The Effect of Moderation on Online Mental Health Conversations

David Wadden Tal August Qisheng Li Tim Althoff
Paul G. Allen School of Computer Science & Engineering
University of Washington, Seattle, WA
{dwadden, taugust, liqs, althoff}@cs.washington.edu

Abstract

Many people struggling with mental health issues are unable to access adequate care due to high costs and a shortage of mental health professionals, leading to a global mental health crisis. Online mental health communities can help mitigate this crisis by offering a scalable, easily accessible alternative to in-person sessions with therapists or support groups. However, people seeking emotional or psychological support online may be especially vulnerable to the kinds of antisocial behavior that sometimes occur in online discussions. Moderation can improve online discourse quality, but we lack an understanding of its effects on online mental health conversations. In this work, we leveraged a natural experiment, occurring across 200,000 messages from 7,000 conversations hosted on a mental health mobile application, to evaluate the effects of moderation on online mental health discussions. We found that participation in group mental health discussions led to improvements in psychological perspective, and that these improvements were larger in moderated conversations. The presence of a moderator increased user engagement, encouraged users to discuss negative emotions more candidly, and dramatically reduced bad behavior among chat participants. Moderation also encouraged stronger linguistic coordination, which is indicative of trust building. In addition, moderators who remained active in conversations were especially successful in keeping conversations on topic. Our findings suggest that moderation can serve as a valuable tool to improve the efficacy and safety of online mental health conversations. Based on these findings, we discuss implications and trade-offs involved in designing effective online spaces for mental health support.

1 Introduction

Over 400 million people globally struggle with mental health challenges, with approximately 300 million experiencing depression (WHO 2018b). Depression leads to economic costs totalling more than $100 billion annually in the United States alone (Twenge et al. 2019). Rates of serious psychological distress – including suicidal ideation and suicide attempts – have increased 71% in adolescents and young adults since 2005 (Twenge et al. 2019). Although psychotherapy and social support can be effective treatments (Wampold and Imel 2015; WHO 2018a), vulnerable individuals often have limited access to therapy and counseling (Bose et al. 2018).

Instead, more and more people are turning to online mental health communities to express emotions, share stigmatized experiences, and receive helpful information (Eysenbach et al. 2004). These communities offer an accessible way for users to connect to a large network of peers experiencing similar challenges. Participants unable to access other treatment options can find social support and relief through these conversations (De Choudhury and De 2014; Sharma and De Choudhury 2018; Naslund et al. 2016). Recently, social support networks have begun to offer a more personalized experience by matching people sharing similar struggles in live, private conversations for support (Althoff, Clark, and Leskovec 2016).

While online mental health communities can provide a valuable setting for giving and receiving support, the quality of support provided by peers is less well-characterized. Can conversation participants temporarily assume the role of a psychological counselor to assist those in serious distress? In addition, the often unrestricted and anonymous environment of online discussions can become a platform for antisocial behavior, such as online abuse or harassment (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015; Zhang et al. 2018). Are these concerns relevant in the setting of an app designed expressly for mental health discussion? Perhaps users of this platform are more thoughtful and considerate than the average forum participant. On the other hand, if bad behavior is an issue, moderation has been shown to be effective tool to combat undesirable behavior in online discussions (Seering et al. 2019; Matias 2019; Lampe et al. 2014; Seo 2007). But little is known about the effectiveness of moderation in the context of mental health applications. Do moderators need to be highly *involved* to keep users safe? Or does simply the knowledge that a moderator is *present* influence behavior without active intervention? Furthermore, what roles do moderators assume in mental health discussions? Are they mostly discipline-keepers, or do they also act as counselors and facilitators?

In this work, we investigated how moderation affected online mental health conversations by identifying a natural experiment (DiNardo 2016) occurring in the chat logs of a popular online mental health mobile application. The user community for this application generated an extensive chat history of roughly 200,000 messages, constituting the largest dataset of its kind. After roughly two years of hosting unmoderated conversations, the app platform switched to moderated conversations, assigning a trained modera-

tor to supervise and contribute to every conversation. We leveraged this platform change as a natural experiment, in which the assignment to moderated conversations suddenly changed while user population and application design remained largely identical. By comparing the linguistic attributes of the unmoderated and moderated conversations, we were able to extract data-driven insights about the effects of moderation on the civility, supportiveness, and outcomes of mental health discussions. To ensure the validity of these findings, we performed a number of additional experiments to confirm that our results were attributable to the switch in moderation status, and were not artifacts caused by other factors such as the time frame of data collection or the number of conversation participants. In addition, we took advantage of naturally occurring variation in the degree of moderator activity – i.e. how frequently moderators post in conversations – to disentangle whether effects were simply due to moderator *presence* or truly their active *involvement.* Finally, we compared the behavior of unmoderated users to the behavior of moderators to determine whether users ever engage in counseling-like behavior to support their peers in settings without a moderator present.

Our findings include:

1. Users were more engaged in the conversation when a moderator was present, writing twice as many messages on average (Section 4.1).
2. Moderators and users exhibited distinctive word usage patterns indicative of their roles in conversations. Moderators often acted as mental health counselors, emphasizing words involving perception and understanding others. Most users focused on explaining their concerns and struggles. They tended to disclose negative emotions more openly when a moderator was present. Interestingly, in the absence of a moderator, some users assumed a counseling-like role for other users (Section 4.2).
3. Conversations were much more likely to remain civil and free of toxic or harmful language when a moderator was present (Section 4.3).
4. Users coordinated more to one another in the presence of a moderator, suggesting stronger group cohesion and social support (Section 4.4).
5. Participants experienced positive perspective changes as measured by several psycholinguistic indicators. These improvements were larger on average in moderated conversations (Section 4.5).
6. *Actively* moderated conversations tended to stay more on-topic than passively moderated ones (Section 4.6).

In summary, moderation may positively impact online mental health conversations by encouraging more civil, on-topic conversations with higher user engagement. Based on these findings, we discuss design implications for online mental health platforms and how they may best utilize moderation to keep psychological support group conversations civil and beneficial for everyone (Section 5).

2 Related Work

Our study is motivated by work analyzing the effect of moderation on online discourse (Section 2.1), examining the risks and benefits of online mental health communities (Section 2.2), and understanding digital mental health interventions (Section 2.3).

2.1 The effect of moderation on discourse

Civility and politeness are important elements of online communities (Danescu-Niculescu-Mizil et al. 2013a; Burke and Kraut 2008), and can cause derailment of otherwise healthy discussions (Zhang et al. 2018). Moderation as a way of encouraging civility is ubiquitous in online discussions, where anonymity can invite harassment and anti-social behavior (Kraut and Resnick 2012). Moderators use a number of tools to maintain productive discussions, such as example-setting, posting rules on discussion threads, and restrictive bans when necessary (Seering et al. 2019; Matias 2019). Lampe et al. (2014) showed that moderation reduces uncivil and inflammatory rhetoric and encourages civil conversations in online discussions. In the context of classrooms, Seo (2007) found that students engage more actively and stay more on topic in peer-moderated online discussions than in unmoderated forums. Seering et al. (2019) interviewed 56 moderators across 3 major online platforms, finding that moderator involvement in a community can range from very active to only intervening when a serious transgression occurred. This also highlights different roles a moderator can fill: either as a facilitator, supporting conversations proactively, or solely keeping the peace of the online space (Seering et al. 2019). In this work, we conduct the first large-scale analysis on the role of moderators in online mental health conversations and examine the effect of moderator *activity* (in addition to presence) on discourse quality. In addition, this work represents the first study of the effects of moderation in the online mental health setting.

