause anxiety. These situations were presented ambiguously and resolved in either a threatening (negative) or non-threatening (positive/benign) way. Of the scenarios used, 50% were related to social anxiety, 20% to psychophysical or physical anxiety symptoms, 10% to health anxiety, and 20% to anxious thinking. Each scenario consisted of three parts: scenario title and image, description, and reading comprehension question(s). For example, after being presented with the scenario title (e.g., Spotting a neighbor) and visual image, participants read the scenario description (e.g., "As you are walking down a crowded street, you see your neighbor on the other side. You call out, but they do not answer you. Standing there in the street, you think that this must be because they were... distr_cted" [positive]). The last word of each scenario had a missing letter(s) (e.g., "a" in the above scenario); participants selected the letter(s) that correctly completed the word fragment in order to continue. As participants advanced throughout the training, the difficulty of these exercises increased, such that only one letter was chosen in Sessions 1 and 2, two letters in Sessions 3 and 4, two letters for half of the scenarios in Session 5, and the last word was filled for the remaining half of the scenarios. Following this, participants answered a reading comprehension question (e.g., "Did your neighbor purposely ignore your call to them in the street?") that

reinforced the positive or negative meaning of the scenario. The format of the comprehension questions varied across sessions. Participants answered a yes/no question in Sessions 1, 3, and 5 or chose one of two interpretations that completed a sentence statement related to the scenario storyline in Sessions 2 and 4. Each participant had multiple chances to answer the comprehension question correctly to advance.

3.4 Measures

Anxiety Symptoms Measures

Depression, Anxiety, Stress Scales-Short Form: Anxiety Subscale (DASS-21 AS)

The DASS-21 AS (adapted from P. Lovibond and S. Lovibond 1995) is used to assess the negative emotional state of anxiety. Seven anxiety symptom statements are presented to participants, and they are asked to rate the extent to which each statement has applied to them over the past week. Each response is measured on a scale from 0 ("Not at all") to 3 ("Most of the time"). Scores closer to 3 indicate higher anxiety symptoms. The DASS-21 AS was used as the eligibility screener and measured at four timepoints: baseline (eligibility screener), Session 3, Session 5, and two-month Post Follow-Up. Internal consistency for the analyzed sample using complete item-level data at baseline (eligibility screener) was acceptable, with a Cronbach's $\alpha = 0.727$.

Overall Anxiety Severity and Impairment Scale (OASIS)

OASIS (adapted from Norman et al. 2006) is a short five-item measure used to assess anxiety severity and related impairment. Anxiety severity is captured through items asking about anxiety frequency and intensity. Impairment is captured through items asking about avoidance and work and social interference. Each item is measured on a scale of 0 (lowest impairment/severity) to 4 (highest impairment/severity), with higher

scores indicating more severe and impairing anxiety. The OASIS was measured at all six timepoints: baseline (pretreatment), Sessions 1-5, and two-month Post Follow-Up. Internal consistency for the analyzed sample using complete item-level data at baseline (pretreatment) was good, with a Cronbach's $\alpha = 0.827$.

3.4.1 Interpretation Bias Measures

Recognition Ratings Task (RR)

The RR task (modified from Mathews and Mackintosh 2000) was designed to measure interpretation bias. In this task, participants were first asked to read and imagine themselves in nine emotionally ambiguous scenarios describing social situations and then assess their interpretation of each of these scenarios. These scenarios are similar to those in CBM-I training. The only differences are that a title is included for each scenario and that the scenarios remain ambiguous even after correctly completing the word fragment exercise. An example scenario in this task is "The Loud Noise: You are woken up in the middle of the night by a loud noise. You are not sure what caused the noise and leave your bedroom to see what happened. You walk... downst_irs." After correctly completing the last word (e.g., selecting "a" to complete "downstairs"), participants answered a reading comprehension question for each of the nine scenarios (e.g., "Have you been woken up in the middle of the night?"). Following the comprehension questions, for each scenario, participants saw the title (e.g., THE LOUD NOISE), a starting sentence (e.g., "As you walk downstairs..."), and four disambiguated interpretations. Of the four disambiguated interpretations, two were threat-related (negative: "You feel afraid, and worry that you cannot handle the fear."; positive: "You feel afraid, but you know that you can tolerate the feeling.") and the other two were threat-unrelated (negative: "You feel cold, and think about how the house needs better heating."; positive: "You feel happy, and think about how lovely your house is.").

Participants rated how similar each disambiguated interpretation was to the original ambiguated scenario. The rating was done on a scale of 1 ("Very different") to 4 ("Very similar"); scores closer to four indicate greater levels of negative and positive interpretation bias, respectively. The RR measure was assessed at four timepoints: baseline (pretreatment), Session 3, Session 5, and two-month Post Follow-Up. Following Ji et al. 2021, the negative interpretation bias score was calculated by averaging participants' endorsements of threat-relevant negative disambiguated interpretations at each timepoint. Similarly, the positive interpretation bias score was calculated by averaging participants' endorsements of threat-relevant positive disambiguated interpretations at each timepoint. Internal consistency for the analyzed sample using complete item-level data at baseline (pretreatment) was acceptable for both negative interpretation bias (Cronbach's $\alpha = 0.743$) and positive interpretation bias (Cronbach's $\alpha = 0.730$).

Brief Body Sensations Interpretations Questionnaire (BBSIQ)

The BBSIQ (modified from Clark et al. 1997) is a 14-item measure that assesses negative interpretation bias for internal body sensations and external events. Participants are presented with fourteen situations (e.g., "A friend suggests that you change the way that you're doing a job in your own house. Why?") that are ambiguous and potentially threat-related. For each situation, three different explanations for why the situation could have taken place are shown. One explanation is always negative (e.g., "They think you're incompetent."); the other two are either positive (e.g., "They are trying to be helpful.") or neutral (e.g., "They have done the job more often and know an easier way."). Participants rated how likely they considered the explanation to be true had they found themselves going through the situation on a scale of 0 ("Not at all likely") to 4 ("Extremely likely"). Scores closer to four indicate greater levels of negative interpretation bias. The BBSIQ was assessed at four timepoints: baseline (pretreatment), Session

3, Session 5, and two-month Post Follow-Up. The negative interpretation bias score was calculated by averaging participants' likelihood ratings for all negative explanations at each timepoint (following Steinman and Teachman 2010; Steinman and Teachman 2015; Ji et al. 2021). Internal consistency for the analyzed sample using complete item-level data at baseline (pretreatment) was excellent, with a Cronbach's $\alpha = 0.904$.

3.4.2 Engagement Markers

We defined engagement markers based on intervention completion rate (Li et al. 2022; Baee et al. 2022) and time spent on training components and assessment measures (Eberle, Meyer, et al. 2018; Baee et al. 2022). Other studies have used frequency of use as an engagement marker that includes features like number of logins and number of activities completed. The training sessions for this study were designed to be completed in one sitting. As expected, most participants, on average, logged in once per training session. In addition, all participants had to complete the same number of training and assessment items, so there was not going to be much variability in number of activities completed that was not already accounted for in completion rate. Due to this fact, we did not include the features related to frequency of use.

Intervention completion rate

The MindTrails Calm Thinking study consisted of 5 training sessions and 63 assessment measures for a total of 68 intervention components. The intervention completion rate for a given participant was calculated by dividing the number of training sessions and assessment measures completed by the total number of intervention components. This feature accounts for the number of training sessions and assessment measures completed.

Time spent on training components

Training component features were calculated using reaction time, defined as the time it takes for a participant to get a correct response. The three training features analyzed were the lemon exercise, anxious imagery prime exercise, and individual CBM-I training scenarios. Details for these features are provided below. For participants with repeated entries in a given training component item, the average reaction time for that specific item was calculated first, unless they were duplicated entries for which we removed the duplicated values.

Time spent on the Lemon Exercise: This exercise was given in the first training session. Its goal is to help participants practice imagination-based thinking. Participants are asked to use all senses to imagine what it feels like to hold a lemon in their hand. Time spent was calculated by adding the reaction times from start to whatever point in the exercise the participant reached.

Time spent on the Anxiety Imagery Prime Exercise: This exercise was given in the first training session. Participants are asked to imagine themselves in an anxiety-provoking situation that they are most likely to experience. Time spent was calculated by adding the reaction times from start to whatever point in the exercise the participant reached.

Average time spent on individual CBM-I training scenarios across training sessions: Each CBM-I training session consisted of 40 unique scenarios. A scenario is broken down into three components: the scenario title and image, the description of the scenario with the word fragment, and at least one reading comprehension question. Participants need to complete the word fragment correctly and answer the comprehension question(s) to advance. First, time spent on an individual scenario was calculated by adding together the reaction times of the three scenario components. Then, we averaged the time spent on the available scenarios for a given training session. Finally, we used those values to get the average time spent on individual CBM-I training scenarios across training

sessions for a given participant.

Time spent on assessment measures

Time spent on completing assessment measures was recorded throughout the study. We analyzed measures that appeared before the start of the first training session so that participants would have at least one data point. For repeated measures, the average was calculated adjusting for the total number of measures that the participant had completed. The analyzed measures together with a brief description of what they are meant to assess are summarized in Table 3.1 (assessments are listed in alphabetical order). Time spent on BBSIQ was removed from the analysis due to its high correlation (Pearson correlation value greater than 0.7) to time spent on RR, Mechanisms, and Wellness measures.

TABLE 3.1: Assessment measures for which time spent was calculated. Descriptions come from the Calm Thinking Measures Appendix (ver. 3).

Assessment Measure	**What does it assess?**
Anxiety Identity	Identification between anxiety and self
Anxiety Triggers	Levels of anxiety in certain situations
BBSIQ	Interpretation bias
Comorbid	Depressed mood and influence of drinking
Credibility	Belief that the intervention will help
DASS-21 AS	Anxiety
Demographics	Socio-demographic characteristics
Mechanisms	Cognitive flexibility/reappraisal, avoidance, uncertainty intolerance
Mental Health History	Past and present mental health treatment history
OASIS	Anxiety
Pre-Affect	Current levels of anxiety before training starts
RR	Interpretation bias
Technology Use	Frequency of what devices are used
Wellness	Life satisfaction, self-efficacy, changes in thinking, optimism

3.4.3 Covariates

We included the response to one of the two questions in the first part of the Readiness Rulers measure (modified from Borkovec and Nau 1972) assessed before the start of the first training session (pretreatment). This question rated participants' confidence that the DMHI will be effective in helping them reduce their anxiety on a scale of 0 ("Not at all" confident) to 4 ("Very" confident). The exact question read: "How confident are you that an online training program will reduce your anxiety?" Past MindTrails studies have shown that there is evidence to suggest that greater baseline confidence ratings were associated with lower dropout rates and greater improvements on certain outcomes (Hohensee et al. 2020). Based on these results, we added this question response as a covariate variable to account for the influence that confidence in DMHIs may or may not have on participants' interaction with the program and overall outcomes. We refer to this covariate variable as credibility online throughout the work.

3.5 Statistical Analysis

Intermediately clean data was obtained from the OSF of the MindTrails Calm Thinking study generated from version 1.0.0 of Eberle, Baee, et al. 2022's centralized data cleaning script. Statistical significance was set a $p < 0.05$. All analysis, except for data imputations, were done using `R` (version 4.1.1; R Core Team 2021). Data imputations were done using `Blimp Studio` (version 1.3.6; Enders, Keller, et al. 2022). Analysis and coding scripts followed the work of Eberle, Boukhechba, et al. 2020.

Outlier Detection and Handling

The outlier detection and handling procedure consisted of three parts: outlier detection and removal of individual CBM-I training scenarios, outlier detection for time engagement features, and capping of extreme values. We used the median absolute deviation for outlier detection since it is a more robust measure of dispersion affected less by outliers as compared to the mean and standard deviation (Leys et al. 2013). We defined an outlier as a value that was three median absolute deviations away from the median. First, we identified training scenarios within participants' training sessions whose time spent was flagged as an outlier. Approximately 10% of the scenarios were marked as outliers. We removed these scenarios (following Eberle, Meyer, et al. 2018) and calculated the average time spent on individual CBM-I training scenarios across sessions. Through a boxplot visualization, we found one participant that only completed two training scenarios and spent more than 20 minutes on them, so we decided to treat this participant as if they had done no scenarios. Second, we kept track of the participants who were flagged as outliers for each of the time-related engagement marker features and calculated the proportion of outlier features for each participant. We removed two participants from the engagement analysis since they were outliers in more than 70% of the time-related engagement marker features. Finally, in order to reduce the impact that outliers could have in the clustering analysis (Guha et al. 1998) while still reflecting patterns of spending too little or too much time on training and assessment measures, we capped extreme values. The bottom 1% was set equal to the 1st percentile and the top 1% to the 99th percentile. We applied this procedure to all of the time-related engagement marker features.

Cluster Analysis

Clustering is an unsupervised learning method commonly used as an exploratory classification technique to find unknown subgroups in data. We were interested in grouping participants based on engagement markers to identify distinct engagement patterns. To do this we evaluated three clustering algorithms: K-means, partitioning around medoids (PAM), and agglomerative hierarchical clustering.

