Training TikTok creators in mental health communication benefits their audience, too: An analysis of TikTok user comments

Yuning Liu
PhD Candidate
Center for Health Communication
Harvard T.H. Chan School of Public Health
yuning_liu@g.harvard.edu

Matt Motta
Assistant Professor
Dept. of Health Law, Policy, & Management
Boston University School of Public Health
mmotta@bu.edu

Amanda Yarnell
Senior Director
Center for Health Communication
Lecturer
Department of Social and Behavioral Sciences
Harvard T.H. Chan School of Public Health
ayarnell@hsph.harvard.edu

Abstract

Background: Social media platforms like TikTok are pivotal for mental health communication (MHC), providing accessible, trusted content. Despite known benefits of creator training programs in improving content quality, it's uncertain if or how these enhancements affect people who watch that content. This study examines how training programs for TikTok creators influence TikTok users' mental health knowledge construction and intended behaviors by analyzing user comments on TikTok videos.

Method: We employed LLM-assisted content analysis to annotate TikTok video comments for mental health knowledge construction and behavioral intentions. We then analyzed the impact of a creator training program on users' mental health knowledge construction through within-subject field experiments. Additionally, we extended this method to measure how TikTok's Mental Health Awareness Month (MHAM) in May 2023 affected user mental health knowledge construction and behavioral intentions.

Results: LLM-assisted analysis showed that asynchronous toolkits significantly enhance mental health knowledge construction among video commenters, particularly among creators with high engagement, fewer followers, who are licensed mental health professionals, or who do not offer paid coaching services. MHAM increased mental health knowledge construction among users, though the audiences of the 7 institutions and 10 content creators that were spotlighted by TikTok during MHAM did not show this effect.

Conclusions: Asynchronous toolkits improve mental health knowledge construction among video viewers. Future training should focus on emerging creators and separate toolkit learning from conference training to enhance training efficacy. In addition, promoting hashtags such as #MentalHealthAwareness can drive meaningful discussions and knowledge construction in video comments, but there is an opportunity for platforms to refine their creator selection criteria to improve on-platform mental health knowledge construction.

Introduction

Social media platforms like TikTok are becoming essential tools for mental health communication (MHC), offering users access to relevant content that they increasingly consider a trusted source of health information. These platforms facilitate a wealth of mental health discussions, enabling users to acquire information, share personal stories and concerns, and find community support. Notably, the hashtag #MentalHealth on TikTok saw its views skyrocket from 25.3 billion in 2022 to nearly 44 billion in 2023, almost doubling within just one year (Basch et al., 2022). Meanwhile more and more people are seeking health information online. Data from the U.S. CDC's National Health Interview Survey for July–December 2022 indicates that 58.5% of adults used the internet for health or medical information (Wang, X., & Cohen, R. A., 2023). TikTok particularly resonates with younger demographics: 67% of U.S. adolescents aged 13-17 use the app, with 16% using it daily, and 62% of young adults aged 18-29 use it, which is significantly higher than the 33% usage rate among all U.S. adults (Emily A. Vogels et al., 2022; Jeffrey Gottfried, 2024). Moreover, the trust users place in the health content they find online is noteworthy. According to the KFF Health Misinformation Tracking Poll from April 2024, more than half of young adults aged 18-29 (53%) trust health information they find on TikTok (Alex Montero et al., 2024). Furthermore, younger adults are more likely than older adults to act on this information by consulting a doctor (19%) or seeking mental health treatment (26%) as a result of content viewed on the platform. The increasing ubiquity of TikTok use and the trust users have in health content on the app underscore TikTok's potentially pivotal role in enhancing online mental health communication.