2.2 Online mental health communities

Online mental health communities are a valuable resource for peer-to-peer support (Eysenbach et al. 2004) due to their ease of accessibility, low cost, and the ability to remain anonymous. Work has shown that the ability to remain anonymous in computer mediated communication can increase self-disclosure (Joinson 2001), and Andalibi et al. (2016) showed this in mental health communities by exploring how people shared stigmatized experiences in online mental health communities on Reddit. They found that many users use throwaway accounts (i.e., accounts with no personal information) for sharing these experiences as a way of maintaining anonymity. Throwaway accounts also have been reported to share content with increased negativity and self-focus, and lower self-esteem, supporting their use for self-disclosure (Pavalanathan and De Choudhury 2015).

Many factors play a role in a user self-disclosing, including a desire to manage impressions, online group size, and tie strength (Wang, Burke, and Kraut 2016). Newman et al. (2011) interviewed people to see how they share mental health information online, and identified two competing tensions of wanting to share information concerning a health

issue and managing their own self-presentation. Work has also shown that mental health disclosure online can lead to positive outcomes such as emotional and informational support (De Choudhury and De 2014; Sharma and De Choudhury 2018), and positive cognitive change (Pruksachatkun, Pendse, and Sharma 2019).

Webb, Burns, and Collin (2008) and Lederman et al. (2014) describe the process of creating online mental health forums for adolescents with general mental health issues and psychosis, respectively. They identified roles for moderators including fostering a positive atmosphere, reporting crisis posts, setting boundaries, and encouraging users to practice cognitive and behavioral self-care skills. Zhang and Danescu-Niculescu-Mizil (2020) studied how mental health counselors in a crisis text line balanced responding empathetically and moving a conversation towards a resolution. Previous works have generally analyzed interactions occurring on public message boards or in crisis counseling. In this work, we instead leverage our naturally occurring dataset to understand how moderation affects private mental health discussions.

2.3 Digital mental health interventions

Researchers have explored more active interventions for supporting mental health. Shing et al. (2018) developed an automated suicide risk assessment model based on Reddit posts rated by clinicians. De Choudhury et al. (2013) built a classifier to predict the onset of depression from a users' social media posts. Saha et al. (2019) used social media data to evaluate the effects of psychiatric drugs, showing the feasibility of augmenting clinical studies with large-scale social media analyses of drugs' effects. Other recent work has characterized the ethical tensions of automated mental health interventions (Chancellor et al. 2019), identifying issues such as construct validity and bias, and data privacy.

3 Dataset

We describe the dynamics of the online mental health mobile application, the chat log dataset studied in this work, and basic preprocessing steps used to filter out low-quality data.

3.1 Dataset description

Our data consist of two sets of conversation logs, NoMod and Mod, collected from a mobile application that provides high-quality mental health discussion for distressed individuals. Approval to analyze the dataset was obtained from the Institutional Review Board at our institution.

Moderators were present for conversations in the Mod collection. The data collection timeline is shown in Figure 1. The life cycle of conversations in NoMod and Mod is similar and is shown in Figure 2.

In NoMod, a *starting user* wrote a post on a public topic page, creating a chat room. Other *joining users* could view the subject line of the post and were free to join the chat room to discuss the content of the post. The conversation ended when all users exited the chat. Starting and joining users are referred to collectively as *unmoderated users*.

In Mod, conversations took place in persistent "chat rooms" presided over by a single moderator. Moderators

Figure 1: Data collection timeline.

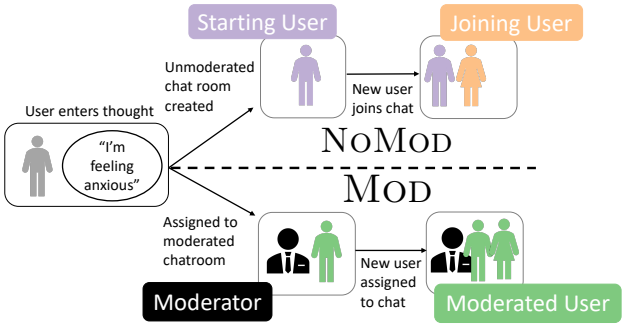


Figure 2: The life cycle of a conversation in NoMod (top) and Mod (bottom). Participants are colored based on their conversation roles. In NoMod, participants are either *starting users* (purple), or *joining users* (orange). In Mod, they are either *moderators* (black) or *moderated users* (green).

were undergraduate or graduate students pursuing degrees in psychology, who had completed training by the platform on peer support facilitation. They were paid $15 / hour. Upon opening the app, *moderated users* were asked to write a sentence describing the issue they were struggling with. They were assigned to a chat room discussing topics relevant to their issue (e.g. depression, anxiety, relationship problems). The assignment was performed automatically by the app using natural language processing techniques. Depending on the number of users on the app and the similarity of their issues, a single chat room could host multiple users with related concerns, or host a one-on-one conversation between the room's moderator and a user. All chat participants were aware of the presence and identity (i.e. chat username) of the moderator.

In both NoMod and Mod, users were assigned to conversations based on their interests and concerns. In NoMod this assignment was performed by the users themselves, while in Mod it was performed by an algorithm. The app designers report that Mod users were generally happy with their automatically-generated room assignments. Thus we have no reason to believe that this minor difference in assignment mechanism affected any of the outcomes studied.

3.2 Data preprocessing

Our unit of analysis in this work is a single conversation. Therefore, we segmented the long-running chat room logs in Mod into self-contained conversations by assuming that an interval of more than 15 minutes without a message indicated the end of a conversation[1].

We applied filters to remove very short, low-quality conversations. For NoMod, we kept all conversations with at least 2 participants and at least 10 messages long. For Mod,

[1]This cutoff was chosen in consultation with the app creators and validated with manual inspection.

	NoMod	Mod
N conversations	7,079	317
N users	7,344	558
N messages	183,570	14,985
Median messages per chat	15	30
Median tokens per message	8	8

Table 1: Summary statistics for filtered NoMod and Mod.

we kept all conversations with at least 1 user, one moderator, and 10 messages. For both datasets, we kept conversations with a median message length of at least 5 tokens. 34% of all messages in NoMod passed our filters, compared to 66% for Mod. All experiments were performed on the filtered data, for which Table 1 provides summary statistics. The NoMod dataset is roughly 13 times larger than Mod, and as a result the confidence intervals on Mod estimates in the following sections are larger. While the smaller number of Mod samples increases the *variance* in the parameter estimates for this dataset, there is no reason to believe it would *bias* our results in a particular direction. We address sources of potential bias in the next subsection.

3.3 Considering potential threats to validity

We considered multiple potential threats to the validity of our reported findings. If conditions are not randomized to participants, it is possible that any observed effects may be due to factors unrelated to the conditions of interest. We identified four potential factors and performed additional experiments to ensure that our findings were robust to them.

Time frame. As shown in Figure 1, the NoMod period ran for nearly 2 years, while Mod ran for roughly 9 months. We performed experiments on appropriately-chosen subsets of NoMod to confirm that our findings were not driven by seasonality effects or differences in the lengths of the NoMod and Mod data collection time frames. All experiments gave qualitatively similar results matching those reported in the main paper, which use all the available data (see Appendix A.1 for details).

Shifts in discussion topic. To establish that our findings were not an artifact of shifts in discussion topics between NoMod and Mod, we confirmed with the app creators that there were no deliberate changes in the topics discussed on the app. Most discussions centered around everyday challenges related to, for instance, anxiety or depression related to work or relationships. As an additional check, we fit an LDA topic model and confirmed that the topic distributions were quite similar for Mod and NoMod (see Appendix A.2). This suggests that the topics of discussion were comparable between the two conditions.

