Metrics for Peer Counseling: Triangulating Success Outcomes for Online Therapy Platforms

Tony Wang
Georgia Institute of Technology
Atlanta, Georgia, USA
tywang@gatech.edu

Haard K. Shah
7 Cups of Tea
Parlin, NJ, USA
haard@7cups.com

Raj Sanjay Shah
Georgia Institute of Technology
Atlanta, Georgia, USA
rajsanjayshah@gatech.edu

Yi-Chia Wang
Independent Researcher
San Mateo, CA, USA
yichia.wang@gmail.com

Robert E. Kraut
Carnegie Mellon University
Pittsburgh, PA, USA
robert.kraut@cmu.edu

Diyi Yang
Stanford University
Stanford, CA, USA
diyiy@stanford.edu

ABSTRACT

Extensive research has been published on the conversational factors of effective volunteer peer counseling on online mental health platforms (OMHPs). However, studies differ in how they define and measure success outcomes, with most prior work examining only a single success metric. In this work, we model the relationship between previously reported linguistic predictors of effective counseling with four outcomes following a peer-to-peer session on a single OMHP: retention in the community, following up on a previous session with a counselor, users' evaluation of a counselor, and changes in users' mood. Results show that predictors correlate negatively with community retention but positively with users following up with and giving higher evaluations to individual counselors. We suggest actionable insights for therapy platform design and outcome measurement based on findings that the relationship between predictors and outcomes of successful conversations depends on differences in measurement construct and operationalization.

CCS CONCEPTS

• **Human-centered computing** → Empirical studies in collaborative and social computing.

ACM Reference Format:
Tony Wang, Haard K. Shah, Raj Sanjay Shah, Yi-Chia Wang, Robert E. Kraut, and Diyi Yang. 2023. Metrics for Peer Counseling: Triangulating Success Outcomes for Online Therapy Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3544548.3581372

1 INTRODUCTION

People in need of mental health support have reported benefits from interacting with peers through online mental health platforms (OMHPs) [49]. These platforms have been growing in popularity [68], are accessible and cost-effective [56], reduce stigma about mental health by building anonymous connections among individuals [10, 51], empower the sharing of individual journeys [45, 78], and enable individuals in times of need to find advice and support for their problems [64, 73, 79]. Prior research has provided valuable insights on effective support strategies for those in need by studying a variety of success metrics such as satisfaction with support [3, 70], mental health status [20], community participation [73, 74], and linguistic behavior such as amount of self-disclosure [72, 78]. The use of diverse metrics offers online platforms the ability to track the impact of peer counseling holistically.

Despite the multiplicity of perspectives in the study of OMHPs, prior research has tended to use singular outcomes without examination of a larger body of potentially meaningful metrics. Examining a single outcome may lead to non-robust and non-generalizable findings, and peer counseling strategies could correlate with different success outcomes in inconsistent or conflicting patterns, limiting potential applications in the design of OMHPs. In the related area of computational psychology, interest in measurement validity has led to calls for metrics triangulation. Chancellor and De Choudhury [18] noted that lack of transparency in the operationalization of predictor variables raises concerns regarding validity, algorithm choice, and replicability of research that aims to predict mental health status using social data. Ernala et al. [27] suggested triangulation of diagnostic signals for predictive models to remedy issues in the validity and contextualization of predictor variables.

Outcome triangulation can be used to study conversation success for OMHPs, which are simultaneously community platforms, on-demand counseling services, and clinical interventions. Some previous studies have researched how a user's continued engagement with an OMHP constitutes a successful outcome in terms of a community's ability to support those seeking help [73, 78]. Others have found that online mental health support is sought out during major life transitions, arguing that a user may leave a community based on many factors including when a user has received sufficient support [47, 80]. This raises an important question for healthcare technology design: How do we balance the notion that peer support increases user engagement with a community yet decreases the probability of users staying on a platform? Triangulation can begin addressing this question by allowing us to compare the social and clinical outcomes of OMHP design.

To the best of our knowledge, no studies have paid attention to systematically triangulating outcomes in studying peer counseling

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581372

on OMHPs. We expect that using multiple metrics will reveal tensions among outcomes and predictors of successful peer counseling. We examine this expectation by proposing two research questions (RQs):

- RQ1: Does triangulating across multiple outcomes provide novel insights for finding indicators of counseling success?
- RQ2: Do widely used predictors of counseling success have consistent relationships with separate outcome metrics?

We first review prior research and note inconsistencies and ambiguities in the way that different authors define and operationalize OMHP outcomes. Next, we document our process for identifying four measures of peer counseling success for therapy-focused platforms: retention in the community, following up on a previous session with a counselor, users' evaluation of a counselor, and changes in users' mood. Then, using statistical analysis of a large dataset of one-to-one chats between support seekers and support providers, we compare the relationships of several widely used predictor variables with these four measures. Results show that predictors of successful conversations correlate negatively with community retention and correlate positively with the likelihood of users following up and giving higher counselor evaluations.

Contribution. In this paper, we make an original contribution toward unifying diverse approaches to online peer counseling research by triangulating different success outcomes and how they correlate with widely used predictors for counseling. Concretely, we leverage a dataset of 1.7M chat sessions between support seekers and support providers to conduct a large-scale regression analysis of peer counseling on 7 Cups of Tea, an online therapy-focused platform. By investigating the relationship between widely-used linguistic predictors of effective counseling and multiple outcome metrics, we validate different prior reported outcomes using large-scale data from an online therapy platform. We discuss the implications of results and provide novel insights for therapy platform design with peers.

2 RELATED WORK

We review prior literature to support our modeling process by identifying work on measuring peer counseling success, quantifying peer counseling strategies, and triangulation methodology for interpreting multiple metrics.

2.1 Measuring Peer Counseling Success

A variety of social and clinical outcomes with different constructs and operationalization methods have been reported in OMHP research. Social outcomes examine factors that influence seeker behavior towards the group after receiving support, reflecting theories of participation in groups to make sense of illness [43, 47]. For example, engagement has been defined as posting within the same conversation thread [63] or posting in any thread across an entire platform [73] after an initial post to an online forum. Engagement may also be broken down into its own factors, such as engagement with specific types or groups in a community and how diverse that engagement is across multiple topics [61]. Other studies have used attitudinal measures to capture seeker attitudes toward individuals or the broader community [39, 70, 78].

Clinical approaches emphasize the potential of online platforms as interventions, focusing on individual changes to mental health status using a diverse range of psychological and linguistic outcomes. Ideally, status may be measured directly through the use of validated clinical questionnaires [20]. However, challenges in sampling and demographic differences among various platforms inhibit the use of lengthier clinical instruments that cause drop off in responses. Mood has been used as an alternative construct in some prior studies, either as a single question metric [3] or as a behavioral pattern based on the types of topics or conversations users have [41]. Other constructs include cognitive-behavioral changes as a result of counseling such as moments of cognitive change, a measure of whether a seeker learns to reframe a problem [57].

The same construct may be operationalized differently, posing a challenge in examining the validity of previously reported outcomes. For example, depression can be measured from seekers' self-reports using a PHQ-9 clinical questionnaire [20] or behaviorally using word choice and writing style [61]. The operationalization method is critical to OMHP research as biases can be introduced during data collection and feature engineering. [27] found that predictive models of mental health status using social media data had strong internal validity during model building, but showed weak external validity when used for clinical diagnosis due to sampling bias and unclear construct validity. If outcomes for peer counseling are meant to capture the effectiveness of therapeutic techniques, then a gap exists in understanding when, where, and how to select outcomes to evaluate platform-wide counseling success.

2.2 Quantifying Peer Counseling Strategies

In order to study factors that lead to successful counseling online, recent work has also begun quantifying effective counseling strategies. Supporters on therapy platforms such as 7 Cups of Tea[1] and TalkLife[2] receive little to no counseling training [81] in contrast to professional therapists who are trained using multiple feedback channels [13, 15, 75] and volunteer crisis counselors who undergo dozens of hours of practice and receive feedback from supervisors [30, 53]. Motivated by the fact that peer counselors often do not have the same feedback mechanisms or training as traditional therapists, automatic methods have been leveraged for capturing counseling expertise on a variety of platforms such as crisis hotlines [3, 83], counseling platforms [62, 63], and social media [61, 64].

Effective counseling depends on therapist behaviors, the ways in which a conversation is carried out by a therapist, and therapist processes, the strategies a therapist chooses to work on with a client [21]. For crisis hotlines, where one-on-one conversations are the primary method of seeker-provider interaction, successful volunteer counselors have been found to be less rigid and more adaptable in their control of the flow of conversation. This is achieved by using more diverse language, writing more in response to ambiguous messages, and changing the topic of conversation more deftly to deal with diversity in support seeker needs [3, 83]. The study of linguistic and social norms on OMHPs with public forums has examined factors such as the amount of informational support and emotional support [41, 64, 70, 73] and self-disclosure [78] in conversations.

[1] www.7cups.com
[2] www.talklife.com

Other research has used word distributions to study the amount of affect used in communication [41, 57, 61], topical distributions [19], or writing behaviors such as linguistic style [61, 64].

2.3 Triangulation Methodology

Triangulation allows researchers to better ensure the validity of their findings [22, 38] and has a rich history in both HCI [46, 55, 69] and healthcare research [5, 11, 23, 52]. A range of methodological triangulation approaches have been used for studying the social and clinical outcomes of online health communities that involve peer support. Yao et al. [80] cross-examined interviews, surveys, and behavioral logs to develop a holistic view of journeys on the Cancer Survivors Network (CSN), a forum for discussion of cancer-related topics. Ma et al. [45] used classification methods and survival analysis to examine the role of expressive journal writing in user engagement and associated survival rates on CaringBridge, a blogging platform. Zhang et al. [84] used behavioral logs and clinical questionnaires to capture the correlation between mental health app usage habits and better clinical outcomes to define clinically meaningful use of technology. In these studies, triangulation occurs in corroborating multiple predictors of specific outcomes.