In K-means clustering, one must first specify the number of clusters K to then partition the data into K distinct, non-overlapping clusters. This partitioning minimizes the total within-cluster variation (defined using squared Euclidean distance). K-means clustering is done in two iterative steps. First, the algorithm is initiated by randomly assigning all observations to a number from 1 to K. Then, the centroid is calculated for the K clusters, and observations are reassigned to the cluster with the nearest centroid. This second step is repeated until the observation assignments no longer change, indicating that the model has converged (James et al. 2013). The K-medoids algorithm, PAM, tries to minimize the sum of dissimilarities between observations making it more robust than K-means to noise and outliers (Madhulatha 2011). Instead of calculating the centroids for each cluster as in K-means, actual data points are selected as medoids. A medoid is a data point in a cluster with the lowest average dissimilarity to all the other points. The PAM algorithm consists of two phases: build and swap. K points are randomly initiated in the build phase as medoids and observations are assigned to the cluster with the closest medoid. A different non-medoid data point is randomly selected as a medoid in the swap phase and swapped with the initial medoid. If the swapping minimizes the objective function, then a new set of medoids are defined, and observations are reassigned. This process is repeated until the medoids stop changing (Schubert and Rousseeuw 2019). For both of these clustering methods, we used the Euclidean distance.

Agglomerative hierarchical clustering is another common clustering technique. Each observation starts as an individual cluster, and the pairwise inter-cluster dissimilarities are calculated using a distance metric such as Euclidean distance. Then, the algorithm proceeds to merge the two clusters with the highest similarity. The new pairwise inter-cluster dissimilarities are calculated for the remaining clusters based on the linkage criterion, which is a function of the distance metric. This is repeated until all of the observations fall under one single cluster (James et al. 2013). Finally, the outcome is visualized in a dendrogram. To extract the clusters, one must specify at what height to cut the dendrogram. It is important to note that different distance metrics and linkage methods yield varying clustering results. For our analysis, we used Euclidean distance and Ward's linkage.

Before clustering, we visualized the histogram of the time-related engagement marker features and noticed that most had a right-skewed distribution due to outliers. Hence, we log-transformed these features to get their distributions closer to normal and proceeded to standardize all of the engagement marker features, including completion rate. We assessed internal and stability validation metrics and visualized the cluster partitions to determine which clustering algorithm to use. We used the `clValid` package (ver. 0.7; Brock et al. 2008) to calculate internal and stability validation measures for the different clustering algorithms with two to four clusters. The internal stability measures calculated were connectivity, silhouette width, and Dunn index. These measures account for the compactness, connectedness, and separation of the clustering groups. Connectivity values should be minimized, while silhouette width and Dunn index values should be maximized (Brock et al. 2008). The stability validation measures calculated were average proportion of non-overlap (APN), average distance (AD), average distance between means (ADM), and figure of merit (FOM). These stability measures compare the results from clustering with all the features versus clustering by removing one feature at a time. All of these measures should be minimized (Brock et al. 2008). We

ran the `fviz_cluster` function in the `factoextra` package (ver. 1.0.7; Kassambara and Mundt 2020) to visualize the different clusters in two dimensions using principal components analysis.

Handling Missing Data

There were two patterns of missing data encountered in the analysis. A non-monotone missing data pattern at the item level resulted from participants' endorsement of the "prefer not to answer" option choice. Nonetheless, this was infrequent; across the five outcomes of interest, the proportion of item responses with missing data ranged from 0.10% to 0.62%. For cases with item level missing data, the mean of the available items was used to calculate participants' scale scores following Eberle, Boukhechba, et al. 2020.

A monotone missing data pattern at the scale level resulted from attrition; participants who dropped out of the study did not complete the remaining intervention assessments. For the analyzed sample, the proportions of missing data at the scale level across the five outcomes ranged from 55.7% to 57.6% (see Table 3.2 for descriptive statistics of outcomes over time for analyzed sample and engagement groups). Therefore, we opted to impute missing scale scores given the degree of missing data. To determine auxiliary variables to include in the imputation analysis, we analyzed differences in the mean proportion of missing assessments on categorical demographic features (following Eberle, Daniel, et al. 2022). Features were considered potential auxiliary variables if their mean proportion difference between two levels with more than 100 participants was greater than 0.1. Device used throughout the intervention (one device vs. multiple devices) and gender (male vs. female) were the only features that met these criteria. Device used had a mean proportion difference that was greater than 0.2. Gender's mean proportion difference was greater than 0.1; a one-way ANOVA test confirmed that this

TABLE 3.2: Descriptive Statistics of Outcomes by Engagement Group Over Time for Analyzed Sample

Outcome	Timepoint	Analyzed Sample			Less Time Spent			More Time Spent		
		n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
OASIS	Baseline [a]	697	2.31	0.70	386	2.28	0.71	311	2.34	0.69
	Session 1	549	2.19	0.74	288	2.18	0.74	261	2.20	0.73
	Session 2	397	1.74	0.74	219	1.72	0.74	178	1.76	0.73
	Session 3	356	1.76	0.79	199	1.79	0.82	157	1.73	0.76
	Session 4	295	1.56	0.82	164	1.56	0.81	131	1.56	0.84
	Session 5	272	1.54	0.78	152	1.58	0.77	120	1.50	0.78
	Post Follow-Up	244	1.57	0.83	140	1.59	0.81	104	1.54	0.86
DASS21	Baseline [b]	697	1.61	0.54	386	1.65	0.54	311	1.56	0.53
	Session 3	355	1.00	0.61	198	1.06	0.62	157	0.92	0.59
	Session 5	272	0.86	0.57	152	0.95	0.56	120	0.75	0.57
	Post Follow-Up	244	0.81	0.60	140	0.87	0.59	104	0.73	0.60
BBSIQ	Baseline [a]	697	1.54	0.84	386	1.59	0.84	311	1.46	0.83
	Session 3	347	0.89	0.73	195	0.93	0.74	152	0.84	0.72
	Session 5	270	0.79	0.68	151	0.83	0.71	119	0.74	0.63
	Post Follow-Up	240 [c]	0.79	0.62	137	0.88	0.66	103	0.66	0.54
RR Negative Bias	Baseline [a]	696 [d]	2.31	0.53	386	2.31	0.52	310	2.31	0.53
	Session 3	351	2.87	0.46	196	2.86	0.48	155	2.88	0.44
	Session 5	270	2.87	0.48	151	2.88	0.45	119	2.86	0.52
	Post Follow-Up	241	2.75	0.47	137	2.77	0.47	104	2.73	0.47
RR Positive Bias	Baseline [a]	696 [d]	2.90	0.53	386	2.95	0.54	310	2.83	0.51
	Session 3	352	2.53	0.55	197	2.52	0.57	155	2.53	0.54
	Session 5	270	2.55	0.55	151	2.56	0.57	119	2.55	0.54
	Post Follow-Up	241	2.52	0.55	137	2.57	0.54	104	2.45	0.56

[a] Assessed during pretreatment
[b] Assessed during eligibility screener
[c] One participant endorsed "prefer not to answer" for all BBSIQ items at Post Follow-Up
[d] One participant endorsed "prefer not to answer" for all RR items at baseline

difference was significant. Based on this analysis, we included these two features as auxiliary variables and assumed that data are missing at random for the imputation and multilevel models. For the imputation analysis, gender was collapsed into male, female, and transgender/other, with prefer not to answer responses treated as missing values; device used was collapsed into one device (e.g., desktop only) and multiple devices (e.g., mobile and desktop) and was then refactored as a binary outcome (one device or not) to help with model convergence.

To impute missing scale scores we used `Blimp Studio` (version 1.3.6; Enders, Keller,

et al. 2022) to conduct fully Bayesian model-based multiple imputations (Enders, Du, et al. 2020) following Wyatt et al. 2021. We built a separate imputation model for each target outcome that accounted for the multilevel structure of our data (Grund, Lüdtke, et al. 2018). By study and analysis design, assessment time, engagement group, and device used were complete; therefore, we specified these variables as fixed to assist in model convergence and computation speed (Keller and Enders 2021). Credibility online was not specified as a fixed variable since it had missing values. Following Eberle, Boukhechba, et al. 2020, time was represented as two piecewise linear trajectories: one for the training trajectory labeled as time_{TR} (baseline to Session 5) and the other for the Post Follow-Up trajectory labeled as time_{FU} (Session 5 to Post Follow-Up). These two trajectories were coded differently depending on the number of assessment timepoints. For the OASIS outcome, which was assessed at all timepoints, time_{TR} was coded as 0 for Baseline, 1-5 for Sessions 1 through 5, and 5 for Post Follow-Up; time_{FU} was coded as 0 for Baseline through Session 5 and 1 for Post Follow-Up. For the rest of the outcomes, which were assessed at the same four timepoints, time_{TR} was coded as 0 for Baseline, 3 for Session 3, 5 for Session 5, and 5 for Post Follow-Up; time_{FU} was coded as 0 for Baseline, Session 3, and Session 5, and 1 for Post Follow-Up. This coding scheme (0,3,5,5) interprets the time_{TR} slope the same across all outcomes; the results are still in terms of going from one session to the next. For the multilevel model specification, we included the fixed effects of engagement group, time_{TR}, time_{FU}, engagement group $\times$ time_{TR}, engagement group $\times$ time_{FU}, credibility online, gender, device used, a random intercept, and random slopes for time_{TR} and time_{FU}. For the variable specification, we defined device used as ordinal, and gender and engagement groups as nominal. The variable credibility online was grand mean centered prior to the imputations.

For each outcome, we imputed 100 datasets. Imputations were saved every 5,000 iterations after the burn-in period, which varied between outcomes and ranged from 10,000 to 100,000 burn-in iterations (OASIS: 10,000; DASS21: 20,000; BBSIQ: 50,000; RR

Negative Bias: 100,000; RR Positive Bias: 50,000). These burn-in iterations were selected based on convergence diagnosis. We diagnosed convergence by checking that the split-chain potential scale reduction factor was less than 1.05 at the final burn-in interval and that the effective number of Markov chain Monte Carlo (MCMC) samples was greater than 100 (Keller and Enders 2021). We used the default software imputation configurations: two MCMC chains with random starting values, homogeneous within-cluster variances, and priors for the dependent (`prior2`) and predictor (`xprior2`) variables (Keller and Enders 2021). Given that the imputation models produced values outside the permitted scale range, we assessed the mean percentage of out-of-range values for each imputed dataset (see Table S2). For any scale at a given timepoint, the mean percentage did not go over 12%. We considered these values not to be large enough to inflate variance estimates (Enders 2010).

Finally, a monotone missing data pattern at the session level for time-related engagement features also resulted from attrition. Rather than imputing the data, we took the mean of the available times. If a participant had no time values in a given feature, we considered time spent to be zero (e.g., participants who did not complete any CBM-I training scenarios).

Multilevel Modeling

Multilevel models for repeated measure studies have been used in psychotherapy research to understand different within-person and between-person trajectories over time. One advantage of these types of models is that they account for the hierarchical structure of the data (e.g., measures assessed at different time points nested in participants) (Tasca and Gallop 2009). For our research, in order to understand the relationship between engagement groups and outcomes over time, we ran five separate multilevel models, using the `nlme` package (ver. 3.1-152; Pinheiro et al. 2021). Each model was fit

by maximizing the restricted log-likelihood. In order to handle convergence errors we adjusted some of the control parameters in `lmeControl` following advice from Wyatt et al. 2021 and Brown 2021. For the OASIS, DASS21, BBSIQ, and RR Positive Bias multilevel models, we switched from the default `nlminb` optimizer to the `optim` optimizer. For the RR Negative Bias model we used the `optim` optimizer, increased the maximum number of iterations for the optimization algorithm (`msMaxIter = 1e9`), and increased the number of iterations for the expected maximization algorithm (`niterEM = 1000`). For each multilevel model, we segmented time into two piecewise linear trajectories: one for the training trajectory labeled as time_{TR} (baseline to Session 5) and the other for the Post Follow-Up trajectory labeled as time_{FU} (Session 5 to Post Follow-Up). We included the fixed effects of engagement group, time_{TR}, time_{FU}, engagement group $\times$ time_{TR}, engagement group $\times$ time_{FU}, and credibility online, a random intercept, and random slopes for time_{TR} and time_{FU}. The covariate variable credibility online was grand mean centered. Engagement group was dummy coded with "Less Time Spent" as the reference group (0 = Less Time Spent, 1 = More Time Spent).

Since we preformed data imputations, for each outcome we pooled the results of the imputed datasets with the `mitml` package (ver. 0.4-3; Grund, Robitzsch, et al. 2021), which follows Rubin's rules (Hayes 2019). We adjusted the `df.com` parameter in the `testEstimates` function with Barnard and Rubin's (1999) procedure to cap the degrees of freedom at values below those had we worked with complete data. Parameter final estimates are reported in terms of unstandardized b, together with their respective 95% confidence interval obtained from the `confint` function.

4 Results

4.1 Baseline characteristics

Baseline demographic characteristics for the 697 participants are shown in Table S1. The mean age of the analyzed sample was 35 years. The majority of the participants were female (80.9%), White/European origin (70.4%), Not Hispanic or Latino (81.8%), and from the United States (91.7%).

4.2 Cluster Analysis

From visual inspection of Figure S2, we noticed that the three algorithms with four clusters and hierarchical clustering with three clusters produced a group of participants who, for the most part, stopped at the first session, meaning that only outcome baseline data were available. This type of grouping raised imputation convergence issues; therefore, we did not select these options. For the internal validation measures, hierarchical clustering with two clusters has a similar connectivity value to K-means and PAM with two clusters but lower Dunn Index and Silhouette values; hence, we did not select this option. K-means has lower values in most stability measures than PAM and better connectivity and silhouette values. After analyzing these two figures, we chose the K-means algorithm as our clustering method (see Figure S1 for internal and stability validation values). We later ran the `NbClust` package (ver. 3.0; Charrad et al. 2014) that

uses a variety of indices to calculate the optimal number of clusters for K-means, which turned out to be two. In summary, we chose the K-means algorithm with two clusters.