Number of conversation participants. Conversation size varied in both NoMod and Mod, and tended to be smaller for Mod (Section 4.1). To account for this, we initially stratified all analyses by the number of conversation participants. The results were qualitatively similar across different numbers of participants (see Appendix A.3 for an example). For ease of presentation, we show the unstratified results.

	Participant	Dataset	Q25	Q50	Q75
Messages	User	NoMod	3.4	5.3	8.5
	User	Mod	5.3	10.0	19.0
	Moderator	Mod	8.0	13.0	29.0
Tokens	User	NoMod	4.0	8.0	17.0
	User	Mod	3.0	7.0	14.0
	Moderator	Mod	5.0	9.0	15.0

Table 2: Messages per chat participant, and tokens per message.

First-time vs. repeat users. We confirmed in discussions with the app creators that, due to the 8-month gap between the end of the NoMod version of the app and the launch of Mod version, most Mod chat participants were new users unfamiliar with the older NoMod version of the app[2]. Thus, our findings are unlikely to be influenced by differences in behavior between new versus returning users.

3.4 Data anonymization

Extra precautions were taken to anonymize all conversations, posts, and discussion topics. Following best practices dealing with stories around abuse (Matthews et al. 2017), we anonymized all presented quotes and conversations by removing or deliberately changing any identifying information, including generalizing specific mentions of places or people. We also added and removed filler words or rephrased content with less unique word choices or phrases. This process was repeated independently by two of the authors. Due to the sensitive nature of the data, we are unable to provide screenshots of the mobile app as this could compromise the anonymity of participants. Instead, we are sharing several anonymized conversations based on the process described above (see Section 4.2).

4 Effects of moderation on conversation dynamics and outcomes

We study the word usage patterns of moderators and users (Section 4.2), and the effects of moderation on user engagement (Section 4.1) and civility (Section 4.3). We then explore how moderation promoted linguistic coordination, which is suggestive of supportiveness and group cohesion (Section 4.4), and facilitated positive changes in user perspective (Section 4.5). Finally, we examine whether moderation helped keep conversations more on topic (Section 4.6).

We initially performed all analyses stratified into three groups by moderator *activity*, measured by the fraction of messages sent by the moderator. We present stratified results in Section 4.6. For all other experiments, moderator activity level did not affect the outcome and we present unstratified results. See Appendix B for statistics on moderator activity.

4.1 User engagement and participation

We examined the engagement of users in moderated and unmoderated discussion. As shown in Table 2, users sent

[2]For the launch of Mod, the app creators sent a promotional email to users of the old NoMod platform; fewer than 10 recipients tried the new platform.

roughly twice as many messages per conversation in moderated discussions, indicating greater user engagement. Message length was similar across datasets.

70% of conversations in NoMod had more than two participants, compared to 40% in Mod. One explanation for this change might be that users lost interest in supporting their peers when a moderator was present, instead waiting for a 1-on-1 conversation. On the other hand, the difference could be a simple consequence of the constant availability of the moderators. We return to this issue in Section 4.4.

4.2 Word usage of moderators and users

We expected that moderators and users might play different roles in conversations, with moderators counseling and guiding users through their emotional difficulties, as has been shown in prior work (Zhang and Danescu-Niculescu-Mizil 2020). To evaluate this possibility, we examined whether these different conjectured roles manifested themselves through distinctive word usage patterns. Additionally, we examined whether some unmoderated users might behave in a counseling role to help peers in need of support.

Methods. We categorized the chat messages into four groups based on the roles described in Section 3.1: starting users, joining users, moderated users, and moderators.

We used the LIWC lexicon (Tausczik and Pennebaker 2010) to quantify differences in word usage among our four groups. LIWC defines 76 psycholinguistic categories – for instance "Sad" – and provides a list of words associated with each category. Figure 3 shows an example conversation annotated with LIWC categories. In Figure 3 and throughout, we denote quotations from moderators with *italicized text* and quotations from users with `typewriter text`. For each category, we computed relative changes in word usage for each group of users, compared to moderators. Word usage was measured by computing the fraction of words in each utterance belonging to each LIWC category (see Appendix C).

Results. Figure 4 shows usage of each LIWC category by each of the three user groups, relative to usage by moderators. Overall, we found that moderators and users showed speech patterns consistent with their conversation roles. Joining users also exhibited some counseling behavior similar to moderators. Validity checks confirmed that these results were robust to the factors described in section 3.3 (see Appendix A for some examples).

Users employed more self-focused language. Moderators expressed their interest in the well-being of users by making heavy use of language related to perception (Panels B, D, F), such as "*Oh, I see*" or "*I hear where you're coming from*".

In addition, moderators used the most second-person pronouns (Panels G, H), for instance "*That must make you feel lousy*"), and the fewest first-person pronouns (Panel N). Moderated users employed the most first person pronouns and the fewest second-person, issuing statements like "`Sometimes when I'm nervous, I'll procrastinate`" or "`I've always fought with my siblings`". These word usage patterns indicate that conversations were focused on the issues and concerns of

U: `I'm sad all the time and I don't want to do anything.`
M: *I'm sorry, that sounds like it must be really hard.*
M: *What are some activities that make you feel happy?*
U: `I like playing basketball with my friends and going for walks.`
M: *Have you tried going to play with your friends when they invite you?*
U: `I did, but the whole time I worried that my friends don't like me.`
M: *What about going for a walk?*
U: `Yeah, that's a good idea, it helps clear my head sometimes.`

I / me You Sad Assent / negation Perception Question

Figure 3: A conversation excerpt, with content altered to preserve anonymity.

users. Interestingly, a similar division occurs in unmoderated conversations: the pronoun usage of *joining* users was more similar to moderators, while the usage of *starting* users was more similar to moderated users. This suggests that counseling-type behavior occurred even in unmoderated conversations – albeit by untrained users.

Users experienced negative emotions. Moderated users expressed feelings of anxiety (Panel I; "`I worry that I'm unlovable`") and sadness (Panel P; "`I've been hurt and betrayed over and over`") more readily than both moderators and unmoderated users, suggesting that the presence of a moderator may have allowed conversation participants to speak more openly about difficult emotions.

All users uttered words related to anger (Panel J; "I hate my job") and death / suicide (Panel M) more frequently than moderators. Some death-related utterances were benign, like "`My job is killing me right now`", but others indicated serious distress, for instance "`I came close to killing myself`" or "`I wish I could die`".

In response to the range of negative emotions voiced by users, moderators employed positive, encouraging language (Panel E), for instance "*Physical activities like going for a walk are a good distraction from negative thoughts*", "*It's good that you have friends and family in your corner*".

Word meaning was influenced by context. Body-related (Panel L) and sex-related (Panel Q) words were employed by users for different purposes. In some cases, words related to body parts or physical processes were used literally to discuss physical discomfort or health challenges, such as "`I can't sleep at night because I can't stop thinking`" or "`I get bad headaches when I'm stressed`". Other times, users discussed body parts metaphorically to convey emotions: "`My heart is so lonely`" or "`I get in my own head and second-guess myself`". Bodily functions were also used as curse words: "`After all the sh*t, I still forgive them`" or "`That's crap`".

A similar phenomenon occurred for sex-related words.