Given TikTok's potential for enhancing MHC, research institutes, international organizations, and civil society groups have initiated new efforts to support creators who want to make mental health content. For example, Harvard Chan School’s Center for Health Communication, the World Health Organization (WHO), and YouTube Health have each launched programs for mental-health creators in recent years (Motta et al., 2024; YouTube, n.d.). These programs’ offerings vary but may include expert-vetted mental health information as well as training in mental health communication, evidence-based content practices, and social video production. For instance, YouTube Health's 2025 THE-IQ creator program will offer mental health creators virtual training and resources on mental health communication best practice (YouTube, n.d.). Meanwhile, organizations like the Mental Health Storytelling Coalition and the Center for Health Communication maintain publically available resource libraries for creators who want to make content about mental health topics or their own mental health experiences (USC, 2024).

As resources are increasingly allocated to creator training programs, an important question arises: What are the effects of these programs? Previous research suggests that it is possible to construct interventions that *influence the influencers* to promote evidence-based mental health content. For example, in a randomized controlled trial conducted on TikTok (Motta et al., 2024), we found that the provision of asynchronous educational resources promoting evidence-based mental health talking points (e.g., whether or not videos make reference to the idea that mental health issues can result from intergenerational trauma) can increase the prevalence of evidence-based mental health content on social media. This resulted in tens of thousands of additional views of evidence-based material. Ongoing qualitative assessments of content creators who participated in this study provide suggestive evidence that there may be an added and lasting benefit to adding in-person components to this training (Xu et al., 2024). Creators invited to participate at an in-person summit at the research site reported being highly motivated to make additional effort to include evidence-based mental health research in the content of videos that they produce. These studies suggest that creator training programs can improve the overall quality of MHC content produced on TikTok.

However, a significant research gap remains: While we know that creator training programs can enhance the quality of content produced, it is unclear whether these improvements lead to meaningful changes in audience behavior. According to models of behavioral change in health communication, the information individuals encounter on social media can alter both their perceptions of the threat posed by a specific mental health issue and their perceptions of the effectiveness of recommended behaviors (Janz & Becker, 1984). These perceptions, in turn, may influence how individuals respond behaviorally to the issue (Limbu et al., 2022). Consequently, we hypothesize that enhancements in MHC content on TikTok resulting from creator training programs may also influence users’ attitudes and intended behaviors concerning mental health. However, no previous studies have empirically tested this hypothesis. Therefore, this study aims to explore this unresolved issue.

To examine users' attitudes and behaviors regarding mental health, we analyze user comments on TikTok videos. User comments, as a form of user-generated content, provide a valuable lens for exploring user motives, expectations, and concerns in mental health communication (MHC). Comments offer real-time insights into user engagement with online MHC content, providing a more objective perspective compared to retrospective self-reports (Shahbaznezhad et al., 2021). Comments have been used to analyze product needs and uncover predictive patterns, and in the context of video-sharing platforms, they often feature socialization and idea sharing (H. Nguyen & Diederich, 2023). For example, in a study of comments on selected TikTok mental health videos, 66% of comments contained supportive or validating content, while 56% described other mental health issues or struggles (Basch et al., 2022). In this study, we use an analysis of TikTok video comments to examine user-level MHC features and mental-health related behaviors on TikTok.

Specifically, by analyzing user comments on TikTok, we aim to assess the impact of creator training programs across three key dimensions: user knowledge construction of mental health, mental health misinformation, and mental health behavioral intentions. First, we focus on knowledge construction because TikTok has proven to be a platform conducive to informal learning. Educators have integrated TikTok into learning processes, enhancing participant engagement and learning outcomes (Jacobs et al., 2022). A couple of studies have employed theoretical frameworks such as the Interaction Analysis Model (IAM) to study informal learning and interaction processes on TikTok (Haythornthwaite et al., 2018; H. Nguyen & Diederich, 2023) . IAM suggests that computer-mediated social learning unfolds as users (1) share ideas, (2) explore dissonance, (3) negotiate meanings, (4) test tentative knowledge syntheses, and (5) apply the newly constructed knowledge to new contexts, with each phase escalating in complexity (C. Gunawardena et al., 2016, 2016). IAM provides a valuable framework for understanding the informal learning processes of MHC on TikTok.