While such studies highlight the added interpretability of methods triangulation, outcome triangulation is also valuable in systems and platform research because multiple outcomes provide a comprehensive overview of user engagement [54] and empower design for large numbers of users [31] as a form of data analytics. Rodden et al. [60] proposed happiness, engagement, adoption, retention, and task success as complementary metrics for tracking user experience (UX) in large-scale products, each of which may be operationalized and prioritized differently depending on goals for design, development, and research. Healthcare analytics leverages multiple types of data and insights to support better decisions for healthcare providers [58]. Analytics have also been used to examine challenges to designing effective online mental health support systems [66, 82]. Inspired by this line of work, we focus on the triangulation of multiple outcome measures and their associations with different linguistic predictors of conversation success.

3 TRIANGULATING OUTCOME MEASURES FOR PEER COUNSELING

Our aim is to empower designers and researchers of OMHPs by unifying social and clinical perspectives in the literature and to develop a proposal for platform-level decision-making with regard to identifying success outcomes. In this section, we first conduct a review of outcome measures and organize them by the construct they measure. Next, we outline multiple metrics available on the 7 Cups of Tea platform and discuss our choice of outcome metrics.

3.1 Literature Review

Since our approach blends both social and clinical perspectives, we conduct a review of outcome measures for peer counseling using a keyword search of papers [44] defining outcomes for participation in platforms or communities for online peer counseling, mental health support, and therapy. Studies in which peer support was part of a larger body of mental health-related features were excluded. Qualitative analysis of outcomes was excluded to focus on metrics for platform analytics. Identified outcomes were grouped by community, conversational, and individual levels based on social computing, psychotherapy, and clinical perspectives respectively. We also noticed differences in method of measure that lead to multiple operationalization of the same constructs in the literature. Lastly, we identify prior work that has used multiple outcomes and note differences in their approaches to ours.

3.1.1 Community. Community-level constructs track behaviors or attitudes spanning interactions with a broader group of many users. Examples include continued seeker engagement across multiple conversations [45, 73, 74], commitment [77], support seeking behavior [78, 80], support provision behavior [14, 80], or general attitudes toward the community such as the desire for support [39, 80]. Community metrics may also track broader perceptions about the platform on which the community is built, reflecting the UX nature of measuring attitudes [42]. For example, Alvarez-Jimenez et al. [4] evaluated an online social therapy platform for first-episode psychosis recovery using attitudinal measures of perceived degree of social interaction and platform usability.

3.1.2 Conversation. Conversation-level constructs track outcomes of support provided by specific individuals, capturing the impact of dyadic relationships within public forums or private chats. Prior work in this space has leveraged both attitudinal and behavioral measures. Attitudinal measures such as the Session Rating System [10], counselor helpfulness [83], and counselor rating [62] align with psychotherapy research aimed at evaluating client-therapist alliance [12, 35]. Some papers have examined behavioral outcomes such as amount of self-disclosure in response to what others have said [7, 8, 72, 78], linguistic alignment [71], and amount of support provision [64, 71, 79, 80]. Vlahovic et al. [70] evaluated seeker satisfaction with support using third-party annotation methods.

3.1.3 Individual. Individual-level outcomes capture the impact of receiving support on a single user's cognitive state. Outcomes include moments of change in how an individual thinks about a problem [57, 61], clinical questionnaires of mental health symptoms [4, 20, 41, 84], mood [3, 41], and various psycholinguistic proxies of mental health status such as affective language use and cognitive behavior [61].

3.1.4 Method of Measure. Several outcomes not only spanned multiple construct levels but also had differences in how they were measured. For example, support provision was operationalized as both an annotated, community-level outcome and a behavioral, conversation-level one. The former occurred in the analysis of social roles that an individual adopts (e.g., old-timers who contribute significantly to their communities) among the larger community [14, 80]. The latter was measured within a conversation in which a member of an online community provided informational or emotional support to a seeker [79]. Satisfaction was tracked as an attitudinal metric at the community level with regard to general peer counseling experience on a platform [82], as an attitudinal metric at the conversation level in response to experience with a specific counselor [10], and as an annotated metric at the conversation level based on seeker responses to receiving support [70]. To accommodate a wide range of outcome operationalizations, we note a second

dimension for the method of measure in our literature review to account for overlaps in measured construct but differences in data collection methods.

3.1.5 Multiple Outcomes. Some studies leveraged more than one outcome to measure multiple hypotheses of peer counseling impact [4, 41, 78, 80]. Others report hybrid constructs that span multiple construct levels or methods of measure depending on platform or experiment design. For example, threaded conversations may be conversations when there is only a single user responding to a top-level post but can generalize to include multiple discussions using a nested structure. This has led to engagement outcomes that measure both community and conversational results depending on the number of providers within a thread [73, 74]. We found one example of a composite outcome that used multiple outcomes across individual and community levels to create an overall proxy score for mental health [61]. In general, outcomes were chosen or created based on the theories being tested, which may lead to challenges in measurement validity [9, 18]. This finding supported our hypothesis that systematic outcome triangulation is an opportunity to discover new insights about outcome selection in OMHP research.

3.2 Selecting Outcome Variables for Triangulation on 7 Cups of Tea

We contribute to research on effective peer counseling by studying conversations and multiple outcomes on 7 Cups of Tea (7 Cups), a coping and therapy platform where users can discuss a variety of issues with peer volunteers who are ready to listen. The platform allows support seekers to register as *members* and support providers to register as *listeners*. A user can also register as both roles. 7 Cups requires all listeners to complete a 30-60 minute long initial training that teaches various talk therapy techniques such as active listening, showing empathy, summarizing and reflecting back to members their concerns, and asking guiding questions. Chats on 7 Cups start with a member requesting support. Available listeners then choose to chat with members based on incoming requests, a process that allows listeners to match themselves with requests related to topics they specialize in or have personal experience with.

7 Cups was chosen as a research site for validating prior work on peer counseling because the majority of interactions on 7 Cups are through private, one-to-one conversations where member-listener pairs converse anonymously. Private messages comprise roughly 90% of over 400 million messages sent between users of the platform from January 2020 to August 2022. Other mental health apps for therapy such as BetterHelp and TalkSpace focus on professional therapist services while online communities with public spaces such as Reddit, CSN, or TalkLife center around many-to-many interactions. Although 7 Cups offers these features as well, the emphasis on connecting seekers to providers in chats offers a close environment for replicating prior work on peer counseling strategies conducted on one-to-one channels such as crisis hotlines.

Next, we describe available member outcomes for 7 Cups and our rationale for selecting specific ones for triangulation.

3.2.1 Community. One challenge with platform metrics is identifying meaningful measures of community outcomes as 7 Cups offers a number of individual and social features. In light of prior ambiguities in construct validity, we define community-level outcomes as a variable representing a member's relationship with more than one 7 Cups user. Although 7 Cups administers several attitudinal measures such as a product market fit survey [25] and the net promoter score [59], these metrics may not capture community outcomes since several platform features on 7 Cups do not support peer-to-peer interaction. Similarly, product reviews may contain insights about user experience, but they are likely to reflect high-level perceptions of 7 Cups such as interface usability or app design [2]. As a result, we considered possible behavioral outcomes at the community level based on user logs data.

- **Engagement** is the presence or absence of a member's continued participation in spaces with other 7 Cups users [41, 61, 63, 73, 74].
- **Frequency of participation** is a behavioral metric quantifying the amount by which something occurs such as the number of posts or responses over a member's lifetime on the platform [61, 64, 77, 80].

Retention was chosen as a community-level outcome and operationalized from engagement as whether or not a member chatted with other listeners after a conversation on 7 Cups. This parallels definitions of community engagement as posting or commenting in multiple threads in online health communities research. We select this measure to represent the significant body of prior work on continued participation as a meaningful outcome of receiving support in communities. Frequency of participation, operationalized as the number of past conversations a member had prior to conversing with a listener, was used as a control variable (Section 4.6).

3.2.2 Conversation. Conversation level outcomes represent member outcomes for a single conversation involving a member and listener. In addition to user logs, other metrics were available based on instruments currently deployed on 7 Cups.

- **Engagement** is the presence or absence of a member's continued participation in conversation [73, 74].
- **Rating** is a single-question attitudinal measure of a member's perception towards a listener [62].
- **Hearts** are a feature where members and listeners can react with a heart for any message, similar to emoji functionality for SMS-based text chats and the like button on social media. Hearts are a novel measure that we speculate to represent shallow engagement based on social media research [54].

We found no prior work exploring the use of a similar metric to hearts in peer counseling for OMHPs, although turn or message-level metrics exist in other domains [6, 50]. With our focus on examining previously reported outcomes, we did not pursue adding hearts as a novel outcome due to a lack of interpretability. However, to account for the possibility that hearts serve as a proxy for familiarity with platform functionality and shallow engagement in a conversation, we use the historical usage of hearts as a control variable (Section 4.6).

Two conversational outcomes, follow-up, and rating were chosen to address our primary interest in peer counseling. Follow-up was operationalized from engagement because we hypothesized that support provision should lead to a higher desire to continue a

		Construct Level		
		Individual	**Conversation**	**Community**
Method of Measure	**Attitudinal**	BAI [4] BPRS [4] CDSS [4] GAD-7 [20] PHQ-9 [20, 41] Mood [3, 10, 41]	Rating [62, 83] Satisfaction [10] Session rating scale (SRS) [10] Support provision [80]	Attachment [77] Ease of use [4] Helpfulness [4] Information utility [39] Participation [39] Perceived support [39] Patient empowerment [39] Satisfaction [82] Social interaction [4]
	Behavioral	Affective word use [61] Complexity or repeatibility [61] Psycholinguistic keywords [61] Readibility [61] Readibility [61] Symptomatic word use [61]	Conversation length [7, 8] Engagement [63, 73] Frequency [8] Support provision [79]	Engagement [45, 63, 73, 74, 82] Length of participation [77] Number of posts [61] Number of topics [61] Number of responses [61, 77] Support seeking [78]
	Annotation	Moment of change [41, 57]	Satisfaction [70] Self-disclosure [7, 8, 78, 79] Support provision [64]	Support provision [14, 80]

Table 1: Outcome measures employed in this study (purple) with referent work organized by construct level and method of measurement, contextualized using reported outcomes in peer support or therapy literature. We operationalize using behavioral measures *retention* and *follow-up* from community and conversation engagement respectively.

conversation [61, 63, 73, 74, 79]. Rating data was chosen as a metric for tracking satisfaction towards support provision [10, 70, 83].