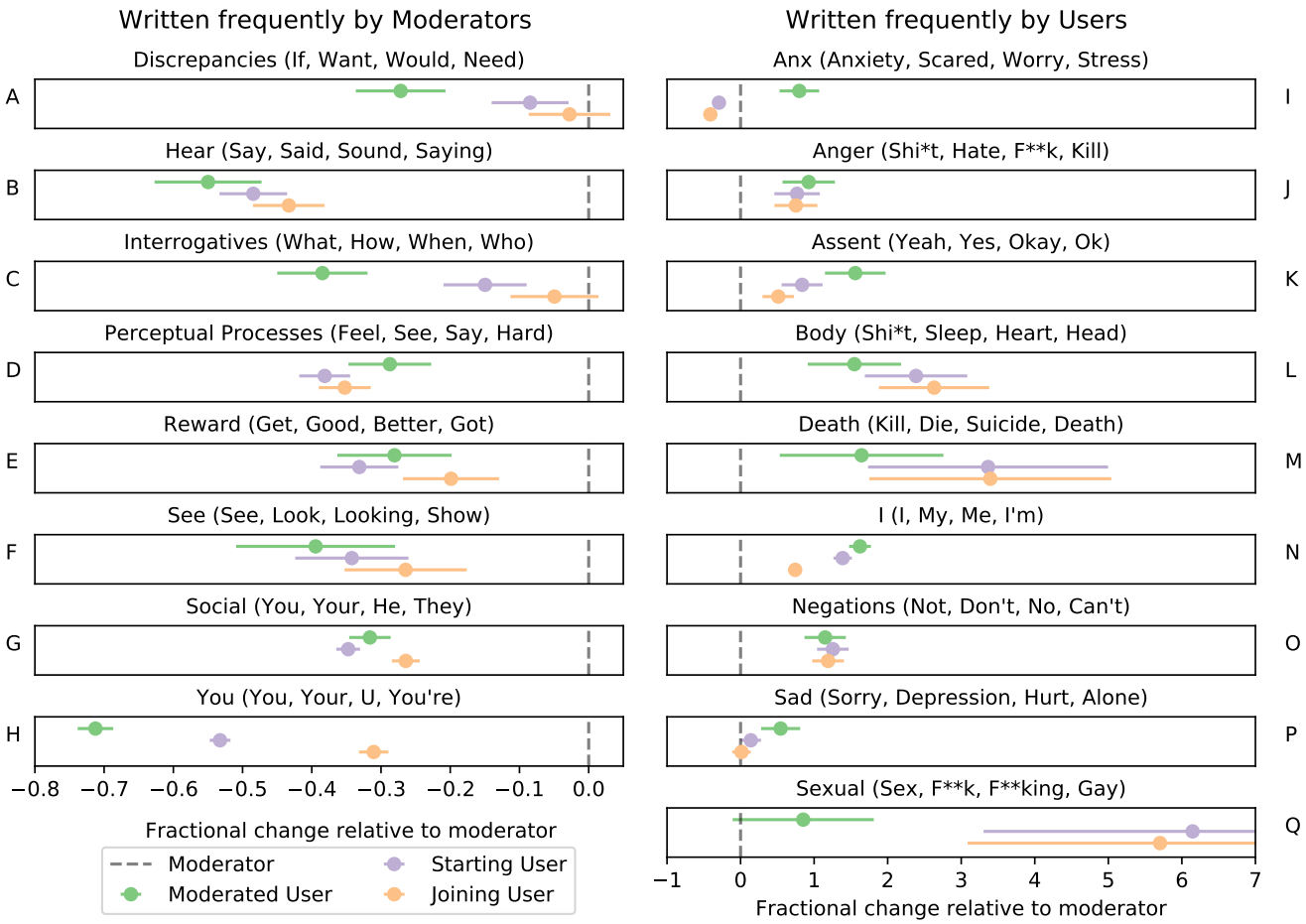


Figure 4: Word usage of different user groups. The x-axis shows the change in word usage of each LIWC category relative to moderator usage. For instance, a value of 1.2 indicates a 120% increase in word usage. Moderator usage is represented by dashed vertical lines. In this figure and throughout the paper, error bars are bootstrapped 95% confidence intervals. Any time the error bars between two groups do not overlap, the word usage between the groups is statistically significant at the 5% level.

Some users wanted to ask advice or discuss their sex lives: "`I just had sex for the first time and I'm not sure how I feel about it`" or "`It's hard to come out, because my family is religious`". Others used words concerning sex or sexuality as insults: "`f*ck you`" or "`That's gay`".

Unmoderated users employed body-related words slightly more frequently than moderated users and uttered sex-related words much more frequently. Much of this difference was due to the higher rates of profanity in unmoderated discussions, as explored in Section 4.3.

Moderators and joining users often acted as counselors. When moderators and users took part in a one-on-one conversation, they frequently followed a pattern of asking questions, making suggestions, and providing feedback. First, the moderator used interrogatives (Panel C) to ask questions and understand the user's situation, and the user responded with clarification. Then, the moderator made suggestions, often employing a LIWC category known as *discrepancies* (Panel A). Moderators utilized discrepancies to give advice or show interest, while allowing the user to disagree or express preference for an alternative. For instance a moderator might say, "*I'm happy to listen if you want to share your thoughts*" instead of "*tell me what you're thinking*". In response, users provided feedback in the form of either assent (Panel K), indicating that they agreed with the moderator's advice or assessment or negation (Panel O), indicating that the advice was not right for them. The conversation excerpt below depicts a moderator asking questions to understand a user's situation and offering advice and resources.

U: `I'm so annoyed. My boyfriend has been seeing other women, and when I confront him he lies.`

M: *So you've tried to talk with him? Does he get mad when you bring it up?*

M: *I would try to really communicate to him how hurtful it is to you. Does that make sense?*

U: `It's just so confusing. He can be so nice to me, but he gets mad whenever I ask him.`

M: *I'm with you. I would ask, too!*
M: *I can send you an article about healthy communication if you'd like?*
U: `Ya I would really appreciate that.`

Just as joining users exhibited different pronoun usage compared to starting users, they also also made slightly greater use of discrepancies and interrogatives. The following excerpt shows a discussion between a user U1 who started a chat with the subject "`I want to hurt myself`", and a joining user U2 who saw the subject line and offered help, acting like an informal counselor. We show messages from U2 in *italics* for readability.

U1: `I want to hurt myself.`
U2: *Please don't!*
U2: *Do you want to tell me about it?*
U2: *I also struggle with this. You need to find a way to focus on other things.*
U1: `I've tried. I wish I could follow that advice but I can't fight the urge.`
U2: *Maybe you could listen to some happy music, or draw?*
U1: `Yes, you're right I should do that.`
U2: *Have you told anyone how you feel?*
U1: `Well I have really great friends and family, but I don't want to burden them.`
U2: *You should tell them how you feel. They care about you and they will want to help.*

4.3 Conversation civility

Due to the potential vulnerability of mental health conversation participants, ensuring that discussions remain free of offensive, profane, and toxic language is a top priority. While previous research has shown that moderation can reduce profanity in online forums (Lampe et al. 2014), no prior work has examined whether moderation is necessary – or effective – in the setting of an application designed exclusively for mental health support.

Methods. We used Google's TensorFlow toxicity identifier[3] to examine the effect of moderation on conversation civility. The model identifies seven types of violent and abusive language: (1) identity attacks, (2) insults, (3) obscenity, (4) sexually explicit content, (5) threats, (6) toxicity, and (7) severe toxicity. We ran this identifier on all messages sent by users in MOD and NOMOD (removing moderator messages)[4]. For the unmoderated conversations, initial experiments did not reveal any differences in profanity between starting users and joining users; therefore, we analyzed all unmoderated messages together. Fortunately, there were no instances of severe toxicity in either dataset, so this category was not considered further.