We conducted a Wilcoxon rank-sum non-parametric test to compare these clusters since our engagement marker features were not normally distributed. Results from this test showed that there were significant differences between the two clusters for all of the features except completion rate (see Table 4.1 for test results and descriptive statistics). Furthermore, we visualized by engagement group the boxplot of time spent on training components (see Figure S3) and time spent on assessment measures (see Figure S4), the density distributions of time spent on assessment measures (see Figure S5), and the density distributions of completion rate (see Figure S6). From this visual inspection, we noticed that the completion rate was balanced in both groups. There was also a difference in time spent across all features, such that one group spent less time in intervention components than the other group. Therefore, we labeled these clusters as "Less Time Spent" and "More Time Spent," respectively.

TABLE 4.1: Descriptive Statistics of Engagement Markers by Engagement Group

Engagement Marker	Less Time Spent (n = 386) Median (IQR)	More Time Spent (n = 311) Median (IQR)	Wilcoxon rank-sum test	p
Completion Rate	0.57 (0.75)	0.49 (0.69)	58520	.56
Time spent on training components				
Average time spent on scenarios across sessions (sec)	11.18 (6.19)	13.88 (5.56)	43202.5	<.001
Time spent on lemon exercise (min)	0.88 (0.61)	1.44 (0.91)	31077	<.001
Time spent on the anxiety imagery prime exercise (min)	1.66 (0.82)	2.54 (1.40)	25833	<.001
Average time spent on assessment measures (min)				
Anxiety Identity	0.19 (0.10)	0.32 (0.16)	18250	<.001
Anxiety Triggers	0.66 (0.33)	1.23 (0.67)	14713	<.001
Comorbid	0.43 (0.21)	0.71 (0.30)	16564.5	<.001
Credibility	0.26 (0.31)	0.85 (0.65)	19807	<.001
DASS-21 AS	0.53 (0.22)	0.83 (0.42)	17764.5	<.001
Demographics	1.05 (0.38)	1.72 (0.76)	15187	<.001
Mechanisms	0.55 (0.24)	0.89 (0.45)	14127.5	<.001
Mental Health History	1.07 (0.50)	1.97 (1.07)	13317.5	<.001
OASIS	0.45 (0.15)	0.73 (0.31)	11735.5	<.001
Pre-Affect	0.13 (0.06)	0.21 (0.12)	20579	<.001
RR	2.58 (0.91)	4.16 (1.53)	11246.5	<.001
Technology Use	0.25 (0.11)	0.40 (0.19)	21272	<.001
Wellness	0.70 (0.28)	1.24 (0.50)	9420.5	<.001

4.3 Longitudinal Piecewise Linear Multilevel Model Results

The pooled results for each piecewise linear multilevel model are shown in Table 4.2. For each outcome, the estimated means of the piecewise linear trajectories at mean level of credibility online over time by engagement group are displayed in Figure 4.1. In addition, for outcomes that did have significant interactions between the different time trajectories and engagement groups, we built separate simple time effects models for the two engagement groups to further understand this relationship (see Table 4.3). For the simple time effects models we used the same outcome multilevel model parameter configurations and included the fixed effects of time_{TR}, time_{FU}, and credibility online grand mean centered, a random intercept, and random slopes for time_{TR} and time_{FU}.

Anxiety

For the OASIS and DASS21 outcomes, there were no significant interactions between the different linear time trajectories and engagement groups.

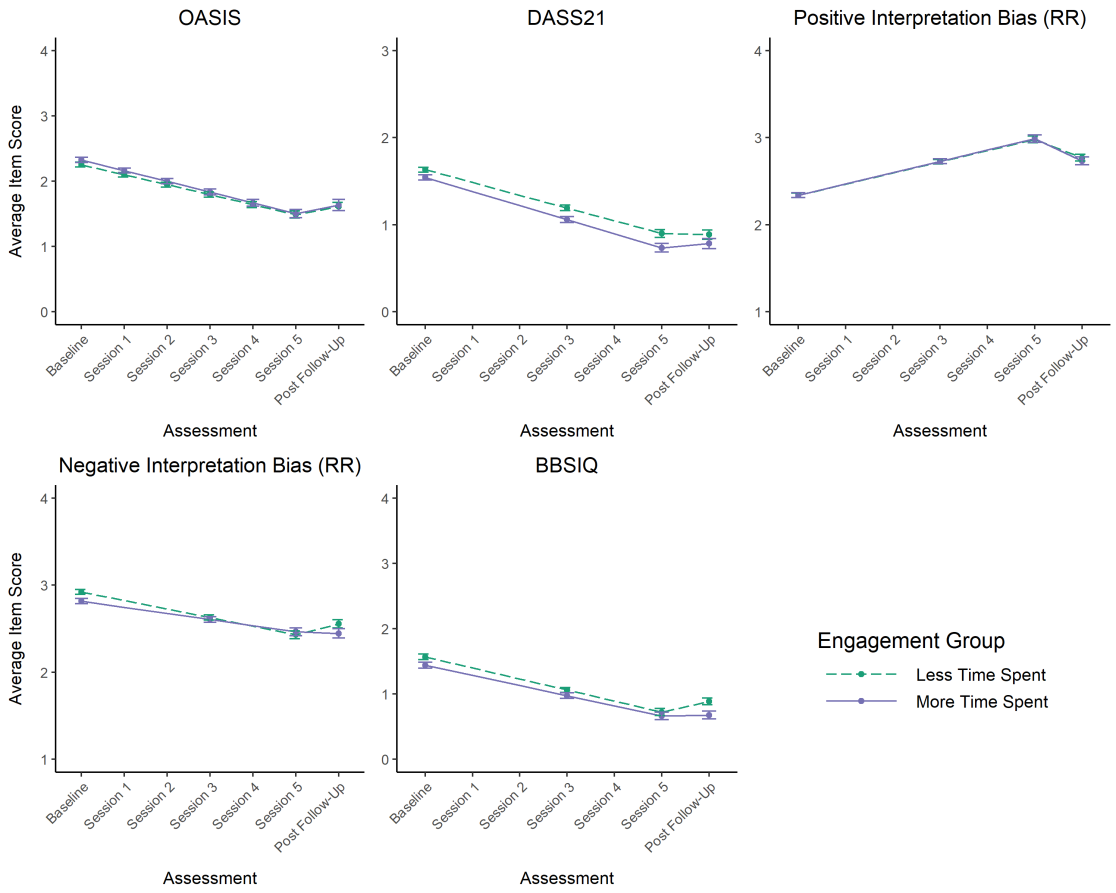
Positive Interpretation Bias

For the RR positive interpretation bias outcome, there were no significant interactions between the different linear time trajectories and engagement groups.

Negative Interpretation Bias

For the BBSIQ outcome, there was no significant interaction between time at the training trajectory and engagement groups. However, from the BBSIQ plot in Figure 4.1, we can see that both engagement groups reduced their negative interpretation bias scores to a comparable degree from Baseline to Session 5. At the Post Follow-Up trajectory we did find a significant interaction between time and engagement groups ($b = -0.15$, $p = .033$). From the simple time effects analysis, the group that spent less time showed

FIGURE 4.1: Piecewise Linear Estimated Means Over Time by Engagement Group for Analyzed Sample



Note. Estimated means ($\pm 1SE$) from the piecewise linear multilevel models at mean level of credibility online based on the analyzed sample. The reference group is Less Time Spent. Followed same procedure as Eberle, Boukhechba, et al. 2020. For each imputed dataset, estimated means were calculated and pooled using the `testConstraints` function of the `mitml` package (ver 0.4-3; Grund, Robitzsch, et al. 2021) that follows Rubin's rules. Plots were created with the `ggplot2` (ver. 3.3.5;Wickham et al. 2022) and `cowplot` (ver. 1.1.1; Wilke 2020) packages. Estimates are only shown for assessed timepoints.

a significant increase in negative interpretation bias ($b = 0.16$, $p = .001$), reflecting some loss in treatment gains. In contrast, the group that spent more time showed no significant change.

For the RR negative interpretation bias outcome, there were significant interactions between time at the training trajectory and engagement groups ($b = 0.03$, $p = .040$) and time at the Post Follow-Up trajectory and engagement groups ($b = -0.15$, $p = .028$). From the simple time effects analysis, both engagement groups showed a significant reduction in negative interpretation bias from Baseline to Session 5, with the group that spent less time showing a significantly more negative slope ($b = -0.10$ vs. $b = -0.07$, $ps =< .001$). Nonetheless, similar to the findings for the BBSIQ outcome, during the Post Follow-Up trajectory, the group that spent less time showed a significant increase in negative interpretation bias ($b = 0.13$, $p = .003$), reflecting some loss in treatment gains. In contrast, the group that spent more time showed no significant change.

TABLE 4.2: Piecewise Linear Multilevel Modeling Results for Individual Outcomes

									r		
Outcome	*Fixed Effect*	*b (SE)*	*t*	*df*	*p*	*95% CI*	*Random Effect*	s^2	1	2	3
OASIS	Intercept	2.25 (0.04)	62.49	3492.21	<.001***	[2.18, 2.32]	1. Intercept	0.38	-		
	$time_{TR}$	-0.15 (0.01)	-14.07	320.5	<.001***	[-0.17, -0.13]	2. $time_{TR}$	0.01	-0.18	-	
	$time_{FU}$	0.12 (0.06)	2.01	250.32	.045*	[0.00, 0.25]	3. $time_{FU}$	0.34	-0.14	-0.40	-
	More Time Spent	0.07 (0.05)	1.38	641.69	.169	[-0.03, 0.18]	Residual	0.18			
	Cred. Online GMC	0.01 (0.03)	0.27	535.14	.789	[-0.05, 0.07]					
	More Time Spent× $time_{TR}$	-0.01 (0.02)	-0.68	306.56	.494	[-0.04, 0.02]					
	More Time Spent×$time_{FU}$	0.00 (0.11)	0.03	180.43	.975	[-0.21, 0.21]					
DASS21	Intercept	1.63 (0.03)	59.82	2022.03	<.001***	[1.58, 1.69]	1. Intercept	0.17	-		
	$time_{TR}$	-0.15 (0.01)	-16.29	232.8	<.001***	[-0.16, -0.13]	2. $time_{TR}$	0.00	0.00	-	
	$time_{FU}$	-0.01 (0.05)	-0.3	157.29	.764	[-0.11, 0.08]	3. $time_{FU}$	0.04	-0.46	0.11	-
	More Time Spent	-0.09 (0.04)	-2.15	673.94	.032*	[-0.17, -0.01]	Residual	0.12			
	Cred. Online GMC	0.02 (0.02)	0.76	503.96	.449	[-0.03, 0.07]					
	More Time Spent×$time_{TR}$	-0.02 (0.01)	-1.19	249.28	.234	[-0.04, 0.01]					
	More Time Spent×$time_{FU}$	0.06 (0.06)	1	190.09	.320	[-0.06, 0.19]					
BBSIQ	Intercept	1.57 (0.04)	37.09	2033.33	<.001***	[1.49, 1.65]	1. Intercept	0.52	-		
	$time_{TR}$	-0.17 (0.01)	-13.46	353.23	<.001***	[-0.19, -0.14]	2. $time_{TR}$	0.02	-0.63	-	
	$time_{FU}$	0.16 (0.05)	3.4	197.54	.001**	[0.07, 0.26]	3. $time_{FU}$	0.03	0.03	-0.71	-
	More Time Spent	-0.13 (0.06)	-2.03	677.37	.043	[-0.25, -0.00]	Residual	0.18			
	Cred. Online GMC	0.04 (0.03)	1.11	354.69	.270	[-0.03, 0.10]					
	More Time Spent×$time_{TR}$	0.01 (0.02)	0.74	336.31	.463	[-0.02, 0.05]					
	More Time Spent×$time_{FU}$	-0.15 (0.07)	-2.14	213.62	.033*	[-0.29, -0.01]					
RR	Intercept	2.92 (0.03)	110.2	1969.37	<.001***	[2.87, 2.97]	1. Intercept	0.14	-		
Negative Bias	$time_{TR}$	-0.10 (0.01)	-10.3	275.16	<.001***	[-0.12, -0.08]	2. $time_{TR}$	0.01	-0.25	-	
	$time_{FU}$	0.13 (0.04)	3.09	201.3	.002**	[0.05, 0.21]	3. $time_{FU}$	0.01	-0.20	-0.38	-
	More Time Spent	-0.11 (0.04)	-2.67	664.33	.008**	[-0.18, -0.03]	Residual	0.14			
	Cred. Online GMC	0.03 (0.02)	1.15	457.54	.250	[-0.02, 0.07]					
	More Time Spent×$time_{TR}$	0.03 (0.01)	2.06	327.22	.040*	[0.00, 0.05]					
	More Time Spent×$time_{FU}$	-0.15 (0.07)	-2.22	173.46	.028*	[-0.28, -0.02]					

RR	Intercept	2.34 (0.03)	89.37	1999.7	<.001***	[2.29, 2.39]	1. Intercept	0.14	-	
Positive Bias	$time_{TR}$	0.13 (0.01)	14.14	366.96	<.001***	[0.11, 0.15]	2. $time_{TR}$	0.01	-0.57	-
	$time_{FU}$	-0.21 (0.04)	-5.21	217.01	<.001***	[-0.29, -0.13]	3. $time_{FU}$	0.03	0.44	-0.85 -
	More Time Spent	0.00 (0.04)	0.06	672.66	.952	[-0.07, 0.08]	Residual	0.13		
	Cred. Online GMC	0.00 (0.02)	0.07	374.71	.946	[-0.04, 0.04]				
	More Time Spent$\times time_{TR}$	0.00 (0.01)	0.07	383.21	.940	[-0.03, 0.03]				
	More Time Spent$\times time_{FU}$	-0.04 (0.07)	-0.67	185.97	.506	[-0.17, 0.09]				

Note. Each outcome was modeled separately. Every model had the fixed effects of engagement group, $time_{TR}$, $time_{FU}$, engagement group$\times time_{TR}$, engagement group$\times$ $time_{FU}$, credibility online grand mean centered, a random intercept, and random slopes for $time_{TR}$ and $time_{FU}$. Engagement group was dummy coded with Less Time Spent as the reference group (0 = Less Time Spent, 1 = More Time Spent).TR = Training trajectory; FU = Follow-Up trajectory; GMC = Grand mean centered. *p < .05. **p < .01. ***p < .001.