3.2.3 Individual. Individual outcomes reflect measures of the member's mental state. 7 Cups administers multiple self-report measures at different frequencies. In addition, user logs were also available for creating psycholinguistic proxy variables of mental health status.

- **General mood** is a single-question instrument that occurs at most once every hour asking a member how they feel at that moment, similar to a mood question used in crisis hotlines [3].
- **PHQ-9** is a nine-item battery of questions for depressed mood [20] administered at most once every two weeks.
- **GAD-7** is a seven-item battery of questions for anxiety [20] administered at most once every two weeks.
- **Psycholinguistic proxies of mental health status** could be used to measure a member's cognition based on language use found in user logs as reported by [61].

Mood was chosen over clinical questionnaires to replicate studies by Althoff et al. [3] and Kushner and Sharma [41], both of which used mood as a proxy for mental health status. While clinical questionnaires are the gold standard in clinical research, the two-week delay in the administration of 7 Cups led to doubts about sensitivity to peer counseling effects. We did not pursue the replication of a psycholinguistic outcome following Saha and Sharma [61], as they conducted a group-level aggregate analysis by counselors with this metric. It was unclear if their method generalized to our session-based analysis while all other outcomes in our study had been applied in similar contexts examining individual counseling sessions. Furthermore, the moment of cognitive change was excluded from modeling as it may be confounded with our goal of identifying successful peer counseling sessions: a high rating or lack of follow-up may be the result of a moment of cognitive change, which may appear before a session is concluded.

3.3 Situating Outcomes in Literature

The review of literature and available measures on 7 Cups illustrated a need to differentiate between construct levels and methods of measure as part of outcome variable operationalization. As the method of measure can have an impact on validity of outcomes, we add a dimension capturing attitudinal, behavioral, and annotation methods of measure. Attitudinal measures offer a more direct connection to a user's perceptions but suffer from issues with response rate or human bias [42]. Behavioral measures such as engagement benefit from being observed [60], avoiding pitfalls with issues in drop-off or bias in reporting. More recently, human annotation using experts or crowd workers has been used as a method of labeling behavioral outcomes for machine learning models [18].

Table 1 situates our four outcome measures among literature reviewed in Section 3.1 organized by construct level and method of

measure. In total, we use four metrics across three different construct levels and two methods of measurement. We position our work as a form of external validation, leveraging regression models to ascribe relative merits to existing constructs and operationalization methods [9, 42]. Considering that human annotation is also a form of external validation and the private nature of 7 Cups chats, we choose not to pursue a secondary validation method through annotation for this particular study and instead focus on investigating attitudinal and behavioral outcomes. An attitudinal community outcome and behavioral individual outcome were not included for triangulation following the selection process in Section 3.2.

4 METHOD

4.1 Ethical Considerations

This research study has been approved by the Institutional Review Board (IRB) at the researchers' institution. Data was provided in collaboration with 7 Cups and the data collection process follows HIPAA and confidentially agreements. Given the private nature of the conversations, the authors adopt additional steps to protect the participants' privacy throughout the research process such as anonymizing data to prevent association with any particular user. All researchers associated with this work have completed Collaborative Institutional Training Initiative (CITI Program) certifications.

4.2 Dataset Creation

We received access to all conversations between March 1st, 2020, and March 7th, 2022 via a data use agreement with 7 Cups of Tea. To study the effectiveness of counseling sessions on 7 Cups, we follow Kushner and Sharma's approach to studying bursts of activity inside conversations [41]. A conversation is split into multiple sessions if two bursts of messages between seekers and support providers are separated by five days or more. The five-day interval represents three standard deviations of time between consecutive messages in the dataset. This resulted in over 7 million sessions. Since a seeker-supporter pair could have multiple sessions, we only analyze the first to ensure that outcome measures are not influenced by the contents of prior sessions. This reduced the dataset to over 4 million sessions. Lastly, we analyzed sessions only up until March 1st, 2022 but tracked both retention and follow-up for one week after this date, following the average of five days for session intervals. We acknowledge this as a limitation in dataset creation.

Next, we examined the length of sessions to focus on modeling conversational factors. The median number of messages per session was 47 (mean 79.8) and 50% of the dataset had 3 or fewer messages. Manual examination of conversations with less than 10 messages revealed a significant portion of discontinued conversations; e,g., Listener: "Hi, welcome to 7 Cups!", Member: "Hey.", Listener: "Is there anything on your mind today?" before the member stops responding. Because we were not interested in factors that could lead conversations to stop prematurely, we removed short sessions from the dataset with less than 20 messages, which is less than half of the median number. This threshold resulted in a final dataset of 1,739,398 sessions containing 583,842 members and 118,701 listeners. Even after filtering sessions with fewer than 20 messages in length, over 1,000,000 sessions were between 20 to 50 messages in length as shown in Figure 1.

4.3 Outcome Variables

Next, we describe concrete operationalizations for outcome variables chosen in Section 3.

4.3.1 Retention. Retention is a community-level behavioral measure that describes whether the member returns to the platform to have another conversational session with a different listener after the member's first session in our dataset. We give a binary label 1 if the member returns for another conversation and 0 otherwise. High retention is considered a positive outcome of support prevision by an online community [73, 78].

4.3.2 Follow-up. Follow-up is a conversation-level behavioral measure that describes whether the member returns to the platform to have another conversational session with the same listener after the first session. We treat this as a binary label of 1 if the member returns to continue a conversation and 0 otherwise. A higher likelihood of follow-up is considered a positive outcome of participating in conversations within a community [73, 78].

4.3.3 Rating. Rating is a conversation-level attitudinal measure that a member can leave for listeners at any point during or after a conversation. Ratings vary from 1 (worst) to 5 (excellent) and can be given anonymously. We ensure that a rating describes the current session by accepting reports during the session or within the five-day window separating sessions.

4.3.4 General Mood. Mood is an individual-level attitudinal measure asking "How are you feeling right now?" with responses varying from 1 (Awful), 2 (Bad), 3 (Okay), 4 (Good), to 5 (Great) that is administered to members independently of conversations with listeners. We capture the post-session mood as an outcome measure and use the pre-session mood as a control. We ensure that mood scores before and after a session are within a 24-hour window, following sampling practice recommended in other clinical evaluations of mood [1, 48]. If a member reports multiple mood self-reports within 24 hours after a session, we take the score closest to the end of the session. Higher mood scores are positive outcomes from the perspective of OMHPs as interventions [3, 20].

4.3.5 Outcome Distributions. The distribution of each outcome is shown in Figure 2. 1,569,185 (90.2%) sessions result in member retention on the platform and 139,769 (8%) of members followed up with the same listener after the initial session, reinforcing Baumel's [10] findings that members on 7 Cups "shop around" to find a listener they would like to hold a long conversation with. 261,940 (13.09%) of sessions have a rating in which a member evaluates a listener. Ratings follow a J-shaped pattern seen in prior research on product and service reviews [36].

While a total of 14,434 (1.5%) of sessions had post-session mood score reported within 24 hours, only 4,839 (0.8%) of sessions have both pre- and post-session mood reports within 24 hours. The low percentage of reported mood data relative to the amount of sessions may be due to a combination of factors. The mood question is the most frequent measure employed on 7 Cups, which may lead to members selectively responding depending on their mental health status due to its availability. This may lead to bias towards negative mood reports. Furthermore, administration is triggered by platform usage rather than conversations, so members are not encouraged to

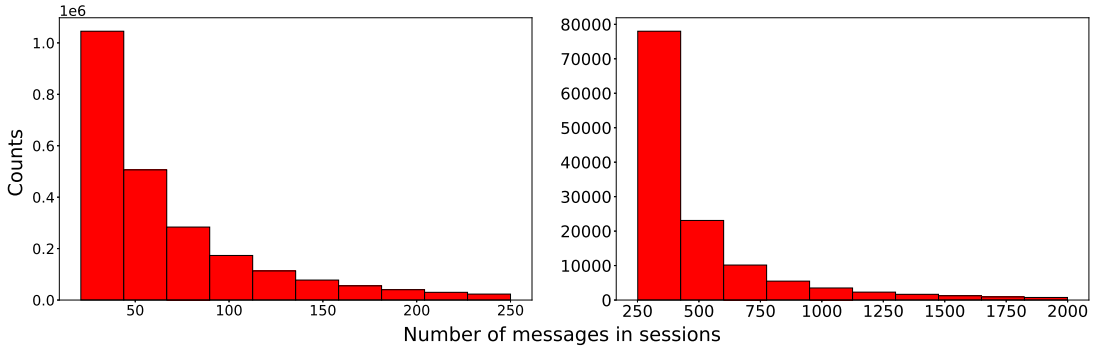

Figure 1: Counts of the number of sessions based on the number of messages in them at different granularities after filtering out sessions less than 20 messages in length. Over 1,000,000 sessions had less than 50 messages (left). Over 100,000 sessions had over 250 messages long and remained a small portion of the dataset (right).