Results. Figure 5 shows the percentage of MOD and NOMOD messages containing each category of offensive

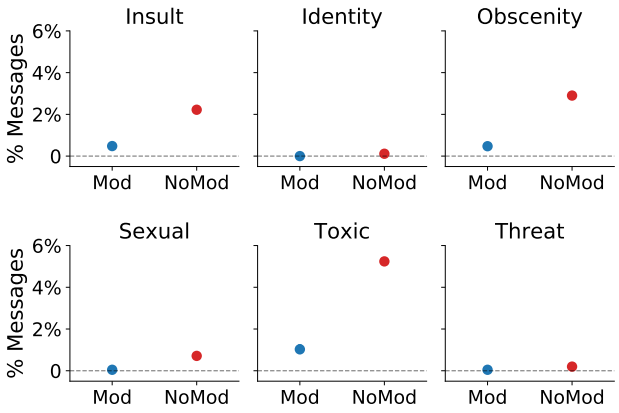


Figure 5: Percentage of user messages in each incivility category. Moderated conversations are almost totally free of toxic and harmful language. Error bars are computed but are too narrow to be visible.

language. Moderation reduces all forms of incivility to 1% of messages or lower, and reduces the occurrence of general toxic language five-fold – from 5.2% of messages to 1.0%. These results suggest that moderation is, indeed, highly effective at combating toxicity and incivility in online mental health settings. Moderators seem to have accomplished this without coming across as disciplinarians and driving users away, as users tended to stay longer in moderated chats compared to unmoderated ones (Section 4.1).

4.4 Coordination and trust-building

Given our results in Section 4.1, we wondered whether users might engage less with their peers when a moderator was present, preferring 1-on-1 counseling conversations instead. To evaluate this possibility, we analyzed the degree to which users in NOMOD and MOD coordinated linguistically toward one another. Past work suggests that higher levels of coordination are indicative of trust-building (Scissors, Gill, and Gergle 2008) and lead to improved group performance (Fusaroli et al. 2012) and social support (Sharma and De Choudhury 2018; Danescu-Niculescu-Mizil et al. 2013b).

Methods. We used the Cornell Convokit[5] to measure linguistic coordination among users in both datasets. Convokit measures linguistic coordination with words processed nonconsciously by listeners and unrelated to topic, such as articles, auxiliary verbs, and conjunctions (Danescu-Niculescu-Mizil et al. 2012).

Results. Surprisingly, the results in Figure 6 demonstrate that on average peer users in MOD actually coordinated significantly *more* toward one another than users in NOMOD. The higher levels of coordination in MOD conversations suggest that users did not lose interest in group discussion. Instead, moderation positively impacted conversations by encouraging coordination.

[3]https://github.com/tensorflow/tfjs-models/tree/master/toxicity

[4]To confirm that our findings were robust to the particular software toolkit used, we also made profanity predictions using the `profanity-check` Python library (https://github.com/vzhou842/profanity-check) and obtained similar results.

[5]http://convokit.cornell.edu/

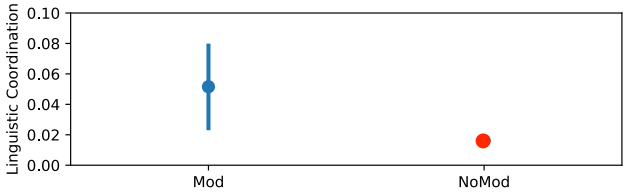


Figure 6: Linguistic coordination in NoMod and Mod. On average peer users in NoMod coordinated significantly less to one another (mean=0.01, med=0, sd=0.07) than users in Mod did to one another (mean=0.05, med=0.01, sd=0.13) ($U = 433419$, $p < 0.01$)

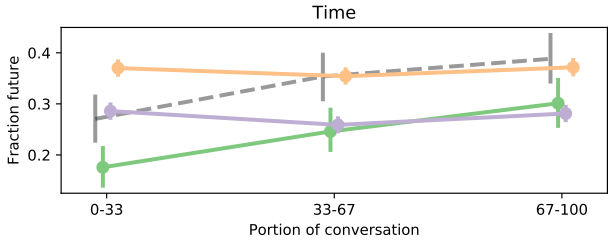
4.5 Positive change in perspective

Previous work has demonstrated that 1-on-1 online counseling conversations can facilitate positive changes in psychological perspective (Althoff, Clark, and Leskovec 2016), but it is not clear how a group setting could influence psychological perspectives differently. We explored whether the different user groups experienced perspective changes consistent with their roles from Section 4.2.

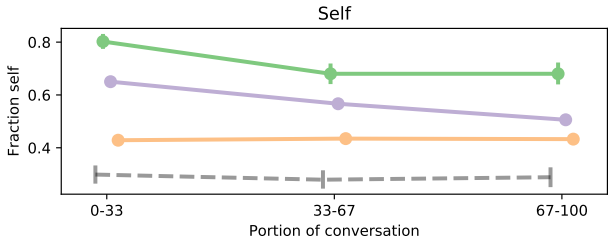
Methods. We leveraged the LIWC lexicons to compute three different measures of psychological perspective, following the approach in Althoff, Clark, and Leskovec (2016).

- **Time**: Psychological research has linked depression with excessive rumination about past events (Pyszczynski, Holt, and Greenberg 1987), while recent work examining 1-on-1 online counseling conversations has found associations between greater future-orientation and more positive conversation outcomes (Althoff, Clark, and Leskovec 2016). We computed the usage of LIWC future words as a fraction of all future and past-related words. Higher values indicate more focus on the future and less on the past, suggesting a more positive perspective.
- **Self**: Individuals experiencing depression can become preoccupied with their own thoughts and have difficulty engaging with others (Pyszczynski, Holt, and Greenberg 1987; Pyszczynski and Greenberg 1987). This tendency can be expressed through heavy use of first-person pronouns. We measured self-focus by computing the fraction of pronouns used by conversation participants that were first-person, as opposed to second or third person. Lower self-focus suggests a more positive perspective.
- **Sentiment**: To measure sentiment, we computed the usage of LIWC words related to positive emotion (PosEmo), as a fraction of the total number of words related to any emotion – positive or negative (NegEmo). Higher values suggest more positive sentiment.

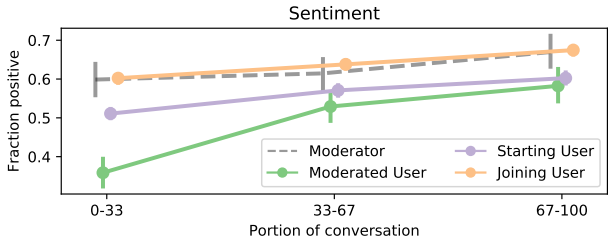
To measure how user perspective changed over the course of a conversation, we divided each conversation into thirds, and computed the perspective measures for each third. To ensure that we had enough data for each time period, we restricted our analysis to conversation with at least 20 messages sent by users. This left 2,541 conversations in NoMod and 219 in Mod. As in Section 4.2, we grouped messages into those sent by moderators, moderated users, starting users, and joining users. Moderator activity did not have a substantial effect on perspective, so all moderated users were analyzed together.



(a) Of the three user groups, moderated users show the largest increase in future focus, rising by 0.13 ± 0.06 from the first third to the final third of each conversation. Reported uncertainties are 2 standard errors of the mean.



(b) Moderated users and starting users show similar decreases in self-focus (-0.12 ± 0.04 and -0.14 ± 0.02, respectively).



(c) Moderated users show the largest sentiment increases (0.22 ± 0.05), followed by starting users (0.09 ± 0.02).

Figure 7: Change in perspective over time. Moderators and joining users exhibit counseling-type behavior. Users experience significant improvement in perspective.