TABLE 4.3: Piecewise Linear Multilevel Modeling Engagement Group × Time Significant Interaction and Simple Time Effects for Analyzed Sample

Outcome	*Fixed Effect*	*b (SE)*	*t*	*df*	*p*	*95% CI*
BBSIQ	More Time Spent $\times$ time$_{FU}$	-0.15 (0.07)	-2.14	213.62	.033*	[-0.29, -0.01]
	Time$_{FU(LessTimeSpent)}$	0.16 (0.05)	3.38	168.46	.001**	[0.07, 0.26]
	Time$_{FU(MoreTimeSpent)}$	0.01 (0.05)	0.27	161.03	.788	[-0.09, 0.11]
RR Negative Bias	More Time Spent$\times$time$_{TR}$	0.03 (0.01)	2.06	327.22	.040*	[0.00, 0.05]
	Time$_{TR(LessTimeSpent)}$	-0.10 (0.01)	-10.05	243.67	<.001***	[-0.12, -0.08]
	Time$_{TR(MoreTimeSpent)}$	-0.07 (0.01)	-7.44	245.17	<.001***	[-0.09, -0.05]
	More Time Spent$\times$time$_{FU}$	-0.15 (0.07)	-2.22	173.46	.028*	[-0.28, -0.02]
	Time$_{FU(LessTimeSpent)}$	0.13 (0.04)	3.06	175.75	.003**	[0.05, 0.21]
	Time$_{FU(MoreTimeSpent)}$	-0.02 (0.05)	-0.38	125.7	.705	[-0.12, 0.08]

Note. For the simple time effects, separate models were fit for each engagement group with fixed effects for Time$_{TR}$, Time$_{FU}$, and credibility online grand mean centered, a random intercept, and random slopes for Time$_{TR}$ and Time$_{FU}$. Simple time effects were calculated and displayed only for significant interactions ($p < 0.05$). TR = Training trajectory; FU = Follow-Up trajectory. *$p < .05$. **$p < .01$. ***$p < .001$.

5 Discussion

5.1 Principal Findings and Implications

This study aimed to understand the relationship between engagement markers and the psychosocial outcomes of anxiety and interpretation bias in participants of the MindTrails Calm Thinking study who received CBM-I training. We defined engagement markers based on completion rate, time spent on training components, and time spent on assessment measures. Using K-means clustering, we identified two engagement patterns characterized by the amount of time spent in training and assessment measure components. One group spent more time in the measured intervention components than the other. Results showed no significant between-group differences in anxiety and positive interpretation bias outcomes; however, symptom reduction for negative interpretation bias significantly differed for both engagement groups during the training and post follow-up phases. During training, participants who spent less time had a more pronounced decrease in negative interpretation bias (RR) than participants who spent more time. As shown in Figure 4.1, the estimated mean scores in negative interpretation bias (RR) for participants that spent less time slightly drops below that of the other group by Session 5. This finding is contrary to the idea that more usage relates to better outcomes (Sieverink et al. 2017) and points towards the construct of effective engagement where a sufficient level of engagement exists such that participants benefit

from the intervention (Yardley et al. 2016). In both RR and BBSIQ outcomes, participants who spent less time had a significant loss in treatment gains during post follow-up, while participants who spent more time had no significant change. Our findings suggest that despite differences in engagement behavior, MindTrails participants still benefited from the intervention, as seen by a reduction in negative interpretation bias for both engagement groups. These findings share some similarities with other studies that have examined engagement patterns and intervention outcomes. For example, Matthews et al. 2018, Sanatkar et al. 2019, Li et al. 2022 identified four, three, and two distinct engagement patterns respectively that described levels of intervention completion and usage in different DMHIs for anxiety and depression. Matthews et al. 2018 reported that improvements in anxiety levels were noticeable across the four patterns of engagement at the early stages of the intervention. Sanatkar et al. 2019 highlighted that there were no significant between-group differences in symptom reduction for depression, anxiety, and stress among the three engagement patterns over time. Li et al. 2022 showed that although both low and high engagement groups reduced their symptoms of depression and stress, the high engagement group had a more pronounced reduction throughout the study. The low engagement group had some loss of gains in depressive symptoms at post follow-up.

These studies used engagement markers that described levels of completion rate and frequency of use, which makes the distinction between low and high engagement clearer to establish and support (e.g., participants who completed more of the intervention were more engaged). We hypothesized that higher levels of engagement with the MindTrails intervention would result in lower anxiety levels and greater improvements in interpretation bias. However, our clustering analysis returned two groups that differed in time spent on the intervention, making it difficult to assume that less or more time spent are indicators of lower or higher levels of engagement. Across the

literature, there have been mixed findings on whether time spent in DMHIs is associated with higher engagement and better outcomes. For example, results from Eberle, Meyer, et al. 2018 suggested that there exists a bidirectional relationship between participants who spent more time on individual CBM-I training scenarios and better outcomes. Donkin, Hickie, et al. 2013 pointed out that time spent on the intervention significantly differed between participants who did and did not obtain a significant change in depression symptoms. Nonetheless, when assessing whether usage was related to changes in outcomes, time spent on the intervention was not significant. Other studies have also shown that time spent is not associated with outcomes of depression and anxiety (Kenardy et al. 2003; Donkin, Christensen, et al. 2011; Zeng et al. 2020). Going back to Nahum-Shani et al. 2022's definition of engagement as a state of energy investment, one could argue that dedicating more time to a DMHI could suggest a greater investment of energy leading to a higher level of engagement, but time itself does not explicitly account for the cognitive and affective aspects included in this definition. Furthermore, there are multiple characteristics at the individual level, like processing speed and reading comprehension skills, that could influence the amount of time spent. Although time spent may help differentiate between levels of intention and interest in a DMHI (scanning through the material vs. taking time to view the content) (Pham et al. 2019), it is challenging to distinguish periods of inactivity from actual time dedicated to completing a digital intervention (Donkin, Christensen, et al. 2011). Taking this into account, we interpret our results based on differences in engagement patterns rather than engagement levels - we are not able to directly answer our hypothesis. The relationship between more time spent and higher levels of engagement is still unclear and studies continue to show that this engagement metric is not related to psychosocial outcomes. By itself, this feature may not be a strong enough indicator of this relationship, which could explain why we did not report any between-group differences in anxiety and positive interpretation bias. Nonetheless, considering that there were significant

between-group differences in negative interpretation bias, we need to further investigate the importance of this feature.

The dose-response effect of engagement on outcomes will be influenced by the markers researchers select for their analysis. These markers will vary depending on the design, structure, and goal of the DMHI (Hanano et al. 2022). Our engagement markers focused on completion rate, time spent on training components, and time spent on assessment measures, with most of the features belonging to the last marker. Although important to the intervention, assessment measures are indirectly related to the amount of training or dose that a participant receives, suggesting some limitations in using these types of features to understand dose-response effects. Instead, completion rate and time spent on training components are more direct indicators of enacted dose. Had we focused on these two markers for the clustering, most likely, we would have ended with a group of participants with low and high completion rates (similar to Li et al. 2022 and suggested by the bimodal distribution of completion rate in Figure S6). However, most of the participants in the low completion rate group only have baseline assessment data available since they did not complete the first training session. As we saw in the early stages of the analysis, this grouping raised issues for imputation convergence due to the high proportion of missing data at the scale level for all group members. This highlights the methodological challenges present in studies where assessment measures come after training and the importance of clearly defining the inclusion threshold for these types of analyses. Nevertheless, our findings and research contribute to the exploration of different engagement markers and help guide future work in understanding the association between engagement and outcomes.

5.2 Limitations

There are several limitations present in this research study. First, no causal claims about the relationship between engagement patterns and target outcomes can be made, even though the analyzed sample originally came from a randomized control trial. Second, we are potentially introducing sampling bias by excluding high-risk CBM-I participants assigned to coaching since the analyzed sample now had a greater proportion of low-risk CBM-I participants. Third, the generalizability of our results are impacted by the fact that the majority of the participants in the analyzed sample were female, White, Not Hispanic or Latino, and from the United States. Moreover, the replicability of our results is affected by our clustering methodology. As we saw from our clustering analysis, the data partitions varied between clustering techniques and number of clusters for the same analyzed sample. Choosing a different clustering method will likely yield different engagement groupings and conclusions. Further, clustering will return the number of clusters specified regardless of whether a pattern is present or not (Jain 2010), so it is important to analyze differences between clustered groups to check for patterns. Fourth, there may be other confounding variables that are not accounted for in our multilevel model, such as participant variability in enrollment motivation and self-regulation skills, which could influence engagement behavior and outcomes (Yardley et al. 2016). In addition, as with other DMHIs, the statistical significance of our conclusions is limited by the large proportion of missing data at post follow-up, which impact we tried to minimize by conducting fully Bayesian model-based imputations (Enders, Du, et al. 2020). Finally, the average time spent on training and measure components and our clusters (Less Time Spent vs. More Time Spent) do not capture engagement variability across the course of the intervention. For example, it could be the case that a participant who on average spent less time in the intervention, took longer to complete the last training sessions, which could be indicative of a different engagement behavior.

By using time spent on individual training sessions, we could gain a better insight into the dynamic nature of engagement and its fluctuations through time (Nahum-Shani et al. 2022).

5.3 Future Work

Understanding engagement behavior in DMHIs like MindTrails will continue to be important. One direction for future work is redefining the inclusion threshold and re-running the primary analysis on this new subset of participants. Although we initially restricted the analyzed sample to participants in CBM-I conditions who had started training, around 10% of the sample did not get to complete an individual training scenario. Hence, they did not receive an actual dose of the intervention, and their average time spent on most of the training component features was zero, which skewed the distribution of these features to the left. By adjusting the threshold to completing at least one training scenario or one training session, the analysis would focus on participants with data for those engagement features and who received some dose of CBM-I training. Hanano et al. 2022 conducted a similar analysis in a DMHI for anxiety and depression in university students. They first focused on the intent-to-treat sample and then on an initiators-only sample. Interestingly, the significant relationships between their engagement metrics and outcomes found in the intent-to-treat sample were no longer present in the initiators-only sample. These findings potentially suggests that analysis inclusion thresholds matter in understanding the association between engagement and outcomes and should be clearly reported. Furthermore, to test the replicability of our results, this research could expand to other MindTrails studies similar in structure, such as the Testing Engagement and Transfer (TET) study launched after Calm Thinking.

In addition to adjusting the intervention sample and analyzing a different participant pool, another direction would be to look at individual engagement metrics instead

of engagement groups. For example, Zeng et al. 2020 built latent growth curve models for each engagement metric of interest (completion rate, frequency of items completed, and time spent) and found that time spent was not significant. Li et al. 2022 built upon the work of Zeng et al. 2020 and excluded time spent from their clustering analysis. Similarly, we could first focus on identifying which engagement metrics are significantly related to outcomes, and then use them to cluster participants. This approach could provide greater insight into what specific measures are related to better results in DMHIs and help in the feature selection process for the clustering analysis.

We could also explore using other engagement features. For example, in their attrition study Baee et al. 2022 looked at behavioral features like time spent during training, time passed between sessions, and time of the day when intervention items were completed. Furthermore, it will be important to include features related to the cognitive and affective components of engagement. For example, in the Calm Thinking study participants completed the subjective units of distress questionnaire before and after the first, third, and fifth training sessions. This questionnaire assesses a participant's current level of anxiety, which relates to the affective component of engagement. These features could be included in our engagement analysis to develop a more complete understanding of how participants interact with MindTrails.

Finally, our analysis does not account for the magnitude of the between-group differences or the variability in engagement behavior (e.g., initial engagement followed by disengagement). To address the first, we need to calculate the standardized effect sizes of these differences in order to increase the interpretability of our results and help other researchers compare their findings to ours (Lorah 2018). In order to analyze variability in engagement behavior across time, we could explore implementing a growth mixture model. In the literature these types of models have been used to identify unobserved sub-groups (e.g., engagement groups) and examine the within-group and between-group differences of said groups (Ram and Grimm 2009; Coa et al. 2018).

These patterns of sustained engagement may be more indicative of participants' true interaction with DMHIs and its relationship with outcomes.