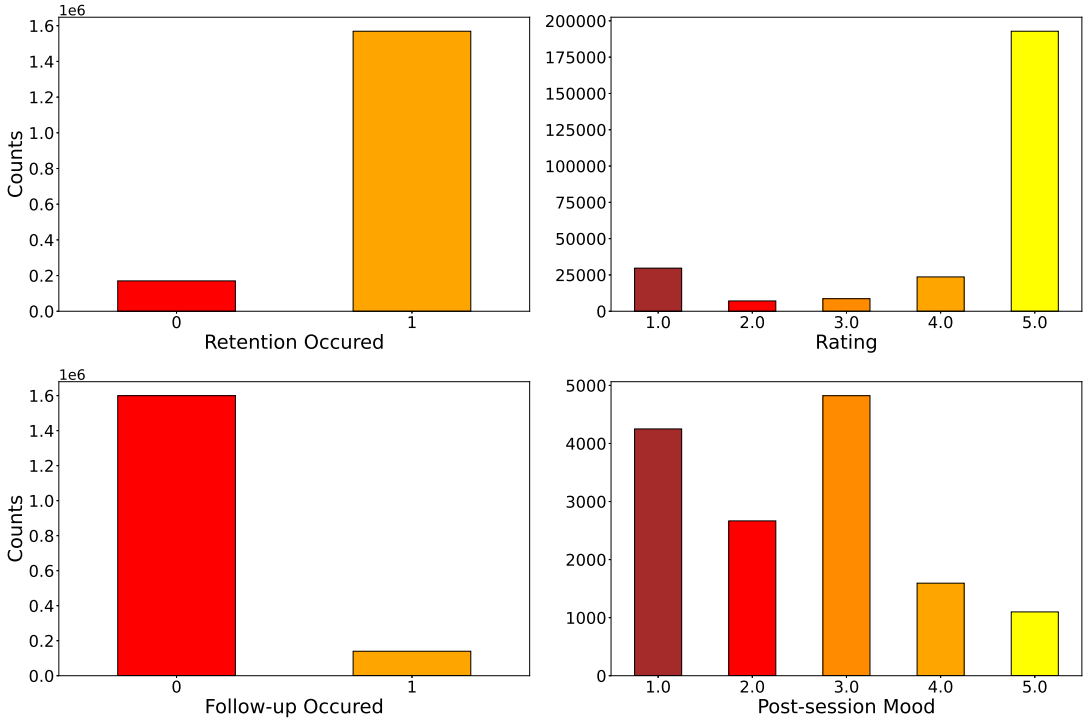

Figure 2: Distribution of outcome variables.

report mood immediately before and after a chat. Lastly, platform design elements such as usability or discoverability may also impact survey response rates.

4.4 Model Choice

Table 2 summarizes our model choices. Behavioral outcomes, retention and follow-up, are binary variables without missing data, so we use logistic regression for modeling. Attitudinal data is difficult to triangulate with other metrics due to selection bias in reporting. For ratings, we employ the Heckman selection model [33] to first model the distribution of missing data (stage 1) and then fit an imputed Gaussian variable that approximates the rating distribution (stage 2). In prior work, eighty percent of conversations in [3] and two-thirds of the samples in [20] are discarded because users did not provide any feedback. We did not use the Heckman model for mood as the low number of samples can lead to significant bias in estimating the distribution of missing data [40]. Instead, we conduct ordinal regression on post-session mood score using pre-session mood score as a control variable. We acknowledge that the mood model uses a much smaller dataset than other models, which may introduce selection bias due to a non-random dropping of data.

Outcome	Model	Description	Samples
Retention	Logistic	Predict likelihood member stays on the platform	1,739,398
Follow-up	Logistic	Predict likelihood member follows-up with the same listener	1,739,398
Rating	Heckman	(Stage 1) Estimate latent variable for censoring based on observed data (Stage 2) Predict rating with Gaussian approximation of latent variable	(Observed) 261,940 (Total) 1,739,398
Mood	Ordinal	Predict post-session mood score	4,839

Table 2: Descriptions of each model and the amount of data used. Due to the small number of observed data for mood, we do not use a Heckman selection model to estimate mood across the entire dataset.

4.5 Explanatory Variables

Explanatory variables are based on prior work identifying what support providers do that influence success in peer counseling for mental health. All explanatory variables are mean standardized. Those with long-tailed distributions were log-transformed prior to standardization. As we did not seek to develop new variables for this work, we note explicitly prior work from which these variables are operationalized.

4.5.1 Total Words. This metric describes the quantity of content exchanged between member and listener in a particular session. A greater number of total words exchanged between users represents conversation progress and is associated with desirable mental health outcomes [3, 20, 72–74].

4.5.2 Member Words Ratio. We capture the proportion of words sent by the seeker relative to the total number of words as a measure of interaction between seekers and supporters [57]. Since 7 Cups trains listeners to listen, the ratio of member words suggests how much a member is willing to engage in conversation and how much a listener is dominating the conversation.

4.5.3 Member & Listener Self-disclosure. This construct describes how much information about oneself is given in a conversation [7, 8]. Previous research [73, 78] suggests that more personal self-disclosure could lead to more continual participation in a thread or on a platform. Following [72], we utilize pre-existing dictionaries from the Linguistic Inquiry and Word Count (LIWC) to capture counts of words that represent a discussion of personal details. Manual additions were made to this corpus for positive and negative affective words from LIWC, positive and negative emoticons seen in 7 Cups chat logs, pronouns that describe how much the user is talking about themselves or loved ones from the "I", "We", "Family", "Friends", "She/He", and "They" categories in LIWC, and common nicknames which users may use when talking about their loved ones such as 'hubby' or 'darling'.

4.5.4 Listener Median Response Time. This variable represents how quickly the listener responds. [63] investigate the effect of a similar measure and show that faster response time leads to favorable retention outcomes on TalkLife posts. We include this measure based on our inspection of shorter messages that were removed from the dataset in which members become disengaged, leave the platform, or leave negative ratings when a listener takes too long to respond or when a listener responds too quickly.

4.5.5 Linguistic Style Matching. We describe coordination between member and listener based on linguistic function words such as prepositions, conjunctions, articles, quantifiers, personal pronouns, and negation. Higher coordination suggests mutual attraction and shared understanding using a variable for linguistic style matching [29]. [64] showed that support seekers that conform to the general linguistic style of a Reddit community receive greater emotional and informational support from others. These results suggest that higher linguistic style matching may be a predictor of seeker-provider alignment and better counseling outcomes.

4.5.6 Topic Matching. The topic matching variable describes coordination between member and listener in terms of discussion content. We obtain various topic dictionaries from the dataset using Empath engine [28]. Topic dictionaries are used to obtain a vector that describes the topic distribution of members' and listeners' messages in a session. The topic-matching score is calculated by computing the euclidean distance between members' and listeners' topic vectors. Specific topics and the procedure for obtaining them will be described in the next session 4.6.

4.6 Control Variables

Control variables are variables representing individual traits such as prior behavior and overall engagement with 7 cups that we hypothesize influence outcome measures. We include various experience and demographic descriptors for both members and listeners as well as topical controls to account for potential variation in needs across topics previously reported on 7 Cups [10]. Control variables were also mean standardized and log-transformed when necessary.

4.6.1 Listener Experience & Demographics. These features allow us to remove confounds in outcomes as a result of individual factors or tenure on 7 Cups. Yao et al. [81] showed that experienced listeners are more competent in being able to navigate novel situations and difficult problems that members want to discuss. Listener experience variables include average hearts received, average rating received, badge count (listeners receive badges for completing listener training-related content), number of past conversations, and number of forum upvotes. One demographic variable, age, was used to control for life skills that older listeners may bring to counseling.

4.6.2 Member Experience & Demographics. Kushner and Sharma [41] show that users that are more persistent or active on a platform are more likely to experience a positive change. Member experience variables include average hearts given, average hearts received, average rating given, number of past conversations, and number

Topic	Session Count	Percent (%)
Self Improvement	579192	33.30
Dating	297088	17.08
Parents	200584	11.53
Depression	162014	9.31
Romantic Relationship	135624	7.80
Lonely	120538	6.93
Suicide	67428	3.88
Pandemic	38114	2.19
Anxiety	28012	1.61
Home	26715	1.54
Family	20285	1.17
Sexuality	20235	1.16
LGBTQ	11331	0.65
Dissociative Identity	9856	0.57
Overwhelming	8004	0.46
Stress	5188	0.30
Health	4920	0.28
Intimacy	4270	0.25
Total	1,739,398	100

Table 3: Empath topical overview of all first sessions on 7Cups from March 2020 to March 2022.

of forum upvotes. Age was used to control for life skills that older members may bring to cooperation with listeners. In addition, the pre-session mood score was used as an extra control variable only when modeling the post-session mood score. Similar to the post-session mood score, pre-session mood is captured within a 24-hour window prior to a session's start. If there is more than one mood score within the window, we use the most recent one.

4.6.3 Topics. Users arrive on 7 Cups with a wide variety of problems such as romantic relationships, loneliness, depression, and anxiety [10]. To this end, we utilize Empath [28] to generate custom lexical categories for various topics seen in our dataset. Empath draws connotations between words and phrases using neural word embeddings trained on large text corpora. To tune the model for our dataset, we use regular expressions to extract the top 18 most frequently discussed topics between new members and a chatbot during new user onboarding. Then, we randomly sample ten sessions in each topic and manually select seed words per topic to feed the Empath engine. Distributions of the primary topic for all sessions can be found in Table 3. The top 18 topics are romantic relationships, dating, pandemic, self-improvement, suicide, depression, parents, anxiety, family, stress, lonely, overwhelming, sexuality, LGBTQ, intimacy, home, dissociative identity, and health.

5 RESULTS

5.1 Models

Figure 3 shows the correlations among the four outcome variables in this study using Kendall's τ. Correlations are low, with the mean

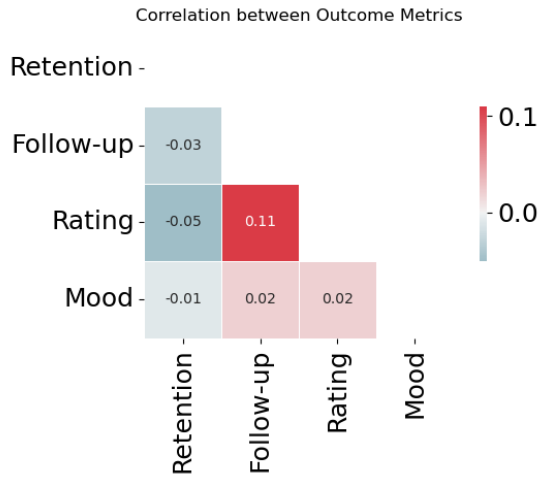

Figure 3: Kendall rank correlation coefficients between four outcome measures.

absolute τ coefficient being .04. Since these alternate measures of counseling success are relatively independent of each other, the low correlations offer the possibility that triangulating across the community, conversation, and individual levels may provide new insights for assessing counselor success (RQ1). The largest positive correlation is between rating and follow-up ($\tau = .11$), supporting our belief that conversation-level constructs should be more highly correlated with each other. The largest negative correlation is between rating and retention ($\tau = -.05$), which validated our choice in the triangulation stage to identify conversational-level outcomes separately from community-level ones. In general, mood and retention show a smaller correlation with other outcomes. Engagement at the community and conversational levels ($\tau = -.03$) can be seen as relatively independent measures with respect to predictors of effective counseling for an online therapy platform.