Results. Figure 7 reveals important differences in perspective trajectory for the three user groups[6]. Joining users behaved very similarly to moderators. They showed positive perspective early on and maintained it throughout their discussions, consistent with their hypothesized role as counselors for distressed conversation starters (Section 4.2). Moderated users showed the most negative perspective initially, consistent with the observation that moderated users express negative emotion more freely (Section 4.2). Fortunately, they also enjoyed the largest overall improvement in perspective, and by the end of discussion their levels of sentiment and future-focus were nearly as high as those of unmoderated users. Their self-focus also decreased, but remained higher than the self-focus of other users.

This is unsurprising in one sense, since moderators

[6]Validity checks confirmed that these differences were robust to variations in time frame and chat room size (Appendix A).

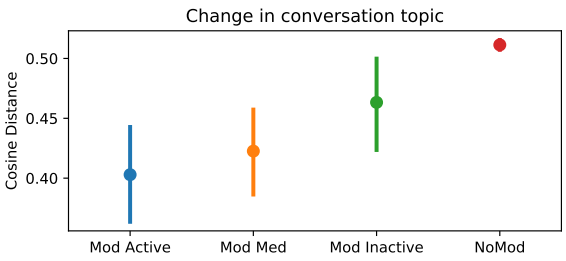


Figure 8: Mean cosine distances between first half and second half of with heavy moderator activity (mean = 0.40), medium activity (mean = 0.42), light activity (mean = 0.46) and with no moderation (mean = 0.51). The pairwise differences among the four groups of conversations are all statistically significant ($p < 0.05$) except for Mod Med compared to Mod Inactive ($p = 0.14$) or Mod Active ($p = 0.49$).

showed the lowest self-focus and likely continue eliciting first-person responses from users throughout their conversations. However, it is also surprising since moderated users coordinated more with one another than unmoderated users (see Section 4.4), suggesting that although users might focus more on themselves they also were focused on engaging with other users in moderated conversations. Starting users had initial perspectives between those of joining users and moderated users. They showed improved sentiment and a greater decrease in self-focus than moderated users, consistent with their conversation role.

4.6 Staying on topic

In an online classroom setting, previous work has demonstrated that conversations remained more on topic when discussions were moderated (Seo 2007). We explored whether this trend persisted in the more open-ended and personal conversations in our dataset, and further, whether *active* moderation was necessary to achieve an effect.

Methods. Motivated by work in topic segmentation (Hearst 1997), we measured the degree to which a conversation remained on-topic by splitting each conversation in half and computing the cosine distance between bag-of-words representations of the two conversation halves. A lower distance indicates that the conversation stayed on a single topic.

In this experiment, we found that the outcome was influenced by the degree of moderator activity. We present results with the moderated data stratified into three equal-sized collections of conversations. MOD Active contains the conversations with greatest moderator involvement, MOD Inactive with the least involvement, and MOD Med in between (see Appendix B for more details).

Results. As shown in Figure 8, we found that conversations stayed significantly more on topic in chats with a moderator present, in agreement with previous findings (Seo 2007). In addition, our results suggest that conversations with an active moderator may remain more on topic than those with an inactive moderator ($p < 0.05$).

5 Discussion

In this work, we performed the first large-scale quantitative analysis examining the effect of moderation on online mental health discussions, using the unmoderated and moderated conversation logs in our data set as control and treatment groups in a natural experiment. We found that:

- Moderation improved user engagement (Section 4.1).
- Moderators and users employed different language consistent with their conversation roles. In the absence of moderators, some users assumed a counseling role to support their peers (Section 4.2).
- Moderation improved conversation civility (Section 4.3).
- Moderation was associated with linguistic coordination, which was indicative of trust building and social support (Section 4.4).
- Users in moderated conversations experienced larger positive perspective changes (Section 4.5) and stayed more on topic (Section 4.6).

To support these conclusions, we performed a number of validity checks (Section 3.3 and Appendix A) and confirmed that:

- The time frames over which the data were collected did not impact our findings.
- The topics discussed were highly similar in both MOD and NOMOD.
- While the number of conversation participants varied, it did not have an effect on our findings.
- Both MOD and NOMOD were predominantly composed of new users and therefore the presence of new vs. repeat users was not significantly different.

Below we discuss some implications of our findings.

Should conversations be moderated? The results in this work are consistent with findings on moderation in other web domains (Matias 2019; Lampe et al. 2014), and indicate that moderation may be an effective approach to ensure that participants stay safe while participating in potentially challenging discussions about mental health. Compared to unmoderated or inactively moderated conversations, our results also suggest that *active* moderators may keep conversations more civil (Section 4.3) and on-topic (Section 4.6).

The role and limitations of peer support. Peer support can be helpful in encouraging personal connections and scaling social support (O'Leary et al. 2018). Indeed, we found ample evidence of *user-as-counselor* behavior in our data (Sections 4.2 and 4.5). Users often responded to peers experiencing crises like suicidal ideation by seeking to understand and offer support, sending messages like "*Please don't hurt yourself*" or "*Why would you want to kill yourself*?" However, there were exceptions. Even the best-intentioned users are not trained to handle mental health crises and risk responding to vulnerable individuals in ways that could do unintended harm. In addition, we observed that some individuals experiencing mental health crises were angry and hostile, and used profane and abusive language (Sections 4.2

and 4.3). We also uncovered some evidence of predatory behaviour in unmoderated chats (e.g., encouraging people to share social media accounts or pictures). Trained moderators are better suited than peers for dealing with acute and severe crises like suicidal ideation, hostility, and predatory behaviour, reducing the risk of harmful conversations.

Who should be a moderator? Moderators for a mental health support site could vary in their training and expertise – ranging from volunteers without psychological training to licensed social workers and psychologists. The undergraduate and graduate psychology students used as moderators in the application studied here occupied a middle ground.

The degree of moderator training required depends on the role that the moderator is expected to play in the conversations. On one extreme, a *moderator-as-supervisor* could simply supervise the conversation, discouraging bad behavior (Section 4.3) and stopping or appropriately escalating any discussions of self-harm. Based on our findings and previous work on moderation in public forums like Reddit (Matias 2019; Seering et al. 2019), it appears that *any* moderator could fill this role.

On the other extreme, a *moderator-as-counselor* would be expected to guide the conversation, identifying participants' issues and offering concrete suggestions (as in Section 4.2), and keeping the conversation on topic (Section 4.6). Through manual analysis of chat logs in our data set, we found that many moderators were in fact working in this capacity (as exemplified by the conversation excerpts from Section 4.2). Moderators serving as counselors would require appropriate training as determined by mental health professionals.

Limitations & Future Work. In this work we explored short-term mental health conversations taking place on one popular application platform, and found that moderation meaningfully improved discourse quality and improved the perspective of conversation participants. As in any observational study, we cannot be totally certain that the assignment of users to the NoMod or Mod group was independent of other factors affecting conversation outcome. However, our validity checks and conversations with the app creators provide good assurance that our findings are correctly attributed and are not spurious artifacts. In addition, we did not identify any obvious idiosyncrasies of the app which would prevent our findings from generalizing to other online mental health forums. This work represents a first attempt at understanding the effects of moderation in the online mental health setting, and we eagerly await the availability of additional datasets on which to validate the generalizability our findings.

In our analysis, we leveraged psycholinguistic tools in order to gain insights into the mental states of users. We found that many words included in LIWC categories had different meanings depending on context (Section 4.2). For instance, a user who remarks "this assignment is killing me" is having a bad day, but one who remarks "I'm thinking about killing myself" needs immediate help. An analysis that counts occurrences of the word "kill" cannot distinguish these cases. Followup research could leverage contextualized word embeddings (Peters et al. 2018; Devlin et al. 2019) to identify which mentions of potentially harmful words demand attention. In addition, future work could complement our linguistically-driven approach by collecting and analyzing direct ratings from users (Wang and Culotta 2019).