6 Conclusion

This study provided an insight into the complex and dynamic nature of engagement and its relationship to psychosocial outcomes. In conclusion, we identified two engagement patterns characteristic of time spent on intervention components. There were no significant between-group differences for anxiety and positive interpretation bias, consistent with the majority of previous studies suggesting that time spent is not significantly associated with outcomes like anxiety. We found significant between-group differences in negative interpretation bias; nevertheless, both groups seem to have significantly improved. These findings further recognize that participants interact differently with DMHIs and still report symptom improvements. The relationship between engagement and outcomes should be further investigated to determine whether specific behaviors are associated with better results. This relationship can then be accounted for in the design of future DMHIs to improve their effectiveness and benefit users.

Bibliography

Borkovec, Thomas D. and Nau, Sidney D. (Dec. 1972). "Credibility of analogue therapy rationales". en. In: *Journal of Behavior Therapy and Experimental Psychiatry* 3.4, pp. 257–260. ISSN: 0005-7916. DOI: 10.1016/0005-7916(72)90045-6.

Lovibond, P.F. and Lovibond, S.H. (Mar. 1995). "The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories". en. In: *Behaviour Research and Therapy* 33.3, pp. 335–343. ISSN: 00057967. DOI: 10.1016/0005-7967(94)00075-U.

Clark, David M., Salkovskis, Paul M., Öst, Lars-Göran, Breitholtz, Elisabeth, Koehler, Katherine A., Westling, Bengt E., Jeavons, Ann, and Gelder, Michael (1997). "Misinterpretation of body sensations in panic disorder". In: *Journal of Consulting and Clinical Psychology* 65.2. Place: US Publisher: American Psychological Association, pp. 203–213. ISSN: 1939-2117. DOI: 10.1037/0022-006X.65.2.203.

Guha, Sudipto, Rastogi, Rajeev, and Shim, Kyuseok (June 1998). "CURE: an efficient clustering algorithm for large databases". In: *ACM SIGMOD Record* 27.2, pp. 73–84. ISSN: 0163-5808. DOI: 10.1145/276305.276312.

Mathews, Andrew and Mackintosh, Bundy (2000). "Induced emotional interpretation bias and anxiety". In: *Journal of Abnormal Psychology* 109.4. Place: US Publisher: American Psychological Association, pp. 602–615. ISSN: 1939-1846. DOI: 10.1037/0021-843X.109.4.602.

Kenardy, Justin, McCafferty, Kelly, and Rosa, Viginia (July 2003). "Internet-Delivered Indicated Prevention For Anxiety Disorders: A Randomized Controlled Trial". en.

In: *Behavioural and Cognitive Psychotherapy* 31.3, pp. 279–289. ISSN: 13524658. DOI: 10.1017/S1352465803003047.

Eysenbach, Gunther (Mar. 2005). "The Law of Attrition". en. In: *Journal of Medical Internet Research* 7.1, e11. ISSN: 1438-8871. DOI: 10.2196/jmir.7.1.e11.

Christensen, Helen and Mackinnon, Andrew (Sept. 2006). "The Law of Attrition Revisited". EN. In: *Journal of Medical Internet Research* 8.3. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada, e558. DOI: 10.2196/jmir.8.3.e20.

Norman, Sonya B., Hami Cissell, Shadha, Means-Christensen, Adrienne J., and Stein, Murray B. (2006). "Development and validation of an Overall Anxiety Severity And Impairment Scale (OASIS)". en. In: *Depression and Anxiety* 23.4, pp. 245–249. ISSN: 1520-6394. DOI: 10.1002/da.20182.

Brock, Guy, Pihur, Vasyl, Datta, Susmita, and Datta, Somnath (2008). "clValid: An R Package for Cluster Validation". In: *Journal of Statistical Software* 25.4, pp. 1–22. DOI: 10.18637/jss.v025.i04.

Ram, Nilam and Grimm, Kevin J. (2009). "Growth Mixture Modeling: A Method for Identifying Differences in Longitudinal Change Among Unobserved Groups". In: *International journal of behavioral development* 33.6, pp. 565–576. ISSN: 0165-0254. DOI: 10.1177/0165025409343765.

Tasca, Giorgio A. and Gallop, Robert (July 2009). "Multilevel modeling of longitudinal data for psychotherapy researchers: I. The basics". en. In: *Psychotherapy Research* 19.4-5, pp. 429–437. ISSN: 1050-3307, 1468-4381. DOI: 10.1080/10503300802641444.

Enders, Craig K. (2010). en. In: *Applied missing data analysis*. Methodology in the social sciences. OCLC: ocn456171131. New York: Guilford Press, p. 265. ISBN: 978-1-60623-639-0.

Jain, Anil K. (June 2010). "Data clustering: 50 years beyond K-means". en. In: *Pattern Recognition Letters*. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR) 31.8, pp. 651–666. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2009.09.011.

Steinman, Shari A. and Teachman, Bethany A. (Jan. 2010). "Modifying interpretations among individuals high in anxiety sensitivity". en. In: *Journal of Anxiety Disorders* 24.1, pp. 71–78. ISSN: 0887-6185. DOI: 10.1016/j.janxdis.2009.08.008.

Beard, Courtney (Feb. 2011). "Cognitive bias modification for anxiety: current evidence and future directions". In: *Expert review of neurotherapeutics* 11.2, pp. 299–311. ISSN: 1473-7175. DOI: 10.1586/ern.10.194.

Donkin, Liesje, Christensen, Helen, Naismith, Sharon L., Neal, Bruce, Hickie, Ian B., and Glozier, Nick (Aug. 2011). "A Systematic Review of the Impact of Adherence on the Effectiveness of e-Therapies". EN. In: *Journal of Medical Internet Research* 13.3. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada, e1772. DOI: 10.2196/jmir.1772.

Madhulatha, Tagaram Soni (2011). "Comparison between K-Means and K-Medoids Clustering Algorithms". In: *Advances in Computing and Information Technology*. Ed. by David C. Wyld, Michal Wozniak, Nabendu Chaki, Natarajan Meghanathan, and Dhinaharan Nagamalai. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 472–481. ISBN: 978-3-642-22555-0.

MacLeod, Colin and Mathews, Andrew (2012). "Cognitive Bias Modification Approaches to Anxiety". In: *Annual Review of Clinical Psychology* 8.1. _eprint: https://doi.org/10.1146/annurev-clinpsy-032511-143052, pp. 189–217. DOI: 10.1146/annurev-clinpsy-032511-143052.

Mathews, Andrew (Feb. 2012). "Effects of modifying the interpretation of emotional ambiguity". en. In: *Journal of Cognitive Psychology* 24.1, pp. 92–105. ISSN: 2044-5911, 2044-592X. DOI: 10.1080/20445911.2011.584527.

Donkin, Liesje, Hickie, Ian B, Christensen, Helen, Naismith, Sharon L, Neal, Bruce, Cockayne, Nicole L, and Glozier, Nick (Oct. 2013). "Rethinking the Dose-Response Relationship Between Usage and Outcome in an Online Intervention for Depression: Randomized Controlled Trial". In: *Journal of Medical Internet Research* 15.10, e231. ISSN: 1439-4456. DOI: 10.2196/jmir.2771.

James, Gareth, Witten, Daniela, Hastie, Trevor, and Tibshirani, Robert (2013). "Chapter 10: Unsupervised Learning". en. In: *An Introduction to Statistical Learning with Applications in R*. 1st ed. Springer Texts in Statistics. Springer New York, NY. ISBN: 978-1-4614-7138-7.

Leys, Christophe, Ley, Christophe, Klein, Olivier, Bernard, Philippe, and Licata, Laurent (July 2013). "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median". en. In: *Journal of Experimental Social Psychology* 49.4, pp. 764–766. ISSN: 0022-1031. DOI: 10.1016/j.jesp.2013.03.013.

Andrade, L. H. et al. (Apr. 2014). "Barriers to mental health treatment: results from the WHO World Mental Health surveys". en. In: *Psychological Medicine* 44.6, pp. 1303–1317. ISSN: 0033-2917, 1469-8978. DOI: 10.1017/S0033291713001943.

Charrad, Malika, Ghazzali, Nadia, Boiteau, Véronique, and Niknafs, Azam (2014). "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set". In: *Journal of Statistical Software* 61.6, pp. 1–36.

Steinman, Shari A. and Teachman, Bethany A. (Dec. 2015). "Training less threatening interpretations over the Internet: Does the number of missing letters matter?" en. In: *Journal of Behavior Therapy and Experimental Psychiatry*. Cognitive bias modification: Challenges and new directions 49, pp. 53–60. ISSN: 0005-7916. DOI: 10.1016/j.jbtep.2014.12.004.

Murray, Elizabeth, Hekler, Eric B., Andersson, Gerhard, Collins, Linda M., Doherty, Aiden, Hollis, Chris, Rivera, Daniel E., West, Robert, and Wyatt, Jeremy C. (Nov. 2016). "Evaluating Digital Health Interventions: Key Questions and Approaches". en. In: *American Journal of Preventive Medicine* 51.5, pp. 843–851. ISSN: 0749-3797. DOI: 10.1016/j.amepre.2016.06.008.

Yardley, Lucy, Spring, Bonnie, Riper, Heleen, Morrison, Leanne, Crane, David, Curtis, Kristina, Merchant, Gina, Naughton, Felix, and Blandford, Ann (Nov. 2016). "Understanding and Promoting Effective Engagement With Digital Behavior Change Interventions". In: *American Journal of Preventive Medicine* 51.5. MAG ID: 2474187140, pp. 833–842. DOI: 10.1016/j.amepre.2016.06.015.

Michie, Susan, Yardley, Lucy, West, Robert, Patrick, Kevin, and Greaves, Felix (June 2017). "Developing and Evaluating Digital Interventions to Promote Behavior Change in Health and Health Care: Recommendations Resulting From an International Workshop". EN. In: *Journal of Medical Internet Research* 19.6, e7126. DOI: 10.2196/jmir.7126.

Perski, Olga, Blandford, Ann, West, Robert, and Michie, Susan (June 2017). "Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis". In: *Translational Behavioral Medicine* 7.2, pp. 254–267. ISSN: 1869-6716. DOI: 10.1007/s13142-016-0453-1.

Sieverink, Floor, Kelders, Saskia M., and Gemert-Pijnen, Julia Ewc van (Dec. 2017). "Clarifying the Concept of Adherence to eHealth Technology: Systematic Review on When Usage Becomes Adherence". eng. In: *Journal of Medical Internet Research* 19.12, e402. ISSN: 1438-8871. DOI: 10.2196/jmir.8578.

Coa, Kisha I, Wiseman, Kara P, Higgins, Bryan, and Augustson, Erik (Apr. 2018). "Associations Between Engagement and Outcomes in the SmokefreeTXT Program: A Growth Mixture Modeling Analysis". In: *Nicotine & Tobacco Research* 21.5, pp. 663–669. ISSN: 1462-2203. DOI: 10.1093/ntr/nty073.

Eberle, Jeremy W., Meyer, Joseph, and Teachman, Bethany A. (2018). *Trial Time; Engagement; and Change in Expectancy Bias, Distress, and Well-Being in Online Cognitive Bias*.

Geramita, Emily M., Herbeck Belnap, Bea, Abebe, Kaleab Z., Rothenberger, Scott D., Rotondi, Armando J., and Rollman, Bruce L. (July 2018). "The Association Between Increased Levels of Patient Engagement With an Internet Support Group and Improved Mental Health Outcomes at 6-Month Follow-Up: Post-Hoc Analyses From a Randomized Controlled Trial". eng. In: *Journal of Medical Internet Research* 20.7, e10402. ISSN: 1438-8871. DOI: 10.2196/10402.

Grund, Simon, Lüdtke, Oliver, and Robitzsch, Alexander (Jan. 2018). "Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations". en. In: *Organizational Research Methods* 21.1, pp. 111–149. ISSN: 1094-4281, 1552-7425. DOI: 10.1177/1094428117703686.

Lorah, Julie (Dec. 2018). "Effect size measures for multilevel models: definition, interpretation, and TIMSS example". en. In: *Large-scale Assessments in Education* 6.1, p. 8. ISSN: 2196-0739. DOI: 10.1186/s40536-018-0061-2.

Matthews, Paul, Topham, Phil, and Caleb-Solly, Praminda (Oct. 2018). "Interaction and Engagement with an Anxiety Management App: Analysis Using Large-Scale Behavioral Data". EN. In: *JMIR Mental Health* 5.4, e9235. DOI: 10.2196/mental.9235.

Enrique, Angel, Palacios, Jorge E, Ryan, Holly, and Richards, Derek (Aug. 2019). "Exploring the Relationship Between Usage and Outcomes of an Internet-Based Intervention for Individuals With Depressive Symptoms: Secondary Analysis of Data From a Randomized Controlled Trial". en. In: *Journal of Medical Internet Research* 21.8, e12775. ISSN: 1438-8871. DOI: 10.2196/12775.

Hayes, Timothy (Oct. 2019). "Flexible, Free Software for Multilevel Multiple Imputation: A Review of Blimp and jomo". In: *Journal of Educational and Behavioral Statistics* 44.5. Publisher: American Educational Research Association, pp. 625–641. ISSN: 1076-9986. DOI: 10.3102/1076998619858624.

McVay, Megan A., Bennett, Gary G., Steinberg, Dori, and Voils, Corrine I. (Dec. 2019). "Dose-Response Research in Digital Health Interventions: Concepts, Considerations, and Challenges". In: *Health psychology : official journal of the Division of Health Psychology, American Psychological Association* 38.12, pp. 1168–1174. ISSN: 0278-6133. DOI: 10.1037/hea0000805.