Second, given the pervasive challenge of health misinformation online and its relative under-examination compared to other health misinformation topics, we further investigate the prevalence of mental health misinformation on TikTok (V. C. Nguyen et al., 2024; Starvaggi et al., 2024). To do so we employ an expert-driven coding process to identify mental health misinformation in video comments.

Lastly, TikTok comments also illuminate various users' mental health behavioral intentions. For instance, TikTok facilitates the formation of online support networks for mental health. Previous research has shown that TikTok's community structure is permeable, enabling self-discovery and understanding not commonly found in traditional online communities (Milton et al., 2023). A systematic review identified three categories of social media use for health purposes by the general public, including tracking and sharing health status, exchanging social support in online communities, and seeking health-related information (Chen & Wang, 2021). Building on this work, we propose a five-question measure of mental health behavioral intentions derived from TikTok video comments, including seeking well-being information, sharing well-being related coping strategies, disclosing professional help received for well-being issues, self-disclosing mental health issues, and encouraging others to seek professional help.

The present study

By analyzing user comments on TikTok videos, the current study aims to explore how enhancements in MHC content on TikTok, stemming from <<institute name dropped for peer review>>'s creator training program, may influence users' construction of mental health knowledge, the prevalence of mental health misinformation in their comments, and their related behavioral intentions. Initially, we develop an approach utilizing LLM-assisted content analysis to annotate TikTok video comments, and then we analyze these annotations in an experimental study setting to examine the impact of a creator training program on user-level outcomes (Case 1). Subsequently, we apply this approach in a natural experiment setting to investigate the effects of TikTok's Mental Health Awareness Month (MHAM) (Case 2) and TikTok's MHAM spotlight on institutions and creators(Case 3) on the three dimensions of user comments, using the LLM tools developed in Case 1. This study seeks to address the following four research questions:
RQ1: What is the feasibility of developing LLM-assisted content analysis as text classifiers for evaluating online mental health communication using user comments on TikTok videos?

RQ2: What are the effects of the mental health creator training program on audience knowledge construction of mental health, mental health misinformation, and mental health behavioral intentions?
RQ3: What are the effects of TikTok's MHAM on these three user mental health outcomes?
RQ4: What are the effects of TikTok's MHAM spotlight on institutions and creators on these three user mental health outcomes?

Result

Developing the LLM-based comment classifier

Following the recruitment process of the randomized control trial documented in the previous study of our team, we identified 62 creators for the study. Using TikTok's Research Tools API, we successfully requested 188,169 comments from 1,882 videos created by 49 creators in March, April, and May 2023. Details of the data collection process are presented in Figure A1. Among the full sample of 1,882 videos, 54 TikTok videos created by mental health content creators (MHCCs) involved in the mental health creator training program were randomly sampled for content analysis, yielding 4,152 comments. Three research assistants performed content analysis on these comments to assess levels of knowledge construction, mental health misinformation, and mental health behavioral intentions (Table A1). Knowledge construction outcomes were categorized into six areas: personal reflection, expressing agreement or disagreement, asking for clarification, reinterpreting knowledge, applying knowledge to new areas, and a cumulative outcome labeled as general knowledge construction, indicating the comment contains at least one out of the six knowledge construction processes. Mental health behavioral intentions encompassed five categories: expressing appreciation to the creator, seeking well-being-related information, disclosing personal mental health issues, disclosing receipt of well-being-related professional services, and sharing well-being coping strategies. The three research assistants achieved an average inter-coder reliability of 0.81, with Gwet's AC of 0.80 (Table A2) across a random sample of 300 comments from the total 4,152 comments.