Finally, future work could examine the long-term effects of online mental health conversations. When do these discussions lead to long-term mental health improvements, and what are the key linguistic traits that turn momentary improvement into long-term progress?

6 Conclusion

We conducted a large-scale analysis examining the effect of moderation on online mental health discussions. We found that moderation improved civility, supportiveness, and coherence. Our findings suggest that moderated mental health support conversations could be a scalable tool to combat the ongoing mental health crisis and set the stage for deeper explorations into the impact of moderator expertise, moderation style, and moderator training on conversation outcomes.

Acknowledgements. This research was supported in part by NSF grant IIS-1901386, Bill & Melinda Gates Foundation (INV-004841), an Adobe Data Science Research Award, the Allen Institute Institute for Artificial Intelligence, and a Microsoft AI for Accessibility grant.

References

[Althoff, Clark, and Leskovec 2016] Althoff, T.; Clark, K.; and Leskovec, J. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *ACL*.

[Andalibi et al. 2016] Andalibi, N.; Haimson, O. L.; De Choudhury, M.; and Forte, A. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *2016 CHI*.

[Bose et al. 2018] Bose, J.; Hedden, S. L.; Lipari, R. N.; Park-Lee, E.; and Tice, P. 2018. Use and Mental Health Indicators in the United States: Results from the 2017 National Survey on Drug Use and Health Recommended Citation Substance Abuse and Mental Health Services Administration. Technical report.

[Burke and Kraut 2008] Burke, M., and Kraut, R. 2008. Mind your ps and qs: the impact of politeness and rudeness in online communities. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 281–284.

[Chancellor et al. 2019] Chancellor, S.; Birnbaum, M. L.; Caine, E. D.; Silenzio, V.; and De Choudhury, M. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *FAT**.

[Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015] Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial behavior in online discussion communities. In *Ninth ICWSM*.

[Danescu-Niculescu-Mizil et al. 2012] Danescu-Niculescu-Mizil, C.; Lee, L.; Pang, B.; and Kleinberg, J. 2012. Echoes of power: Language effects and power differences in social interaction. In *21st WWW*.

[Danescu-Niculescu-Mizil et al. 2013a] Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013a. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

[Danescu-Niculescu-Mizil et al. 2013b] Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013b. No

country for old members: User lifecycle and linguistic change in online communities. In *22nd WWW*.

[De Choudhury and De 2014] De Choudhury, M., and De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth ICWSM*.

[De Choudhury et al. 2013] De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *Seventh ICWSM*.

[Devlin et al. 2019] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *HLT-NAACL*.

[DiNardo 2016] DiNardo, J. 2016. *Natural Experiments and Quasi-Natural Experiments*. London: Palgrave Macmillan UK.

[Eysenbach et al. 2004] Eysenbach, G.; Powell, J.; Englesakis, M.; Rizo, C.; and Stern, A. 2004. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *The BMJ* 328(7449).

[Fusaroli et al. 2012] Fusaroli, R.; Bahrami, B.; Olsen, K.; Roepstorff, A.; Rees, G.; Frith, C.; and Tylén, K. 2012. Coming to terms: quantifying the benefits of linguistic coordination. *Psychol. Sci.* 23(8).

[Hearst 1997] Hearst, M. A. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*

[Joinson 2001] Joinson, A. N. 2001. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European journal of social psychology* 31(2):177–192.

[Kraut and Resnick 2012] Kraut, R. E., and Resnick, P. 2012. *Building successful online communities: Evidence-based social design*. MIT Press.

[Lampe et al. 2014] Lampe, C.; Zube, P.; Lee, J.; Park, C. H.; and Johnston, E. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly* 31(2).

[Lederman et al. 2014] Lederman, R.; Wadley, G.; Gleeson, J.; Bendall, S.; and Alvarez-Jimenez, M. 2014. Moderated online social therapy: Designing and evaluating technology for mental health. *CHI* 21.

[Matias 2019] Matias, J. N. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *NAS* 116(20).

[Matthews et al. 2017] Matthews, T.; O'Leary, K.; Turner, A.; Sleeper, M.; Woelfer, J. P.; Shelton, M.; Manthorne, C.; Churchill, E. F.; and Consolvo, S. 2017. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *2017 CHI*.

[Naslund et al. 2016] Naslund, J. A.; Aschbrenner, K. A.; Marsch, L. A.; and Bartels, S. J. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiol. Psychiatr. Sci.* 25(2).

[Newman et al. 2011] Newman, M. W.; Lauterbach, D.; Munson, S. A.; Resnick, P.; and Morris, M. E. 2011. It's not that i don't have problems, i'm just not putting them on facebook: Challenges and opportunities in using online social networks for health. In *ACM 2011 CSCW*.

[O'Leary et al. 2018] O'Leary, K.; Schueller, S. M.; Wobbrock, J. O.; and Pratt, W. 2018. "suddenly, we got to become therapists for each other": Designing peer support chats for mental health. In *2018 CHI*.

[Pavalanathan and De Choudhury 2015] Pavalanathan, U., and De Choudhury, M. 2015. Identity management and mental health discourse in social media. In *Proceedings of the 24th International Conference on World Wide Web*, 315–321.

[Peters et al. 2018] Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *HLT-NAACL*.

[Pruksachatkun, Pendse, and Sharma 2019] Pruksachatkun, Y.; Pendse, S. R.; and Sharma, A. 2019. Moments of change: Analyzing peer-based cognitive support in online mental health forums. In *2019 CHI*.

[Pyszczynski and Greenberg 1987] Pyszczynski, T., and Greenberg, J. 1987. Self-regulatory perseveration and the depressive self-focusing style: a self-awareness theory of reactive depression. *Psychol. Bull.* 102 1.

[Pyszczynski, Holt, and Greenberg 1987] Pyszczynski, T.; Holt, K.; and Greenberg, J. 1987. Depression, self-focused attention, and expectancies for positive and negative future life events for self and others. *J. Pers. Soc. Psychol.* 52 5.

[Saha et al. 2019] Saha, K.; Sugar, B.; Torous, J.; Abrahao, B.; Kiciman, E.; and De Choudhury, M. 2019. A social media study on the effects of psychiatric medication use. *ICWSM*.

[Scissors, Gill, and Gergle 2008] Scissors, L. E.; Gill, A. J.; and Gergle, D. 2008. Linguistic mimicry and trust in text-based cmc. In *2008 ACM CSCW*.

[Seering et al. 2019] Seering, J.; Wang, T.; Yoon, J.; and Kaufman, G. 2019. Moderator engagement and community development in the age of algorithms. *New Media Soc.*

[Seo 2007] Seo, K. K. 2007. Utilizing peer moderating in online discussions: Addressing the controversy between teacher moderation and nonmoderation. *Am. J. Distance Educ.* 21(1).

[Sharma and De Choudhury 2018] Sharma, E., and De Choudhury, M. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *2018 CHI*.

[Shing et al. 2018] Shing, H.-C.; Nair, S.; Zirikly, A.; Friedenberg, M.; Daumé III, H.; and Resnik, P. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Fifth Workshop on Computational Linguistics and Clinical Psychology*.

[Tausczik and Pennebaker 2010] Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1).

[Twenge et al. 2019] Twenge, J. M.; Cooper, A. B.; Joiner, T. E.; Duffy, M. E.; and Binau, S. G. 2019. Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005-2017. *J. Abnorm. Psychol.* 128 3.