Pham, Quynh, Graham, Gary, Carme Carrion, Carrion, Carme, Morita, Plinio P, Seto, Emily, Stinson, Jennifer, and Cafazzo, Joseph A. (Jan. 2019). "A Library of Analytic Indicators to Evaluate Effective Engagement with Consumer mHealth Apps for Chronic Conditions: Scoping Review". In: *Jmir mhealth and uhealth* 7.1. MAG ID: 2910244894. DOI: 10.2196/11941.

Sanatkar, Samineh, Baldwin, Peter Andrew, Huckvale, Kit, Clarke, Janine, Christensen, Helen, Harvey, Samuel, and Proudfoot, Judy (Nov. 2019). "Using Cluster Analysis to Explore Engagement and e-Attainment as Emergent Behavior in Electronic Mental Health". EN. In: *Journal of Medical Internet Research* 21.11, e14728. DOI: 10.2196/14728.

Schubert, Erich and Rousseeuw, Peter J. (2019). "Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms". In: *arXiv:1810.05691 [cs, stat]* 11807. arXiv: 1810.05691, pp. 171–187. DOI: 10.1007/978-3-030-32047-8_16.

Bethune, Sophie (Nov. 2020). *Psychologists report large increase in demand for anxiety, depression treatment*. en.

Chien, Isabel, Enrique, Angel, Palacios, Jorge, Regan, Tim, Keegan, Dessie, Carter, David, Tschiatschek, Sebastian, Nori, Aditya, Thieme, Anja, Richards, Derek, Doherty, Gavin, and Belgrave, Danielle (July 2020). "A Machine Learning Approach to Understanding Patterns of Engagement With Internet-Delivered Mental Health Interventions". In: *JAMA Network Open* 3.7, e2010791. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2020.10791.

Eberle, Jeremy W., Boukhechba, Mehdi, Sun, Jianhui, Zhang, Diheng, Funk, Dan, Barnes, Laura, and Teachman, Bethany (July 2020). "Shifting Episodic Prediction With Online Cognitive Bias Modification: A Randomized Controlled Trial". en-us. In: type: article. DOI: 10.31234/osf.io/dg7z8.

Enders, Craig K., Du, Han, and Keller, Brian T. (2020). "A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms". In: *Psychological Methods* 25.1. Place: US Publisher: American Psychological Association, pp. 88–112. ISSN: 1939-1463. DOI: 10.1037/met0000228.

Fu, Zhongfang, Burger, Huibert, Arjadi, Retha, and Bockting, Claudi L. H. (Oct. 2020). "Effectiveness of digital psychological interventions for mental health problems in low-income and middle-income countries: a systematic review and meta-analysis". English. In: *The Lancet Psychiatry* 7.10. Publisher: Elsevier, pp. 851–864. ISSN: 2215-0366, 2215-0374. DOI: 10.1016/S2215-0366(20)30256-X.

Hohensee, Nicola, Meyer, M. Joseph, and Teachman, Bethany A. (Sept. 2020). "The Effect of Confidence on Dropout Rate and Outcomes in Online Cognitive Bias Modification". en. In: *Journal of Technology in Behavioral Science* 5.3, pp. 226–234. ISSN: 2366-5963. DOI: 10.1007/s41347-020-00129-8.

Kassambara, Alboukadel and Mundt, Fabian (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*.

Linardon, Jake and Fuller-Tyszkiewicz, Matthew (Jan. 2020). "Attrition and adherence in smartphone-delivered interventions for mental health problems: A systematic and meta-analytic review". eng. In: *Journal of Consulting and Clinical Psychology* 88.1, pp. 1–13. ISSN: 1939-2117. DOI: 10.1037/ccp0000459.

Wilke, Claus O. (Dec. 2020). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*.

Zeng, Yu, Yan Guo, Guo, Yan, Guo, Yan, Li, Linghua, Li, Linghua, Hong, Y Alicia, Li, Yiran, Zhu, Mengting, Zeng, Chengbo, Zhang, Hanxi, Cai, Weiping, Cai, Weiping,

Liu, Cong, Wu, Shaomin, Chi, Peilian, Peilian Chi, Monroe-Wise, Aliza, Hao, Yuantao, and Ho, Rainbow T. H. (2020). "Relationship Between Patient Engagement and Depressive Symptoms Among People Living With HIV in a Mobile Health Intervention: Secondary Analysis of a Randomized Controlled Trial." In: *Jmir mhealth and uhealth* 8.10. MAG ID: 3088200138. DOI: 10.2196/20847.

Brown, Violet A. (Jan. 2021). "An Introduction to Linear Mixed-Effects Modeling in R". en. In: *Advances in Methods and Practices in Psychological Science* 4.1, p. 251524592096035. ISSN: 2515-2459, 2515-2467. DOI: 10.1177/2515245920960351.

Gan, Daniel Z. Q., McGillivray, Lauren, Han, Jin, Christensen, Helen, and Torok, Michelle (2021). "Effect of Engagement With Digital Interventions on Mental Health Outcomes: A Systematic Review and Meta-Analysis". In: *Frontiers in Digital Health* 3. ISSN: 2673-253X.

Grund, Simon, Robitzsch, Alexander, and Luedtke, Oliver (Oct. 2021). *mitml: Tools for Multiple Imputation in Multilevel Modeling*.

Ji, Julie L., Baee, Sonia, Zhang, Diheng, Calicho-Mamani, Claudia P., Meyer, M. Joseph, Funk, Daniel, Portnow, Samuel, Barnes, Laura, and Teachman, Bethany A. (July 2021). "Multi-session online interpretation bias training for anxiety in a community sample". en. In: *Behaviour Research and Therapy* 142, p. 103864. ISSN: 0005-7967. DOI: 10.1016/j.brat.2021.103864.

Keller, Brian T. and Enders, Craig K. (2021). *Blimp user's guide (Version 3)*. en.

Lehtimaki, Susanna, Martic, Jana, Wahl, Brian, Foster, Katherine T., and Schwalbe, Nina (Apr. 2021). "Evidence on Digital Mental Health Interventions for Adolescents and Young People: Systematic Overview". EN. In: *JMIR Mental Health* 8.4, e25847. DOI: 10.2196/25847.

Newby, Katie, Teah, Grace, Cooke, Richard, Li, Xinru, Brown, Katherine, Salisbury-Finch, Bradley, Kwah, Kayleigh, Bartle, Naomi, Curtis, Kristina, Fulton, Emmie,

Parsons, Joanne, Dusseldorp, Elise, and Williams, Stefanie L. (Mar. 2021). "Do automated digital health behaviour change interventions have a positive effect on self-efficacy? A systematic review and meta-analysis". eng. In: *Health Psychology Review* 15.1, pp. 140–158. ISSN: 1743-7202. DOI: 10.1080/17437199.2019.1705873.

Pinheiro, Jose, Bates, Douglas, DebRoy, Saikat, Sarkar, Deepayan, and R Core Team (2021). *nlme: Linear and Nonlinear Mixed Effects Models*.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Santomauro, Damian F et al. (Oct. 2021). "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic". en. In: *The Lancet*, S0140673621021437. ISSN: 01406736. DOI: 10.1016/S0140-6736(21)02143-7.

Wyatt, Kristin P., Eberle, Jeremy W., Ruork, Allison K., Enders, Craig K., and Neacsiu, Andrada D. (2021). "Mechanisms of change in treatments for transdiagnostic emotion dysregulation: The roles of skills use, perceived control, and mindfulness". en.

Baee, Sonia, Eberle, Jeremy W., Baglione, Anna N., Spears, Tyler, Lewis, Elijah, Behan, Henry C., Wang, Honging, Funk, Daniel H., Barnes, Laura E., and Teachman, Bethany A. (2022). "Early Attrition Prediction for Digital Mental Health Studies". In: *Manuscript is ready submit to Journal of Medical Internet Research*.

Eberle, Jeremy W., Baee, Sonia, Behan, Henry C., Baglione, Anna N., Boukhechba, Mehdi, Funk, Daniel H., Barnes, Laura E., and Teachman, Bethany A. (Feb. 2022). *TeachmanLab/MT-Data-CalmThinkingStudy: Centralized Data Cleaning for MindTrails Calm Thinking Study*. DOI: 10.5281/zenodo.6149366.

Eberle, Jeremy W., Daniel, Katharine E., Baee, Sonia, Behan, Henry C., Silverman, Alexandra L., Calicho-Mamani, Claudia, Baglione, Anna N., Werntz, Alexandra, French, Noah J., Ji, Julie L., Hohensee, Nicola, Tong, Xin, Boukhechba, Mehdi, Funk, Daniel

H., Barnes, Laura E., and Teachman, Bethany A. (2022). "Web-based interpretation training to reduce anxiety: A sequential multiple-assignment randomized trial". en.

Enders, Craig K., Keller, Brian T., Du, Han, and Levy, Roy (2022). *Blimp Studio*.

Hanano, Maria, Rith-Najarian, Leslie, Boyd, Meredith, and Chavira, Denise (Mar. 2022). "Measuring Adherence Within a Self-Guided Online Intervention for Depression and Anxiety: Secondary Analyses of a Randomized Controlled Trial". EN. In: *JMIR Mental Health* 9.3, e30754. DOI: 10.2196/30754.

Li, Yiran, Guo, Yan, Hong, Y. Alicia, Zeng, Yu, Monroe-Wise, Aliza, Zeng, Chengbo, Zhu, Mengting, Zhang, Hanxi, Qiao, Jiaying, Xu, Zhimeng, Cai, Weiping, Li, Linghua, and Liu, Cong (Jan. 2022). "Dose-Response Effects of Patient Engagement on Health Outcomes in an mHealth Intervention: Secondary Analysis of a Randomized Controlled Trial". eng. In: *JMIR mHealth and uHealth* 10.1, e25586. ISSN: 2291-5222. DOI: 10.2196/25586.

Wickham, Hadley, Chang, Winston, Henry, Lionel, Pedersen, Thomas Lin, Takahashi, Kohske, Wilke, Claus, Woo, Kara, Yutani, Hiroaki, Dunnington, Dewey, and RStudio (May 2022). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*.

Nahum-Shani, Inbal, Shaw, Steven D., Carpenter, Stephanie M., Murphy, Susan A., and Yoon, Carolyn (2022). "Engagement in digital interventions." In: *American Psychologist* (). Publisher: US: American Psychological Association. ISSN: 1935-990X. DOI: 10.1037/amp0000983.

7 Supplemental Tables and Figures

TABLE S1: Demographic Characteristics by Engagement Group for Analyzed Sample

Characteristic	Analyzed Sample (n = 697)	Less Time Spent (n = 386)	More Time Spent (n = 311)
Age (years): *M (SD)*	35.38 (11.88)	31.94 (9.6)	39.66 (13.01)
Gender: *n (%)*			
Female	564 (80.9)	316 (81.9)	248 (79.7)
Male	116 (16.6)	59 (15.3)	57 (18.3)
Transgender	2 (0.3)	1 (0.3)	1 (0.3)
Transgender Female	0 (0)	0 (0)	0 (0)
Transgender Male	3 (0.4)	2 (0.5)	1 (0.3)
Other	10 (1.4)	7 (1.8)	3 (1)
Prefer not to answer	2 (0.3)	1 (0.3)	1 (0.3)
Race: *n (%)*			
American Indian/Alaska Native	7 (1)	2 (0.5)	5 (1.6)
Black/African origin	69 (9.9)	33 (8.5)	36 (11.6)
East Asian	15 (2.2)	12 (3.1)	3 (1)
Native Hawaiian/Pacific Islander	4 (0.6)	1 (0.3)	3 (1)
South Asian	13 (1.9)	9 (2.3)	4 (1.3)
White/European origin	491 (70.4)	270 (69.9)	221 (71.1)
More than one race	58 (8.3)	41 (10.6)	17 (5.5)
Other or Unknown	32 (4.6)	13 (3.4)	19 (6.1)
Prefer not to answer	8 (1.1)	5 (1.3)	3 (1)
Ethnicity: *n (%)*			
Hispanic or Latino	93 (13.3)	48 (12.4)	45 (14.5)
Not Hispanic or Latino	570 (81.8)	325 (84.2)	245 (78.8)
Unknown	14 (2)	6 (1.6)	8 (2.6)
Prefer not to answer	20 (2.9)	7 (1.8)	13 (4.2)
Country : *n (%)*			
United States	639 (91.7)	367 (95.1)	272 (87.5)
Australia	28 (4)	8 (2.1)	20 (6.4)
Canada	8 (1.1)	3 (0.8)	5 (1.6)
United Kingdom	8 (1.1)	5 (1.3)	3 (1)
Other	13 [a] (1.9)	3 (0.8)	10 (3.2)
Prefer not to answer	1 (0.1)	0 (0)	1 (0.3)
Education: *n (%)*			
Junior High	1 (0.1)	1 (0.3)	0 (0)
Some High School	9 (1.3)	5 (1.3)	4 (1.3)
High School Graduate	59 (8.5)	30 (7.8)	29 (9.3)
Some College	225 (32.3)	112 (29)	113 (36.3)
Associate's Degree	75 (10.8)	45 (11.7)	30 (9.6)

[a]Germany (*n = 2*), India (*n = 2*), Ireland (*n = 2*), Colombia (*n = 1*), Croatia (*n = 1*), Ecuador (*n = 1*), Jordan (*n = 1*), Malaysia (*n = 1*), South Africa (*n = 1*), United Arab Emirates (*n = 1*)