Next, we examine the consistency of predictor variables across outcomes (RQ2). We focus our analysis on consistency in coefficients for several reasons. The aim of our research is not to find the best-fitting model, but to understand the relationship between previously reported measures of effective counseling and measures of counseling success. For ratings and mood, report timing does not directly coincide with the end of a session and has the potential to include confounds within the temporal window. A large amount of observational data with unbalanced outcomes (e.g., 90% of the members in our dataset continue using 7 Cups) increases the likelihood of highly significant variables being found (type 1 error) and reduces the goodness of fit for each individual model. Differences in the logistic, Heckman, and ordinal regression models and the amount of data for each outcome make it difficult to directly compare effect sizes across models.

Table 4 reports regression coefficients for all models. Tests of collinearity for all independent variables are provided in A. Results for stage 1 of Heckman models are provided in B. For logistic regression models, the coefficients represent the log odds of the outcome occurring. For the Heckman model, the coefficients follow

Category	Features	Retention	Follow-up	Rating	Mood	Prior Work	Agree
Explanatory Listener perf.	Total words★	-0.240***	0.735***	0.257***	0.107*	Yes	Yes
	(M) words ratio	-0.0132***	-0.133***	-0.082***	0.118***	Yes	Yes
	(L) self disclosure	-0.035***	0.059***	0.030***	-0.032	Yes	Yes
	(M) self disclosure	-0.044***	-0.019***	0.198***	0.006	Yes	Maybe
	Median response time★	0.005	0.208***	0.062***	0.003	Yes	Yes
	Linguistic style matching	-0.031***	-0.029***	0.054***	0.026	Yes	Maybe
	Topic matching	0.002	-0.018***	-0.204***	0.003	No	
Control User history Demographics	(M) age	0.013***	0.009***	0.008**	-0.059*		
	(M) avg. hearts given★	0.0273***	0.116***	-0.010**	0.011		
	(M) avg. hearts received★	0.138***	-0.053***	0.013***	-0.033		
	(M) # of past conv.★	1.478***	0.251***	-0.008*	-0.076		
	(M) # of upvotes given★	-0.061***	0.042***	-0.047***	-0.026		
	(M) Prior mood				0.023***		
	(L) age	-0.010***	0.085***	0.024***	-0.022		
	(L) avg. hearts received★	-0.016**	0.087***	0.026***	-0.023		
	(L) # of past conv.★	-0.117***	0.186**	-0.005	0.023		
	(L) badge count★	0.044***	0.058***	0.038***	-0.010		
	(L) # of upvotes given★	-0.037***	0.052***	0.012***	-0.024		

Table 4: Coefficients of predictor variables for four different regression models utilizing unique success measures. Rating is reported from the prediction (stage 2) step of the Heckman model. Positive coefficients are colored blue and negative coefficients are colored red for explanatory variables. Log-transformed variables are noted with superscript★. The significance for coefficients are '*' for P < .0001, '**' for P < 0.001, and '*' for P < 0.01. Prior Work cites references used in the operationalization of the predictor. Agree refers to consistency across construct levels in our comparison.**

a standard linear regression in the stage 2 phase. The coefficients for ordinal models represent the log odds of moving one interval up the ordinal scale. Our findings suggest that prior reported predictors of counseling success are not consistent across construct levels but are consistent within the conversational outcome level.

5.1.1 Retention. Retention on the platform is negatively associated with the total number of words exchanged in a session (p < 2e-16), the member words ratio (p < 5e-6), listener self-disclosure (p < 2e-16), member self-disclosure (p < 2e-16), and linguistic style matching (p < 2e-16). Since 90% of the members in our dataset stay on 7 Cups after their first session, this result suggests that predictors of conversation success may have an inverse relationship with the need to participate in more chats on 7 Cups. Our results align with those of [41], who report that users turn to mental health platforms during times of need. We also replicate similar findings to those of Yang et al. [78], who noted that self-disclosure decreases commitment to a community.

5.1.2 Follow-up. Follow-up with the same listener is positively impacted by the total number of words (p < 2e-16), listener self-disclosure (p < 2e-16), and median response time (p < 2e-16). Follow-up shows a negative relationship with member words ratio (p < 2e-16), member self-disclosure (p < 4e-8), linguistic style matching (p < 1e-14), and topic matching (p < 0.0002). We expect higher amounts of member words ratio and member self-disclosure to represent information exchange during support provision [79], so lower amounts of it reducing the chance of follow-up suggests a problematic conversation. The negative impact of linguistic style matching and topic matching aligns with previous findings that more successful peer counselors change the topic or flow of conversation to progress conversations to important topics [3, 83]. 10% of members in our dataset follow up with their listener after the initial session, which suggests that this outcome measure may track long-term relationships built between member and listener in cases of an unfinished initial conversation.

5.1.3 Rating. A higher rating is positively associated with the total number of words (p < 2e-16), listener self-disclosure (p < 2e-16), member self-disclosure (p < 2e-16), the median response time (p <2 e-16), and linguistic style matching (p < 2e-16). Rating is negatively associated with member words ratio (p < 2e-16) and topic matching (p < 2e-16). Rating is the most consistent with predictors of effective counseling with prior literature. Unlike follow-up, member self-disclosure and linguistic style matching are associated with high ratings in this model. This suggests that rating is an appropriate

attitudinal measure of a member's satisfaction as a conversation reaches conclusion.

5.1.4 Mood. Member mood is positively associated with the total number of words ($p < 0.038$) and member words ratio ($p < 6e\text{-}6$), but is the least associated with prior predictors of listener performance among our outcomes. Our findings replicate [41] that individual mood does not change much with effective peer counseling, but contrasts with those of [3]. We discuss the differences in methodology between these two studies in 7.1.

5.1.5 Follow-up vs. Rating. Comparing within construct levels, predictors with the same directionality for both outcomes are total words, member words ratio, median response time, and topic matching. Our results for total words and member words ratio are consistent with findings by Althoff et al. [3], who found that successful supporters have longer message lengths and control the flow of conversation better than unsuccessful ones. Our results for topic matching are consistent with those of Zhang et al. [83], who found that lower amounts of topic matching suggested that a supporter is better at controlling the flow of conversation. In contrast, we found that longer response times lead to a higher likelihood of follow-up and a higher rating, unlike Saha and Sharma's [61] claim that faster response times lead to more engagement within TalkLife threads. These differences may be due to community or channel differences between platforms as TalkLife's interactions are forum-based.

Interestingly, member self-disclosure and linguistic style matching change signs within conversation outcomes, negatively correlating with follow-up but positively correlating with rating. The conflicting directionality in the two predictors suggests that follow-up and rating are related but distinct outcomes. Since prior work has suggested that more self-disclosure leads to more support provision [64, 79] and that more linguistic style matching in both dyadic [29] and group [64] conversations represent alignment between participants in a conversation, our findings show that follow-up sessions occur on 7 Cups when a conversation has not reached closure. A member who has not yet had time to self-disclose information and align with their listener in their first session is more likely to follow up after an idle period.

5.2 Robustness Checks

To rule out bias introduced by our choice of models for self-report data, we use model triangulation to check for robustness for the Heckman selection model and our mood model.

The Heckman model handles missing ratings by estimating a Gaussian variable in the selection step (stage 1) and using it for linear regression in the prediction step (stage 2). One drawback to this model is that it cannot handle missing data in control variables, so we were unable to use a member's prior ratings to control for individual differences in rating reports. We validate the results of our Heckman model using an ordinal regression model with a member's average rating given across all sessions prior to the current session added as an additional control variable. Observations for members that did not have at least one prior rating in addition to a session rating were dropped. This resulted in a 126k session subset which was then modeled using ordinal regression. Results showed agreement between the Heckman model and the ordinal regression model. All explanatory variables in the ordinal regression were significant and had the same directionality of coefficients as the Heckman model.

Post-session mood was modeled using pre-session mood as a control variable, which reduced the dataset size from 14,434 to 4,839 data points. To check that the change in sample size does not impact model coefficients, we ran two comparison models without pre-session mood as a control variable. The first model used the 14,434 data points from members who reported post-session mood but not pre-session mood. The second model used the same 4,839 observations from members who reported both pre-session and post-session mood scores, but with the pre-session mood control variable removed from modeling. Results showed that member words ratio, member self-disclosure, and linguistic style matching are significant variables that impact mood for both of these models. This suggests that the lack of significant variables impacting our mood outcome is not due to changes in the sample size.

6 DISCUSSION

The most important finding in our work is that alternative ways of measuring counseling success used in prior research are not strongly correlated with each other and show different patterns of association with conversational features that others have hypothesized to influence counseling success in online mental health platforms. For **RQ1**, results reinforced our hypothesis that outcome triangulation provided novel insight into interactions on OMHPs by revealing tensions in desirable outcomes previously noted in the literature. For **RQ2**, we find that the directionality of previously reported predictors of counseling success are mostly consistent within construct levels but not across them.

Retention has a negative relationship with almost all predictors of effective counseling previously reported in the peer counseling literature. It also shows a weak negative correlation with conversational outcomes, echoing [78]'s findings that support provision inside a conversation may lead to seekers leaving a community. Although the weak relationship may suggest individual conversations may not strongly influence a member's decision to continue chatting with other listeners on 7 Cups, our findings reinforce reports that users leave platforms when their needs are fulfilled [47, 80].

Both dyad-level outcomes were mostly consistent with predictors but showed nuanced differences. Rating is the most consistent outcome in relation to prior literature on predictors of effective counseling. Contrary to rating, less member self-disclosure and linguistic style matching correspond with an increased likelihood of follow-up. A novel insight from this difference is that some conversations have an idle period, yet are likely to continue if members are not given the opportunity to speak about their problems and receive feedback from a listener. Combined, high ratings and low follow-ups may signal effective single-session counseling on 7 Cups.