[Wampold and Imel 2015] Wampold, B. E., and Imel, Z. E. 2015. *The great psychotherapy debate: The evidence for what makes psychotherapy work*. Routledge.

[Wang and Culotta 2019] Wang, Z., and Culotta, A. 2019. When do words matter? understanding the impact of lexical choice on audience perception using individual treatment effect estimation. *Computing Research Repository*.

[Wang, Burke, and Kraut 2016] Wang, Y.-C.; Burke, M.; and Kraut, R. 2016. Modeling self-disclosure in social networking sites. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW 16, 7485. New York, NY, USA: Association for Computing Machinery.

[Webb, Burns, and Collin 2008] Webb, M.; Burns, J.; and Collin, P. 2008. Providing online support for young people with mental health difficulties: challenges and opportunities explored. *J. Early Interv. Psychiatry* 2 2.

[WHO 2018a] 2018a. Depression, key facts. https://www.who.int/en/news-room/fact-sheets/detail/depression. Accessed: 2019-10-1.

[WHO 2018b] 2018b. Mental health, key facts. https://www.who.int/news-room/fact-sheets/detail/mental-disorders. Accessed: 2019-10-1.

[Zhang and Danescu-Niculescu-Mizil 2020] Zhang, J., and Danescu-Niculescu-Mizil, C. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards.

	1	2	3	4	5
Mod	0.24	0.18	0.14	0.14	0.29
NoMod	0.25	0.14	0.16	0.14	0.32

Table 3: Fraction of utterances assigned to each topic in a 5-topic LDA model.

[Zhang et al. 2018] Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *ACL*.

A Potential threats to validity

A.1 Time frame

To confirm that our results were robust to seasonality, we repeated our analyses using the *matched-seasonality* subset of NoMod shown in Figure 9, which runs for the same months as Mod. Similarly, to verify that the results were not simply an artifact of changing behavioral norms or user demographics, we re-ran all analyses using the *latest window* subset of NoMod which includes the final nine months of NoMod. The findings from the full dataset were robust to these perturbations. As one example, Figure 10 shows an excerpt of Figure 4, re-created using only the *matched-seasonality* subset of NoMod. The same trends are apparent here as in the full data.

A.2 Discussion topics

As a data-driven check that Mod and NoMod centered on similar discussion topics, we fit an LDA topic model on all conversations, and computed the fraction of messages in Mod and NoMod assigned to each topic. The results for a model fitted with 5 topics are shown in Table 3. While the topic distributions are not statistically indistinguishable[7], they are qualitatively quite similar; Topic 5 is the largest, followed by Topic 1, followed by 3 topics of roughly equal size. The same trends hold for a model with 10 topics.

A.3 Conversation size

We initially stratified our analyses by the number of conversation participants. We found that the same qualitative trends were present regardless of the number of participants, and we therefore collapsed the groups for our final analysis. As an example, Figure 11 shows the sentiment changes from Figure 7c, stratified into one-on-one conversations and conversations with at least 3 participants. While starting and joining users are more similar in one-on-one conversations, the same qualitative trends are apparent: user perspective improves over time, and these improvements are largest in moderated conversations.

[7] A χ^2 test rejects the null that Mod and NoMod have identical topic distributions with $p < 0.001$.

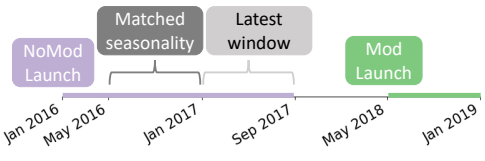


Figure 9: Time frames used for validity checks.

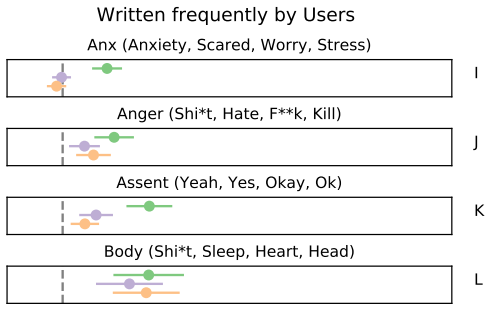
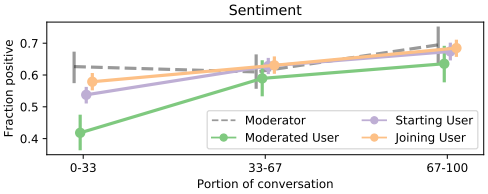
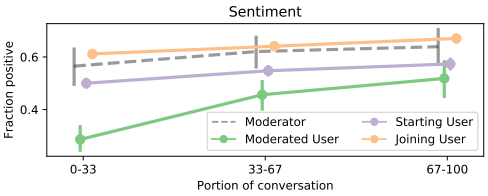


Figure 10: A replication of some rows from Figure 4, using only the *matched-seasonality* subset of NoMod.



(a) Sentiment change in one-on-one conversations.



(b) Sentiment change in multi-user conversations.

Figure 11: A replication of Figure 7c, stratifying by number of conversation participants.

B Moderator activity

We stratified moderated conversations into three groups based on moderator activity, as measured by the fraction of messages in the conversation sent by the moderator. We refer to these three groups of messages as Mod Active, Mod Med, and Mod Inactive. Mod Active contains the top third of conversations with the highest fraction of moderator messages, Mod Med has the middle third, and Mod Inactive the lower third.

The distribution of moderator activity is shown in Figure 12. Activity varied widely, from very frequent posting to virtual absence. Moderator activity had an effect on the experiments presented in Section 4.6, but not elsewhere.

C LIWC Word Usage

We describe the procedure used to compute the error bars on the relative differences in word usage shown in Figure 3. Let g index the user groups and k index the LIWC categories (e.g. "Sad"). Denote the number of messages for each user group as n_g, and the collection of messages for group g as $\{\mathbf{m}_{g,i}\}_{i=1}^{n_g}$, where message $\mathbf{m}_{g,i}$ is a sequence of tokens. For each LIWC category k, define $p_{g,i}^k$ to be the fraction of words in message i of user group g that match some word in LIWC category k. For instance, if the message were "This makes me mad and upset", and the LIWC category "anger" contained the words "mad, upset, angry", then

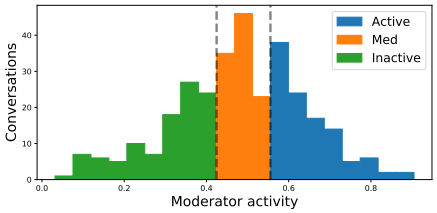


Figure 12: Distribution showing the fraction of messages in each conversation that were written by the moderator. Black dashed lines show 33rd and 67th quantiles.

$p_{g,i}^{\text{anger}} = 2/6$. Then let

$$\bar{p}_g^k = \frac{1}{n_g} \sum_{i=1}^{n_g} p_{g,i}^k, \tag{1}$$

the average fraction of words matching category k for messages from group g. Similarly, let $\bar{p}_m^k$ be the average word fraction for messages sent by moderators. Define the relative change in word usage for LIWC category k by user group g, relative to usage by moderators, as

$$\Delta_g^k = \frac{(\bar{p}_g^k - \bar{p}_m^k)}{\bar{p}_m^k} \tag{2}$$

Resample the user and moderator messages with replacement $n_{\text{boot}} = 1000$ times, compute bootstrapped $\{\tilde{\Delta}_{j,g}^k\}_{j=1}^{n_{\text{boot}}}$ on the resampled messages, and use the 2.5^{th} and 97.5^{th} quantiles as a 95% confidence interval for Δ_g^k.