Using the comments coded in the content analysis, we performed character-level, word-level, and LLM-based sentence-level text augmentation to ensure a rich and balanced sample for LLM model fine-tuning. Table A3 shows that the augmented texts achieved a mean cosine similarity of 0.85 and a median cosine similarity of 0.9 compared to the original text, indicating satisfactory performance. With the augmented texts, we fine-tuned LLMs for 12 outcomes related to mental health knowledge construction and mental health behavioral intentions. We excluded the mental health misinformation category due to the rarity of positive cases in the content analysis, with less than 10 comments containing misinformation (Figure A2). We compared BERT-base and MentalBERT as the based model for fine tuning LLMs and found BERT-base outperformed MentalBERT; thus, we fine-tuned the models using BERT-base for the 12 outcomes with batch size of 32, 20 epochs, and a learning rate of 2e-5. Table A4 presents the performance of the fine-tuned LLMs, showing that each model achieved an average accuracy rate of over 90%. The fine-tuned LLMs are available on HuggingFace: https://huggingface.co/chc-harvard, and the code and data for replicating the analyses are available on OSF: https://osf.io/vbkh8/?view_only=3301fa315ba64f2ebf77876767c19c1f.

Case 1 – Effect of CHC training program on online MHC

After removing comments that only consist of single words, only contain emojis, and only tag other users, 164,855 comments from 1,863 videos created by the 49 creators remained in the following analysis. We first explored the change in the outcomes attributable to exposure to our study's interventions, "Materials Only" condition (MO), and "Conference Plus Materials" condition (CM) for content creators, compared to the control group, shown in Table 1. Regarding the knowledge construction process, the MO intervention for creators significantly enhanced general knowledge construction (β=0.03, SE=0.015) and knowledge construction through reflection of personal experience (β=0.06, SE=0.02) in comments on their videos compared to comments on videos by creators in the control group. But the effect was not observed among creators in the CM intervention branch.

Regarding the mental health behavioral intentions, MO treatment to content creators leads to more commenters disclosing that they are receiving well-being related professional services (β=0.02, SE=0.01) compared to exposure to videos from creators in the control group. Moreover, comments to videos produced by creators in both CM (β=-0.05, SE=0.02) and MO (β=-0.05, SE=0.02) groups are less likely to express appreciation to video creators than comments to videos produced by creators in the control group. Comments expressing appreciation to creators often consist of brief phrases such as 'thank you' or 'heart.' The reduction in expressing appreciation in the comments of videos made by treated creators may suggest that more substantial discussions or content is taking place in these comments. To illustrate, a post-hoc analysis on the average comment length, measured by the number of characters per comment, for comments to videos from creators in the treatment versus control groups, shown in Figure A3, found comments are significantly longer in the treatment group, suggesting users are having a more substantive and concrete mental health conversation.

We further investigated the moderating effects of content creator characteristics on the impact of MO and CM interventions on knowledge construction and mental health behavioral intentions in video comments, shown in Table 2. Understanding these moderated effects will guide the selection of creators for future creator training events. We first study the moderated effect by engagement rate, which is defined as the ratio between the total number of comments, likes, shares, and saves and the total number of views. Social media engagement rate reflects the average number of interactions a creator's content receives per follower. The CM treatment on those creators who have larger engagement rates (>10%) leads to more knowledge construction (β=0.10, SE=0.04), and mentions of receiving well-being related professional services (β=0.06, SE=0.02) during and after the intervention in their video comments than creators with lower engagement rate (<=10%), while the effects are not found in the MO treatment. Both the CM (β=-0.10, SE=0.05) and MO (β=-0.15, SE=0.05) treatment on creators with large engagement rate results in less agreement expression in knowledge construction during and after the intervention, compared to creators with engagement rate lower than 10%.

We then examined how the number of TikTok account followers moderates the effects of creator training on mental health knowledge construction and behavioral intentions. It is important to note that the number of followers does not directly correspond to the level of engagement previously measured (Pourazad et al., 2023). While the number of followers indicates the extent of the creators' content reach and popularity, the engagement rate reflects audience involvement and participation, essentially measuring how 'social' the creator's content is on online platforms. We dichotomized the number of followers at a cutoff of 750k, a threshold at which creators typically begin to employ managers for account and content management. We observed that the CM treatment significantly reduced knowledge construction in video comments for creators with more than 750k followers (β=-0.15, SE=0.06) than their peers with less followers, a change not evident among creators in the MO condition. Conversely, the MO treatment, but not the CM treatment, diminished the knowledge construction process through the reduction of clarification-seeking questions (β=-0.10, SE=0.04) during and after the intervention among creators with larger follower counts compared to those with fewer followers.