General mood shows little relationship with prior predictors. Outcomes that track the impact of a single conversation on an individual member need to be cautiously adopted when analyzing the impact of conversations on individuals. Our results replicate those of Kushner and Sharma [41], who found little change in mood following conversations on TalkLife. Based on the positive relationship between member words ratio and mood, it is possible that counseling expertise does not correlate with positive moods, but

simply chatting with someone does. However, our findings do not necessarily disagree with those of [3] as our methodologies were different. We did not leverage group-level aggregation (Section 3) and our reporting 24-hour reporting window for mood has limitations in terms of temporal causality (Section 4.3). Lastly, the lack of interaction between predictors and mood scores may also be due to issues with sampling or questionnaire administration on 7 Cups.

In summary, our results demonstrate the value of systematic outcome triangulation across construct levels and operationalization methods. Similar to previously reported findings on other platforms [47, 78, 80], we hypothesize that members who have their peer counseling needs met leave 7 Cups based on the negative correlation between retention and both conversational outcomes. There is also a small segment of 7 Cups users who continue a conversation with the same counselor if their needs are not met after an idle period. Rating, an attitudinal metric, becomes a key signal of attitude toward support received when interpreting the behavioral metrics of retention and follow-up. One nuance in interpreting the relationship between retention and follow-up is that seekers on therapy platforms always have the option of finding new supporters to discuss their problems with. This design may lead to trade-offs between community and conversation engagement.

Based on the above, we argue that 7 Cups provides an on-demand service similar to single-session therapy (SST), a therapy delivery method in which the aim is to maximize the efficacy of the first, and sometimes only, session with a walk-in therapist [24, 37]. Yip et al. [82] reported similar findings from an online text-based counseling service for youth called Open Up, noting that 23.6% of 81,654 sessions on the platform came from users that only accessed the service a single time. We provide an even larger dataset suggesting that SST may naturally arise in OMHPs. The lack of significant individual mood change is no longer surprising in light of the single-session perspective as extratherapeutic circumstances are a large factor in the effectiveness of SST as a service [17, 65].

7 IMPLICATIONS

7.1 Interpreting Prior Outcomes

Our findings re-emphasize the importance of individual context in understanding community-level outcomes for OMHPs. In this study, retention was replicated from prior social computing literature as representing a form of engagement with the community beyond individual conversations. Despite differences in the design of online therapy platforms and online health communities, we found evidence on 7 Cups that seekers leave when their needs are fulfilled. This aligns with models of user engagement with technology [34] and disease journeys [47, 80]. Both theories suggest that long-term participants in topical communities may need persistent support compared to seekers with short-term needs. On 7 Cups, peer counseling topics span both short-term and long-term problems. Our retention results may be reflecting the high frequency of topics such as self-improvement, dating, or issues with parents that may be more immediate in terms of care compared to chronic disease diagnoses.

Within the conversation level, the operationalization of follow-up sheds light on unfinished conversations over multiple bursts of time whereas ratings suggest satisfaction with conversations. The predictors of an incomplete conversation on 7 Cups, member self-disclosure and linguistic style matching, are consistent with prior work suggesting that effective counselors on crisis hotlines are better at moving conversations to a close than less effective ones [3, 83]. While these may appear incompatible with findings by Sharma et al. [63] that seeker engagement within discussion threads increases with more mutual discourse between seekers and supporters on TalkLife and Reddit, our findings are compatible if we consider that engagement in a conversation is necessary for successfully concluding a peer counseling session. That is, we want to see high engagement within a session, but the end goal may not be continued long-term engagement itself. Since we did not investigate in-depth the differences between these perspectives, triangulation of these two outcomes provides a direction that future researchers and developers of OMHPs can build off of to study single-session versus long-term seeker-provider relationships.

The value of mood as a proxy outcome for mental health status remains unclear. Our study examined seeker-provider pairs in private chats and found that individual mood is not likely to change. In contrast, Althoff et al. [3] were able to identify effective crisis counselors that improved seeker mood outcomes in a single text-based counseling session. We argue that these findings are not necessarily incompatible. When seeker outcomes are grouped by provider performance, models are now comparing group means similar to controlled trials [3, 20, 61]. Future research on individual outcomes should distinguish and compare their level of aggregation between groups and individuals using experimental design-based approaches for causal inference of observational data to guard against regression to the mean artifacts [32]. Furthermore, the sensitivity of mood scores as an instrument may impact generalizability across platforms. The properties of mood reports from crisis intervention platforms, where seekers may be experiencing intense emotions, may differ from those of reports from 7 Cups, a platform where seekers can have conversations about a variety of topics.

Two implications for individual-level metrics arise out of applying the SST framework to online therapy platforms. Prior literature on walk-in counseling clinics has suggested that SST was most effective for those who had mild severity of illness, motivation to receive counseling, and strong social support for clinical symptoms or reported coping mechanisms [37]. Without access to contextual knowledge of 7 Cups members who respond to questionnaires, it is difficult to know if we should expect a change in individual self-report data such as mood. In addition, those who take questionnaires multiple times on a platform may a priori be individuals who need more involved counseling, thus violating the assumptions of SST. Since our definition of a session allows variable session lengths, extratherapeutic factors may impact reported mood score directly before and after a session if a session lasts for an extended period of time. Caution is necessary for generalizing individual self-report data across platforms due to the potential impact of user illness journeys on measurement.

7.2 Therapy Platform Metrics

If the goal of a counseling platform is to provide support to people in need, leaving the platform is a natural consequence of having their needs met. To increase retention, platforms can leverage user

journeys that capture common pathways to motivate users to continue participating in the community when they are most likely to drop off the platform [16]. For example, seekers looking for conversation in a moment of panic may need an immediate match with a supporter while others who want to discuss long-term mental health diagnoses may benefit from slower but more personalized matchmaking [51]. We found that some members on 7 Cups would frequently start new chats, other members engaged in frequent follow-ups with the same listeners, and some showed both behaviors. A population of long-term users on 7 Cups likely has different needs than individuals seeking timely counseling, and future work can look into how to better support this long-term subset of users through the segmentation of user groups.

Platforms could also develop separate measures of retention for support seekers and support providers in light of the complex interactions that occur between community roles. For example, on 7 Cups, 10% of private conversations are between two listeners. In general, little is known about the transition from seeker to supporter, which may be a form of retention for peer counseling platforms. [60] emphasized user acquisition as a valuable metric to triangulate with retention and engagement. While not explored in this study, the acquisition of supporters, their retention, and their engagement with seekers and other supporters could constitute a community of practice that can be studied and nurtured.

7.3 Peer Counselor Training

Prior research on peer counseling online has tended to focus on measuring seeker outcomes without addressing other desirable measures of dyadic interaction. While no definitive operationalized measures exist for alliance in psychotherapy research [26], rating and follow-up can be paired to train peer counselors. Rating can serve as a signal of support provision, while follow-up may suggest that a conversation has not yet reached a point of satisfaction. Since volunteers lose motivation without feedback that lets them improve their skills [67, 76], user dashboards or additional badge-like features that capture these metrics could give peer supporters more control over how they interpret their effectiveness in supporting seekers who have diverse short-term and long-term needs.

Specifically for 7 Cups, Yao et al. [81] noted that listeners struggle with understanding the impact of their conversations due to a lack of feedback, which in turn can lead to poor mental well-being. For example, listeners may not be aware that it is common for members to stay on a platform to chat with other listeners and that members may pursue multiple conversations to find a better relationship fit [10]. To promote listener well-being, listeners can be informed that drop off may not be a sign of a poor conversation, but instead reflect quality when paired with measures of listener metrics such as rating score. 7 Cups can train listeners to understand their role as single-session counselors or relationship builders while minimizing the connection between a single conversation and individual mood.

8 LIMITATIONS AND FUTURE WORK

Our methodology and analysis focused on correlations across metrics rather than the causal impact of particular predictors on specific outcomes. Future work can examine true effect sizes and predictive models for each individual outcome in this study to help platforms identify the best predictors of key success outcomes. In addition, our choice of using a therapy platform, 7 Cups of Tea, as a research site may limit our analysis to platforms of this type. We urge practitioners to use our findings with caution when applying them to mental health sub-communities on social media or topical platforms such as Breastcancer.org. Design choices in the development of these platforms such as public and private channels, the use of conversation threads versus chats, and illness-specific user needs may change our understanding of what outcomes are important.

Future work can help validate the usefulness of our triangulation methodology by examining multiple outcomes using a similar distinction between construct levels and methods of measurement. In this study, follow-up and rating showed opposite relationships with member self-disclosure and linguistic style matching, which suggests that these two conversational outcomes may track different constructs despite sharing the same relationships with all other predictor variables. There is also room to expand the construct levels to include turn-based metrics. In this study, hearts were used as a proxy control variable for participation in a conversation on 7 Cups but were not used as an outcome measure. It is unclear what a turn-level metric means within the context of peer counseling, unlike in social media where hearts or likes represent positive sentiment and shallow engagement with a post [54]. Although [61] has investigated turn-level predictors of online counseling satisfaction, turn-level outcomes do not appear to be understood in the context of online peer counseling.

9 CONCLUSION

In this study, we examined two research questions around whether triangulating across multiple outcomes provides novel insights for finding counseling success indicators and whether previously reported predictor variables of counseling success track multiple outcome metrics consistently. We analyze demographic, linguistic, and topical features from one-on-one conversations on 7 Cups of Tea with four outcome measures simultaneously to answer these questions. Our findings suggest that community retention and conversational outcomes are relatively independent, follow-up and rating capture two complementary measures of conversation progress, and mood outcomes show little relationship with proposed predictors of counseling success. To the best of our knowledge, this paper is the first to systematically triangulate four outcome measures from the community, conversation, and individual levels to examine effective peer counseling on therapy platforms. Our work shows that research on successful outcomes in peer counseling benefits from a systematic approach to operationalization and measurement that prior work in the literature has not always been fully clear in defining. Based on our findings, we raise questions and discuss future directions for interdisciplinary research on OMHPs.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their helpful comments, members of the SALT Lab at Georgia Tech and Stanford, and 7 Cups for their partnership in this project. This work is funded in part by NSF grants IIS-2247357 and IIS-2112633.

REFERENCES

[1] Adrian Aguilera, Stephen M Schueller, and Yan Leykin. 2015. Daily mood ratings via text message as a proxy for clinic based depression assessment. *Journal of affective disorders* 175 (2015), 471–474.