Bachelor's Degree	164 (23.5)	90 (23.3)	74 (23.8)
Some Graduate School	38 (5.5)	27 (7)	11 (3.5)
Master's Degree	84 (12.1)	53 (13.7)	31 (10)
M.B.A.	13 (1.9)	10 (2.6)	3 (1)
J.D.	3 (0.4)	0 (0)	3 (1)
M.D.	2 (0.3)	1 (0.3)	1 (0.3)
Ph.D.	10 (1.4)	6 (1.6)	4 (1.3)
Other Advanced Degree	12 (1.7)	6 (1.6)	6 (1.9)
Prefer not to answer	2 (0.3)	0 (0)	2 (0.6)
Employment Status: *n (%)*			
Student	96 (13.8)	65 (16.8)	31 (10)
Homemaker	60 (8.6)	29 (7.5)	31 (10)
Unemployed or laid off	33 (4.7)	13 (3.4)	20 (6.4)
Looking for work	38 (5.5)	19 (4.9)	19 (6.1)
Working part-time	107 (15.4)	56 (14.5)	51 (16.4)
Working full-time	295 (42.3)	182 (47.2)	113 (36.3)
Retired	22 (3.2)	4 (1)	18 (5.8)
Other	42 (6)	16 (4.1)	26 (8.4)
Unknown	0 (0)	0 (0)	0 (0)
Prefer not to answer	4 (0.6)	2 (0.5)	2 (0.6)
Annual Income: *n (%)*			
Less than $5,000	33 (4.7)	13 (3.4)	20 (6.4)
$5,000 through $11,999	52 (7.5)	25 (6.5)	27 (8.7)
$12,000 through $15,999	27 (3.9)	15 (3.9)	12 (3.9)
$16,000 through $24,999	63 (9)	32 (8.3)	31 (10)
$25,000 through $34,999	67 (9.6)	36 (9.3)	31 (10)
$35,000 through $49,999	102 (14.6)	58 (15)	44 (14.1)
$50,000 through $74,999	111 (15.9)	68 (17.6)	43 (13.8)
$75,000 through $99,999	68 (9.8)	45 (11.7)	23 (7.4)
$100,000 through $149,999	67 (9.6)	42 (10.9)	25 (8)
$150,000 through $199,999	18 (2.6)	10 (2.6)	8 (2.6)
$200,000 through $249,999	11 (1.6)	5 (1.3)	6 (1.9)
$250,000 or greater	12 (1.7)	9 (2.3)	3 (1)
Unknown	28 (4)	16 (4.1)	12 (3.9)
Prefer not to answer	38 (5.5)	12 (3.1)	26 (8.4)
Marital Status: *n (%)*			
Single	187 (26.8)	103 (26.7)	84 (27)
Dating	83 (11.9)	48 (12.4)	35 (11.3)
Engaged	33 (4.7)	22 (5.7)	11 (3.5)
In a marriage-like relationship	88 (12.6)	51 (13.2)	37 (11.9)
Married	203 (29.1)	117 (30.3)	86 (27.7)
In a domestic/civil union	25 (3.6)	12 (3.1)	13 (4.2)

Separated	16 (2.3)	7 (1.8)	9 (2.9)
Divorced	45 (6.5)	21 (5.4)	24 (7.7)
Widow/widower	7 (1)	1 (0.3)	6 (1.9)
Other	6 (0.9)	3 (0.8)	3 (1)
Prefer not to answer	4 (0.6)	1 (0.3)	3 (1)

TABLE S2: Out of Range Scores Across the 100 Imputed Datasets for Analyzed Sample

				Minimum Score			Maximum Score	
Outome	Timepoint	Scale Range	M% Below	M	Absolute	M% Above	M	Absolute
OASIS	Baseline [a]	[0,4]	0.00	0.00	0.00	0.00	4.00	4.00
	Session 1		0.06	0.00	-0.82	0.11	4.16	5.00
	Session 2		0.16	-0.25	-0.98	0.19	4.27	5.02
	Session 3		0.42	-0.45	-1.43	0.16	4.21	4.98
	Session 4		1.06	-0.66	-1.53	0.15	4.18	4.89
	Session 5		2.47	-1.01	-1.91	0.16	4.16	5.08
	Post Follow-Up		1.81	-0.93	-2.32	0.21	4.28	5.30
DASS21	Baseline [b]	[0,3]	0.00	0.57	0.57	0.00	3.00	3.00
	Session 3		0.95	-0.46	-1.34	0.06	3.05	3.70
	Session 5		5.57	-0.97	-1.41	0.03	2.82	3.40
	Post Follow-Up		5.35	-0.96	-1.91	0.03	2.92	3.33
BBSIQ	Baseline [a]	[0,4]	0.00	0.00	0.00	0.00	4.00	4.00
	Session 3		3.59	-0.96	-1.78	0.00	4.00	4.00
	Session 5		11.10	-1.53	-2.42	0.00	3.93	3.99
	Post Follow-Up		7.64	-1.12	-1.97	0.00	3.12	4.06
RR Negative Bias	Baseline [a]	[1,4]	0.00	1.33	1.33	0.00	4.00	4.01
	Session 3		0.08	0.94	0.34	0.21	4.16	4.80
	Session 5		0.55	0.60	0.00	0.29	4.24	4.83
	Post Follow-Up		0.29	0.76	0.15	0.29	4.22	4.96
RR Positive Bias	Baseline [a]	[1,4]	0.00	1.00	1.00	0.00	4.00	4.00
	Session 3		0.01	0.99	0.85	0.18	4.14	4.51
	Session 5		0.01	1.25	0.57	1.82	4.59	5.54
	Post Follow-Up		0.01	1.26	0.75	0.41	4.25	4.87

[a] Assessed during pretreatment
[b] Assessed during eligibility screener

FIGURE S1: Internal and Stability Validation Plots for K-means, PAM, and Hierarchical Clustering

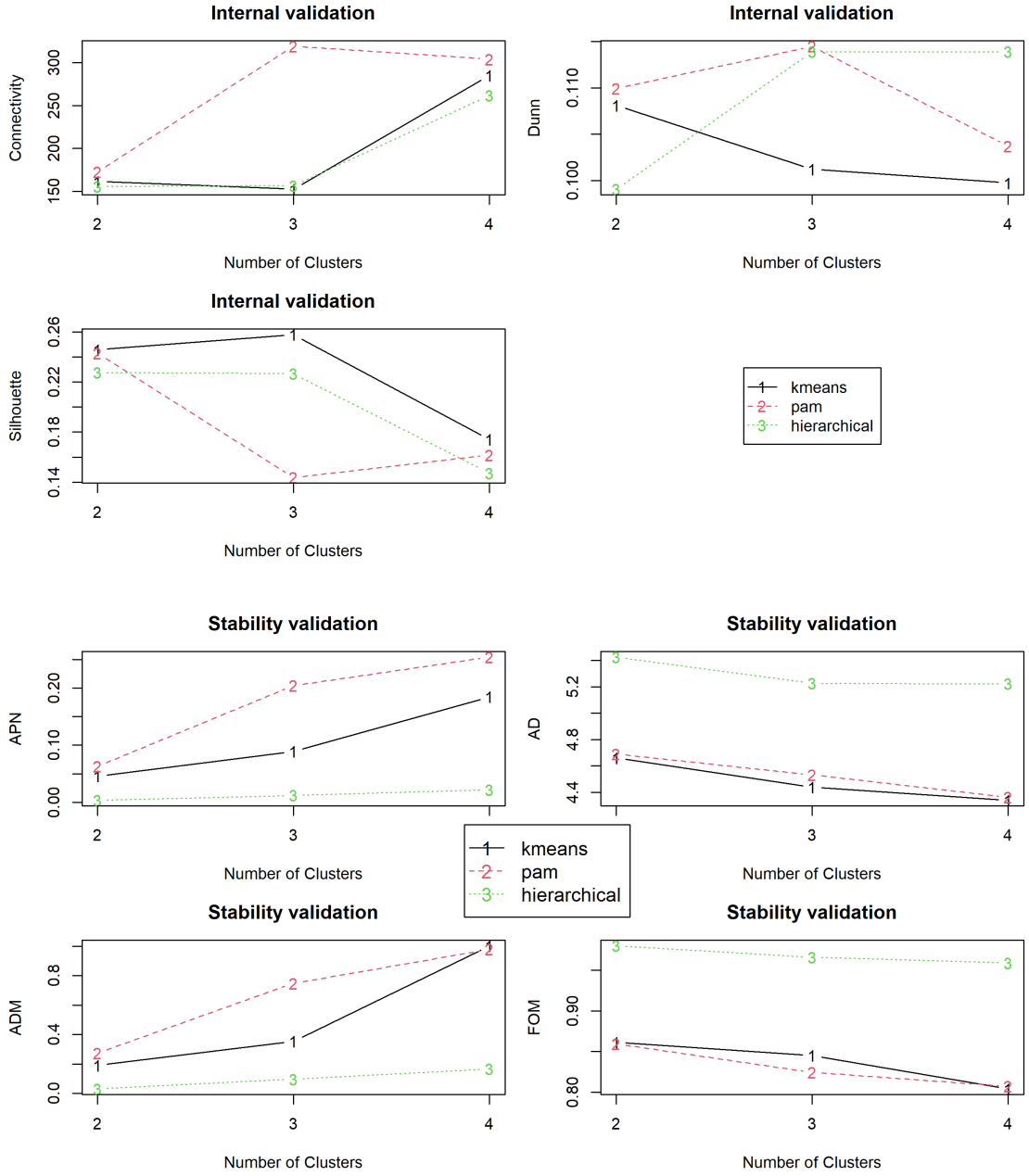


FIGURE S2: Visualization of Clusters Using Principal Components for 2, 3, and 4 Clusters

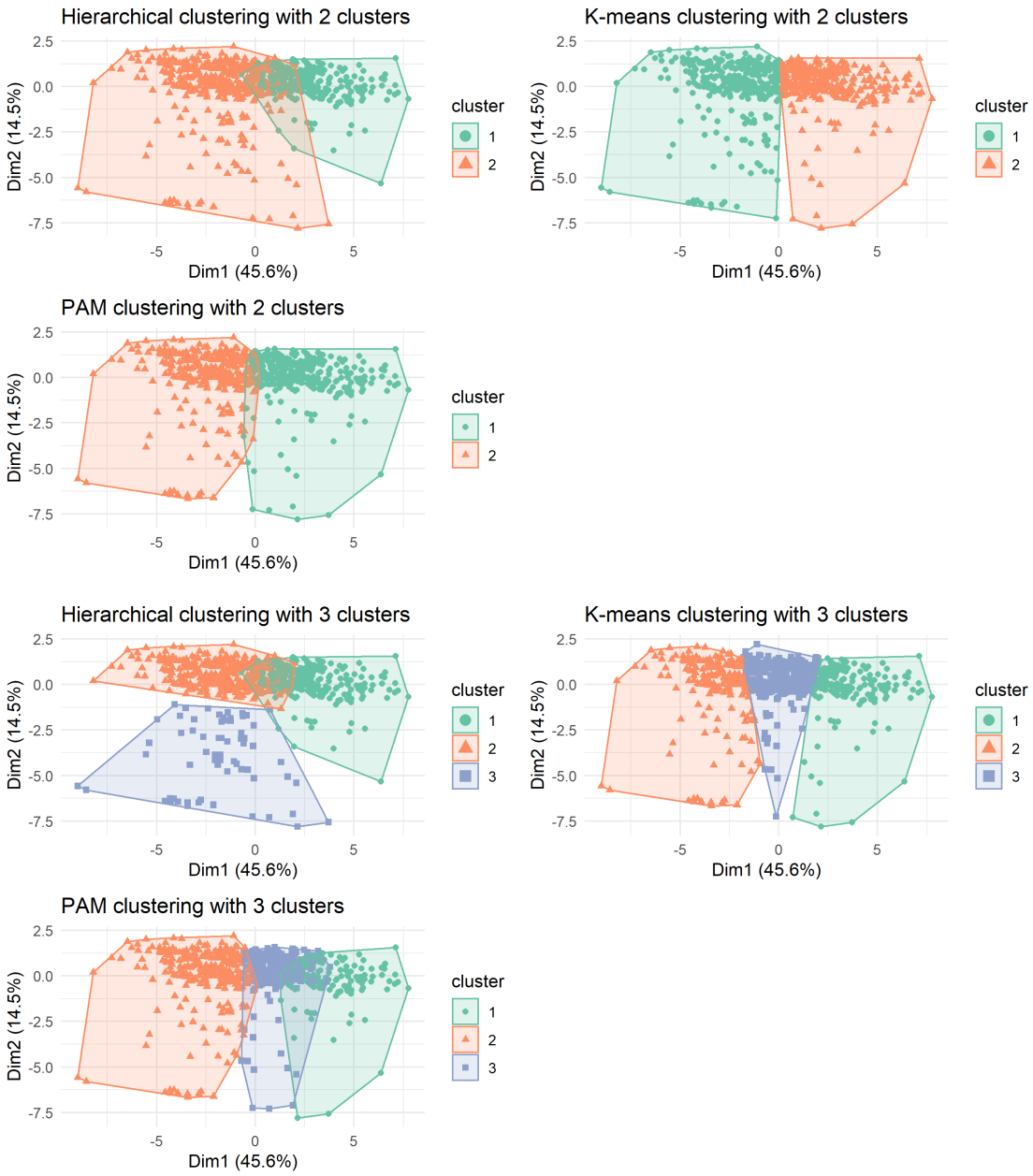


FIGURE S2: Visualization of Clusters Using Principal Components for 2, 3, and 4 Clusters (cont.)