[2] Felwah Alqahtani and Rita Orji. 2020. Insights from user reviews to improve mental health apps. *Health informatics journal* 26, 3 (2020), 2042–2066.

[3] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics* 4 (2016), 463–476. https://doi.org/10.1162/tacl_a_00111

[4] Mario Alvarez-Jimenez, Sarah Bendall, Reeva Lederman, Greg Wadley, Gina Chinnery, Sonya Vargas, M Larkin, Eoin Killackey, PD McGorry, and JF Gleeson. 2013. On the HORYZON: moderated online social therapy for long-term recovery in first episode psychosis. *Schizophrenia research* 143, 1 (2013), 143–149.

[5] Elske Ammenwerth, Carola Iller, and Ulrich Mansmann. 2003. Can evaluation studies benefit from triangulation? A case study. *International journal of medical informatics* 70, 2-3 (2003), 237–248.

[6] Jannis Androutsopoulos. 2017. Online data collection. In *Data collection in sociolinguistics*. Routledge, 233–244.

[7] JinYeong Bak, Chin-Yew Lin, and Alice Oh. 2014. Self-disclosure topic model for classifying and analyzing Twitter conversations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1986–1996. https://doi.org/10.3115/v1/D14-1213

[8] Jin Yeong Bak, Suin Kim, and Alice Oh. 2012. Self-Disclosure and Relationship Strength in Twitter Conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2* (Jeju Island, Korea) *(ACL '12)*. Association for Computational Linguistics, USA, 60–64.

[9] Javier A Bargas-Avila and Kasper Hornbæk. 2011. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2689–2698.

[10] Amit Baumel. 2015. Online emotional support delivered by trained volunteers: users' satisfaction and their perception of the service compared to psychotherapy. *Journal of Mental Health* 24 (2015), 313 – 320.

[11] Abir K Bekhet and Jaclene A Zauszniewski. 2012. Methodological triangulation: An approach to understanding data. *Nurse researcher* (2012).

[12] Thomas Berger. 2017. The therapeutic alliance in internet interventions: A narrative review and suggestions for future research. *Psychotherapy research* 27, 5 (2017), 511–524.

[13] Jeffrey L Binder. 1993. Is it time to improve psychotherapy training? *Clinical Psychology Review* 13, 4 (1993), 301–318.

[14] Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. 2014. Identifying emotional and informational support in online health communities. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*. 827–836.

[15] James F Boswell and Louis G Castonguay. 2007. Psychotherapy training: Suggestions for core ingredients and future research. *Psychotherapy: Theory, Research, Practice, Training* 44, 4 (2007), 378.

[16] Vincent Bremer, Philip I Chow, Burkhardt Funk, Frances P Thorndike, Lee M Ritterband, et al. 2020. Developing a process for the analysis of user journeys and the prediction of dropout in digital health interventions: machine learning approach. *Journal of Medical Internet Research* 22, 10 (2020), e17738.

[17] Alistair Campbell and Samantha Hemsley. 2009. Outcome Rating Scale and Session Rating Scale in psychological practice: Clinical utility of ultra-brief measures. *Clinical Psychologist* 13, 1 (2009), 1–9.

[18] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine* 3, 1 (2020), 1–11.

[19] Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1171–1184.

[20] Prerna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E. Palacios, Derek Richards, and Anja Thieme. 2020. *Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376341

[21] Joe Curran, Glenys D Parry, Gillian E Hardy, Jennifer Darling, Ann-Marie Mason, and Eleni Chambers. 2019. How does therapy harm? A model of adverse process using task analysis in the meta-synthesis of service users' experience. *Frontiers in psychology* 10 (2019), 347.

[22] Norman K Denzin. 2017. *The research act: A theoretical introduction to sociological methods*. Routledge.

[23] Louise Doyle, Anne-Marie Brady, and Gobnait Byrne. 2009. An overview of mixed methods research. *Journal of research in nursing* 14, 2 (2009), 175–185.

[24] Windy Dryden. 2018. *Single-session therapy (SST): 100 key points and techniques*. Routledge.

[25] Sean Ellis and Morgan Brown. 2017. *Hacking growth: how today's fastest-growing companies drive breakout success*. Currency.

[26] Rachel Elvins and Jonathan Green. 2008. The conceptualization and measurement of therapeutic alliance: An empirical review. *Clinical psychology review* 28, 7 (2008), 1167–1187.

[27] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–16.

[28] Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4647–4657. https://doi.org/10.1145/2858036.2858535

[29] Amy L. Gonzales, Jeffrey T. Hancock, and James W. Pennebaker. 2010. Language Style Matching as a Predictor of Social Dynamics in Small Groups. *Communication Research* 37, 1 (2010), 3–19. https://doi.org/10.1177/0093650209351468 arXiv:https://doi.org/10.1177/0093650209351468

[30] Madelyn S Gould, Anthony Pisani, Carlos Gallo, Ashkan Ertefaie, Donald Harrington, Caroline Kelberman, and Shannon Green. 2022. Crisis text-line interventions: Evaluation of texters' perceptions of effectiveness. *Suicide and Life-Threatening Behavior* (2022).

[31] Robert B. Grady. 1994. Successfully applying software metrics. *Computer* 27, 9 (1994), 18–25.

[32] Gemma Hammerton and Marcus R Munafò. 2021. Causal inference with observational data: the need for triangulation of evidence. *Psychological medicine* 51, 4 (2021), 563–578.

[33] James J. Heckman. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47, 1 (1979), 153–161. http://www.jstor.org/stable/1912352

[34] Kasper Hornbæk and Morten Hertzum. 2017. Technology acceptance and user experience: A review of the experiential component in HCI. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 5 (2017), 1–30.

[35] Adam O Horvath and Lester Luborsky. 1993. The role of the therapeutic alliance in psychotherapy. *Journal of consulting and clinical psychology* 61, 4 (1993), 561.

[36] Nan Hu, Jie Zhang, and Paul A Pavlou. 2009. Overcoming the J-shaped distribution of product reviews. *Commun. ACM* 52, 10 (2009), 144–147.

[37] Peter Hymmen, Carol A Stalker, and Cheryl-Anne Cait. 2013. The case for single-session therapy: Does the empirical evidence support the increased prevalence of this service delivery model? *Journal of Mental Health* 22, 1 (2013), 60–71.

[38] R Burke Johnson, Anthony J Onwuegbuzie, and Lisa A Turner. 2007. Toward a definition of mixed methods research. *Journal of mixed methods research* 1, 2 (2007), 112–133.

[39] Allen C Johnston, James L Worrell, Paul M Di Gangi, and Molly Wasko. 2013. Online health communities: an assessment of the influence of participation on patient empowerment outcomes. *Information Technology & People* (2013).

[40] Robert Jonsson. 2012. When does Heckman's two-step procedure for censored data work and when does it not? *Statistical Papers* 53, 1 (2012), 33–49.

[41] Taisa Kushner and Amit Sharma. 2020. Bursts of Activity: Temporal Patterns of Help-Seeking and Support in Online Mental Health Forums. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 2906–2912. https://doi.org/10.1145/3366423.3380056

[42] Effie Lai-Chong Law, Paul Van Schaik, and Virpi Roto. 2014. Attitudes towards user experience (UX) measurement. *International Journal of Human-Computer Studies* 72, 6 (2014), 526–541.

[43] Reeva Lederman, Greg Wadley, John Gleeson, Sarah Bendall, and Mario Álvarez-Jiménez. 2014. Moderated online social therapy: Designing and evaluating technology for mental health. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 1 (2014), 1–26.

[44] Alessandro Liberati, Douglas G Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C Gøtzsche, John PA Ioannidis, Mike Clarke, Philip J Devereaux, Jos Kleijnen, and David Moher. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology* 62, 10 (2009), e1–e34.

[45] Haiwei Ma, C Estelle Smith, Lu He, Saumik Narayanan, Robert A Giaquinto, Roni Evans, Linda Hanson, and Svetlana Yarosh. 2017. Write for life: Persisting in online health communities through expressive writing and social support. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–24.

[46] Wendy E Mackay and Anne-Laure Fayard. 1997. HCI, natural science and design: a framework for triangulation across disciplines. In *Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques*. 223–234.

[47] Michael Massimi, Jackie Bender, Holly O. Witteman, and Osman Hassan Ahmed. 2014. Life transitions and online health communities: reflecting on

adoption, use, and disengagement. In *Proc. CSCW 2014* (proc. cscw 2014 ed.). https://www.microsoft.com/en-us/research/publication/life-transitions-and-online-health-communities-reflecting-on-adoption-use-and-disengagement/

[48] Mark Matthews and Gavin Doherty. 2011. In the mood: engaging teenagers in psychotherapy using mobile phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2947–2956.

[49] J. A. Naslund, K. A. Aschbrenner, L. A. Marsch, and S. J. Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and Psychiatric Sciences* 25, 2 (2016), 113–122. https://doi.org/10.1017/S2045796015001067

[50] Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska De Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics* 42, 3 (2016), 537–593.

[51] Kathleen O'Leary, Arpita Bhattacharya, Sean A. Munson, Jacob O. Wobbrock, and Wanda Pratt. 2017. Design Opportunities for Mental Health Peer Support Technologies. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1470–1484. https://doi.org/10.1145/2998181.2998349

[52] Lawrence A Palinkas, Sarah M Horwitz, Patricia Chamberlain, Michael S Hurlburt, and John Landsverk. 2011. Mixed-methods designs in mental health services research: a review. *Psychiatric Services* 62, 3 (2011), 255–263.

[53] Amber Paukert, Brian Stagner, and Kerry Hope. 2004. The assessment of active listening skills in helpline volunteers. *Stress, Trauma, and Crisis* 7, 1 (2004), 61–76.

[54] Kay Peters, Yubo Chen, Andreas M Kaplan, Björn Ognibeni, and Koen Pauwels. 2013. Social media metrics—A framework and guidelines for managing social media. *Journal of interactive marketing* 27, 4 (2013), 281–298.

[55] Ingrid Pettersson, Florian Lachner, Anna-Katharina Frison, Andreas Riener, and Andreas Butz. 2018. A Bermuda triangle? A Review of method application and triangulation in user experience evaluation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–16.

[56] John Powell and Aileen Clarke. 2007. Investigating internet use by mental health service users: interview study. *Studies in health technology and informatics* 129, 2 (2007), 1112.

[57] Yada Pruksachatkun, Sachin Pendse, and Amit Sharma. 2019. Moments of Change: Analyzing Peer-Based Cognitive Support in Online Mental Health Forums. In *2019 CHI Conference on Human Factors in Computing Systems*. ACM CHI, ACM. https://www.microsoft.com/en-us/research/publication/moments-of-change-analyzing-peer-based-cognitive-support-in-online-mental-health-forums/

[58] Wullianallur Raghupathi and Viju Raghupathi. 2013. An overview of health analytics. *J Health Med Informat* 4, 132 (2013), 2.

[59] Fred Reichheld. 2011. *The ultimate question 2.0 (revised and expanded edition): How net promoter companies thrive in a customer-driven world.* Harvard Business Review Press.

[60] Kerry Rodden, Hilary Hutchinson, and Xin Fu. 2010. Measuring the user experience on a large scale: user-centered metrics for web applications. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2395–2398.

[61] Koustuv Saha and Amit Sharma. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 590–601. https://ojs.aaai.org/index.php/ICWSM/article/view/7326

[62] Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert Kraut, and Diyi Yang. 2022. Modeling Motivational Interviewing Strategies On An Online Peer-to-Peer Counseling Platform. *Proceedings of the ACM on Human-Computer Interaction 6* (2022). https://doi.org/10.1145/3555640

[63] Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 2020. Engagement patterns of peer-to-peer interactions on mental health platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 614–625.

[64] Eva Sharma and Munmun De Choudhury. 2018. *Mental Health Support and Its Relationship to Linguistic Accommodation in Online Communities*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174215

[65] Arnold Slive. 2008. Walk-in single session therapy. *Journal of systemic therapies* 27, 4 (2008), 1–4.

[66] Michael Jae Song, John Ward, Fiona Choi, Mohammadali Nikoo, Anastasia Frank, Farhud Shams, Katarina Tabi, Daniel Vigo, and Michael Krausz. 2018. A process evaluation of a web-based mental health portal (WalkAlong) using google analytics. *JMIR mental health* 5, 3 (2018), e8594.

[67] Frederick Sundram, Thanikknath Corattur, Christine Dong, and Kelly Zhong. 2018. Motivations, expectations and experiences in being a mental health helplines volunteer. *International journal of environmental research and public health* 15, 10 (2018), 2123.

[68] Keith Tudor and David Murphy. 2021. Online therapies and the person-centered approach. , 283–285 pages.

[69] Arnold POS Vermeeren, Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. 2010. User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries*. 521–530.

[70] Tatiana A. Vlahovic, Yi-Chia Wang, Robert E. Kraut, and John M. Levine. 2014. Support Matching and Satisfaction in an Online Breast Cancer Support Community. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1625–1634. https://doi.org/10.1145/2556288.2557108

[71] Yafei Wang, John Yen, and David Reitter. 2015. Pragmatic alignment on social support type in health forum conversations. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*. 9–18.

[72] Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling Self-Disclosure in Social Networking Sites. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work &; Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 74–85. https://doi.org/10.1145/2818048.2820010

[73] Yi-Chia Wang, Robert Kraut, and John M. Levine. 2012. To Stay or Leave? The Relationship of Emotional and Informational Support to Commitment in Online Health Support Groups. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 833–842. https://doi.org/10.1145/2145204.2145329

[74] Yi-Chia Wang, Robert E Kraut, and John M Levine. 2015. Eliciting and Receiving Online Support: Using Computer-Aided Content Analysis to Examine the Dynamics of Online Social Support. *J Med Internet Res* 17, 4 (20 Apr 2015), e99. https://doi.org/10.2196/jmir.3558

[75] Myrna M Weissman, Helen Verdeli, Marc J Gameroff, Sarah E Bledsoe, Kathryn Betts, Laura Mufson, Heidi Fitterling, and Priya Wickramaratne. 2006. National survey of psychotherapy training in psychiatry, psychology, and social work. *Archives of general psychiatry* 63, 8 (2006), 925–934.

[76] Galit Ventura Yanay and Niza Yanay. 2008. The decline of motivation?: From commitment to dropping out of volunteering. *Nonprofit management and Leadership* 19, 1 (2008), 65–78.

[77] Diyi Yang, Robert Kraut, and John M. Levine. 2017. Commitment of Newcomers and Old-Timers to Online Health Support Communities. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6363–6375. https://doi.org/10.1145/3025453.3026008

[78] Diyi Yang, Z. Yao, and R. Kraut. 2017. Self-Disclosure and Channel Difference in Online Health Support Groups. *Proceedings of the ... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media* 2017 (2017), 704–707.

[79] Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. The Channel Matters: Self-Disclosure, Reciprocity and Social Support in Online Cancer Support Groups. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300261

[80] Zheng Yao, Diyi Yang, John M Levine, Carissa A Low, Tenbroeck Smith, Haiyi Zhu, and Robert E Kraut. 2021. Join, Stay or Go? A Closer Look at Members' Life Cycles in Online Health Communities. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–22.

[81] Zheng Yao, Haiyi Zhu, and Robert E. Kraut. 2022. Learning to Become a Volunteer Counselor: Lessons from a Peer-to-Peer Mental Health Community. (2022). CSCW '22 pre-print.

[82] Paul Siu Fai Yip, Wai-Leung Chan, Christian S Chan, Lihong He, Yucan Xu, Evangeline Chan, Yui Chi Chau, Qijin Cheng, Siu-Hung Cheng, Florence Cheung, et al. 2021. The opportunities and challenges of the first three years of Open Up, an online text-based counselling service for youth and young adults. *International journal of environmental research and public health* 18, 24 (2021), 13194.

[83] Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019. Finding Your Voice: The Linguistic Development of Mental Health Counselors. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

[84] Renwen Zhang, Jennifer Nicholas, Ashley A Knapp, Andrea K Graham, Elizabeth Gray, Mary J Kwasny, Madhu Reddy, David C Mohr, et al. 2019. Clinically meaningful use of mental health apps and its effects on depression: mixed methods study. *Journal of Medical Internet Research* 21, 12 (2019), e15644.

A MULTICOLLINEARITY TEST

Table 5 reveals the relationships between predictor variables using the Variance Inflation Factor (VIF) as a measure of multicollinearity. All variables are mean normalized before computing the VIF and variables with superscript ★ are log transformed to reduce skew. Results show that some of our variables have more covariance than others, suggesting that some measures track similar constructs or may have underlying interaction factors. For example, listener badge count (VIF = 4.966) was included as a control for engagement with 7 Cups training and listener number of past conversations (VIF = 5.377) was included as a control for experience, but they may both be tracking time spent on the platform.

B TOPICAL AND HECKMAN STAGE 1 COEFFICIENTS

Table 6 lists the coefficients of the topical control variables for all models and the selection step (stage 1) of the Heckman model. The Heckman model leverages user experience and demographics control variables in addition to topical controls.

Features	VIF
Total Words★	4.093
Member Words / Total	1.548
Listener's Self Disclosure	1.221
Member's Self Disclosure	1.168
Response Time★	1.036
Linguistic Style Matching	1.493
Topic Matching	1.713
Member age	1.009
Member avg. hearts given★	2.016
Member # of past conversations★	2.147
Member hearts received★	1.062
Member # of forum upvotes given★	1.145
Listener age	1.018
Listener avg. hearts received★	1.681
Listener badge count★	4.966
Listener # of forum upvotes given★	1.548
Listener # of past conversations★	5.377
Romantic_Relationship★	1.850
Dating★	1.552
Pandemic★	1.443
Self_Improvement★	1.197
Suicide★	1.597
Depression★	1.880
Parents★	2.690
Anxiety★	2.507
Family★	2.595
Stress★	2.573
Lonely★	1.489
Overwhelming★	1.321
Sexuality★	1.724
LGBTQ★	1.083
Intimacy★	1.887
Home★	1.512
Dissociative_Identity★	1.462
Health★	1.265

Table 5: Variation Inflation Factor for all independent features. All variables are mean normalized before computing the VIF and variables with superscript★ are long transformed to reduce skew

Categories	Features	Retention	Follow-up	Rating	Mood
Control User history Demographics	(M) age			0.008**	
	(M) avg. hearts given★			-0.004201***	
	(M) avg. hearts received★			0.144***	
	(M) # of past conversations★			-0.130***	
	(M) # of forum upvotes given★			-0.038***	
	(M) Prior mood			0.080***	
	(L) age			0.002	
	(L) avg. hearts received★			0.009***	
	(L) # of past conversations★			0.074***	
	(L) badge count★			0.013***	
	(L) # of forum upvotes given★			-0.098***	
Topics	Romantic Relationship★	0.013***	-0.033***	0.014***	-0.053
	Dating★	0.025***	-0.052***	0.026***	0.000
	Pandemic★	0.015***	0.012***	0.041***	0.041
	Self-improvement★	0.018***	-0.035***	0.065***	0.002
	Suicide★	-0.004	-0.025***	0.051***	-0.084**
	Depression★	-0.012**	-0.042***	0.020***	-0.021
	Parents★	0.006	-0.004	0.006**	-0.078
	Anxiety★	0.003	0.016***	0.024***	-0.0129***
	Family★	-0.006	0.025***	0.016***	0.012
	Stress★	-0.016***	0.024***	0.034***	0.085*
	Lonely★	-0.021***	-0.022***	0.034***	0.004
	Overwhelming★	0.004	-0.053	0.072	0.017
	Sexuality★	0.016***	-0.032***	-0.028***	0.063
	LGBTQ★	0.005	0.003	0.019	0.013
	Intimacy★	-0.021***	-0.036***	0.003	-0.056
	Home★	-0.014***	-0.019***	0.031***	0.009
	Dissociate Identity★	-0.005	-0.011**	0.013***	-0.090***
	Health★	-0.015***	-0.005	0.015***	0.024

Table 6: Coefficients for topical control variables for each model and the selection step (stage 1) of the Heckman model.