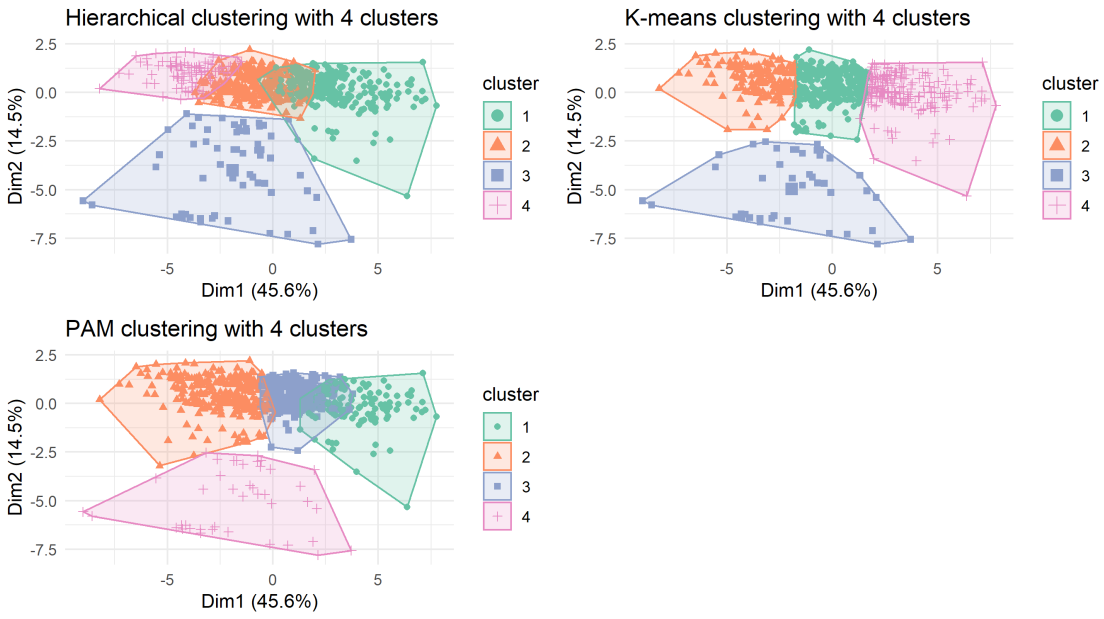


FIGURE S3: Boxplot and Violin plot Visualization of Time Spent on Training Components by Engagement Group

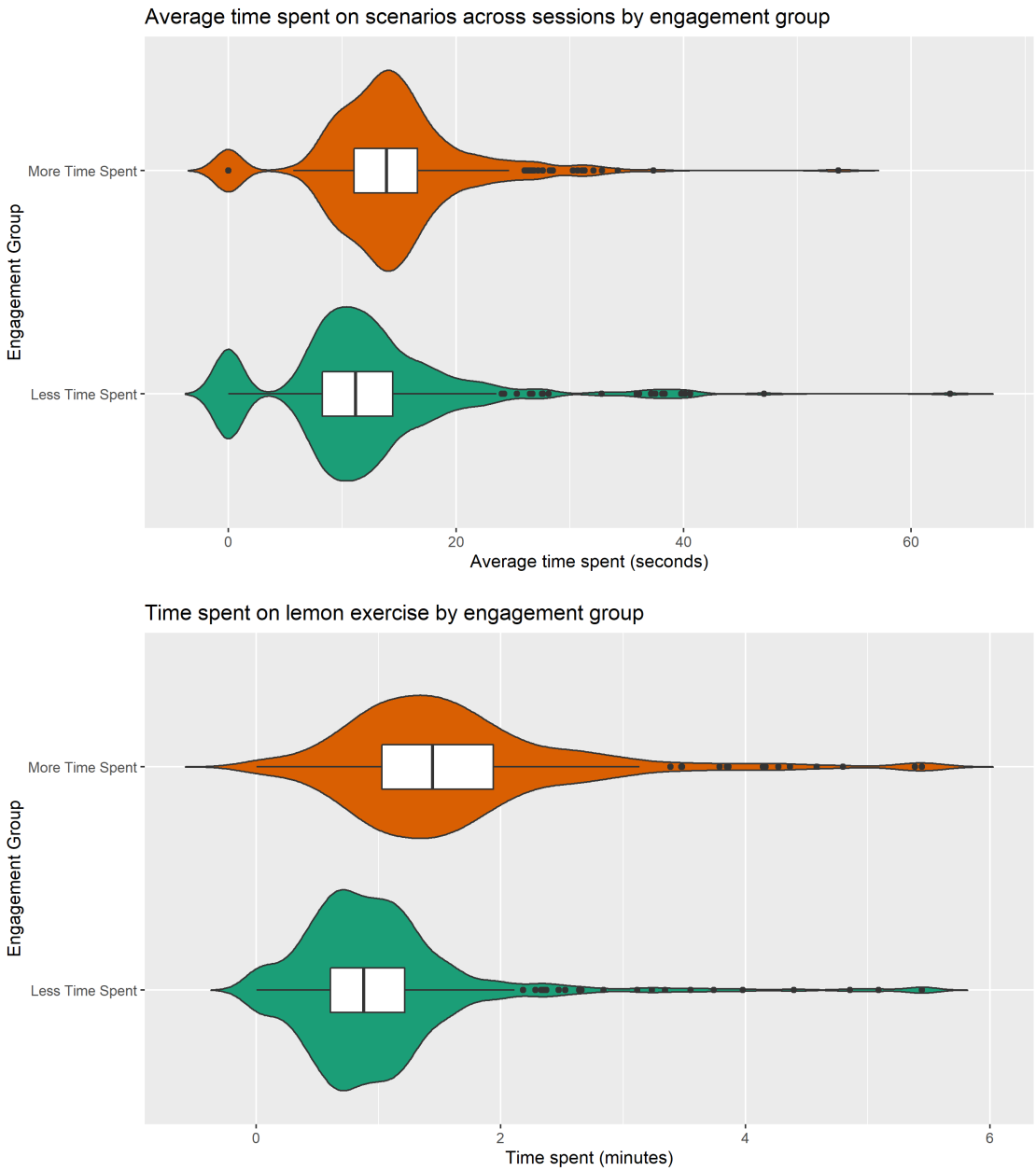


FIGURE S3: Boxplot and Violin plot Visualization of Time Spent on Training Components by Engagement Group (cont.)

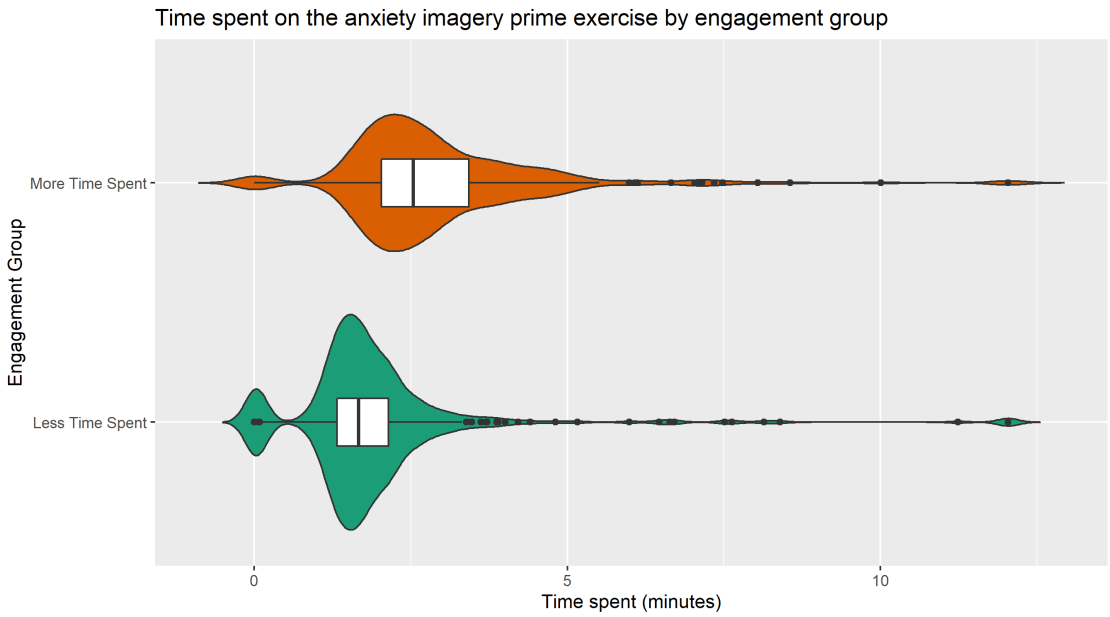


FIGURE S4: Boxplot Visualization of the Log of Average Time Spent on Measures by Engagement Group

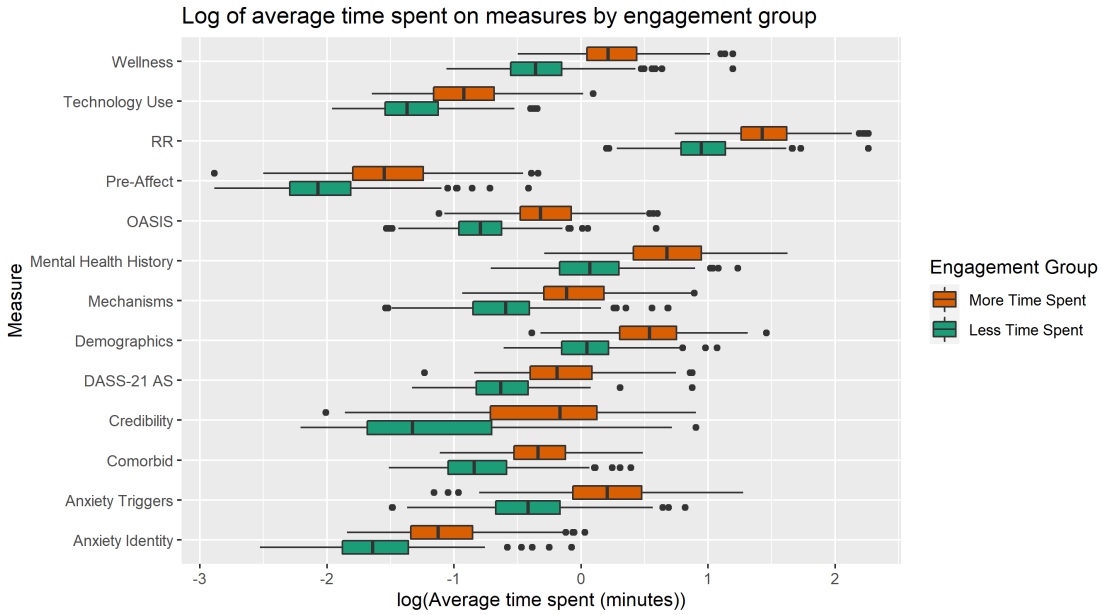


FIGURE S5: Density Distribution of the Log of Average Time Spent on Measures by Engagement Group

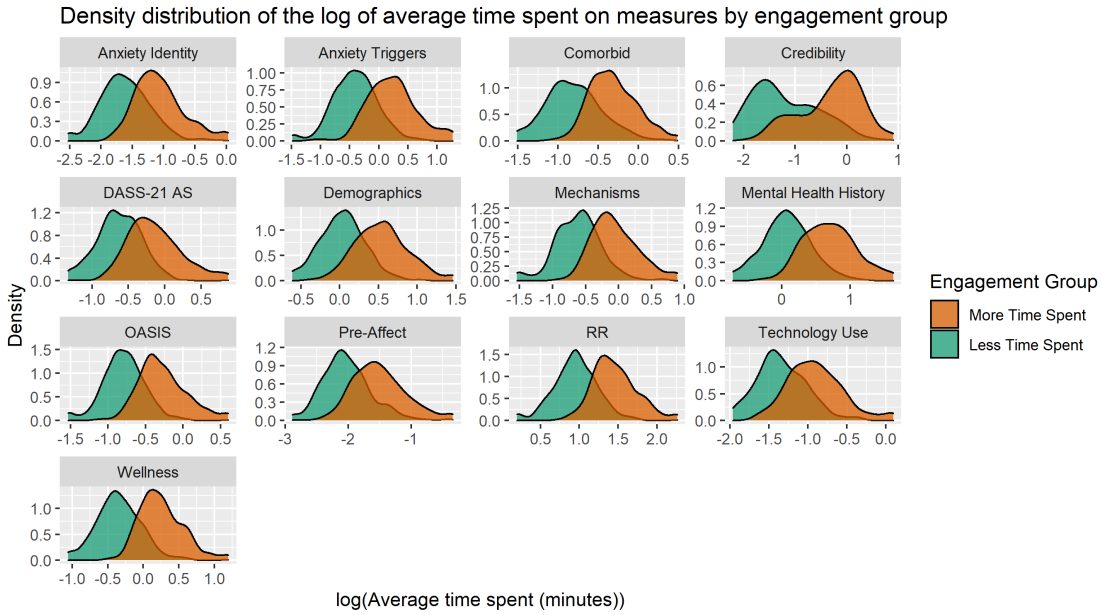


FIGURE S6: Density Distribution of Completion Rate by Engagement Group