We further checked the moderated effects by whether the creators are 1) licensed mental health professional and 2) offer paid coaching services. The licensing status and paid-coaching status of each content creator was collected based on their self disclosure on their social media accounts or websites. We observed that the MO treatment, but not the CM treatment, significantly enhanced knowledge construction through expressing disagreement (β=0.05, SE=0.02) and asking for clarification questions (β=0.04, SE=0.02), as well as reduced appreciation expression (β=-0.13, SE=0.04) in video comments for licensed creators than creators who are not licensed. On contrary, the MO treatment, but not the CM treatment, significantly reduced knowledge construction through expressing disagreement (β=-0.06, SE=0.03) and asking for clarification questions (β=-0.07, SE=0.03) in video comments for creators who currently offer paid coaching services than creators who do not.

We further investigated the average marginal effect (AME) of treatment assignment on three specific outcomes: the probability of knowledge construction, expression of appreciation, and disclosure of receiving well-being-related professional services. Additionally, we examined how engagement rate, number of followers, creator licensing, and the status of offering paid coaching services moderate the AMEs, as detailed in Figures 1 and A4-A6. The AME represents the average change in the probability of the outcome that would occur if all comments in the sample were from videos produced by content creators in the respective treatment and moderated groups, namely MO, CM, and control.

Overall, our analysis concludes that the MO treatment enhances knowledge construction and the disclosure of receiving wellbeing-related professional help in video comments. These effects are more pronounced among creators with higher engagement rates, fewer followers, who are licensed, or who do not offer paid coaching services. Conversely, both the MO and CM treatments reduce the expression of appreciation in video comments, which suggests videos from treated creators engender more substantive discussion.

	CM*POST			MO*POST			VPC
	Beta	**SE**	**P value**	**Beta**	**SE**	**P value**	
Knowledge construction							
Summary measures							
Total knowledge construction (binary)	0.02	0.02	0.262	**0.03**	**0.02**	**0.038***	0.67
Total knowledge construction (continuous	0.03	0.04	0.533	0.03	0.04	0.420	0.56
Individual measures							
Personal reflection	0.03	0.02	0.099	**0.06**	**0.02**	**<0.001***	0.57
Express agreement	-0.01	0.02	0.767	0.00	0.02	0.82	0.64
Express disagreement	-0.01	0.01	0.607	-0.02	0.01	0.082	0.66
Ask clarification questions	0.00	0.01	0.921	-0.01	0.01	0.183	0.72
Reinterpret knowledge	0.02	0.02	0.419	0.01	0.02	0.554	0.55
Apply knowledge to new area	0.01	0.02	0.782	0.01	0.02	0.579	0.53
Mental health behavioral intention							
Express appreciation to the creator	**-0.05**	**0.02**	**0.028***	**-0.07**	**0.02**	**0.001***	0.43
Seek wellbeing-related information	-0.01	0.01	0.558	-0.02	0.01	0.108	0.89
Disclose mental health problems	-0.02	0.01	0.128	0.00	0.01	0.658	0.59
Disclose receiving wb-related prof. service*	0.02	0.01	0.057	**0.02**	**0.01**	**0.010***	0.75
Share wellbeing-related coping strategy	-0.01	0.01	0.631	0.00	0.01	0.641	0.60

Table 1. Effect of MO and CM Interventions Assignment on Commenting Behaviors.

* wb-related prof. service: well-being-related professional service

Parameter estimates are standardized and from a three-level multilevel Linear Probability Models (LPM), with 164,855 comments nested in 1,863 videos that are created by 49 creators (CM group N=20, MO group N=14, Control group N=15). The primary outcomes are 13 dichotomized variables on mental health knowledge construction and mental health behavioral intentions measured by content analysis assisted-LLM. The models identify the effect of change in the outcomes attributable to exposure to our study's interventions by interacting dichotomous fixed effect indicators of treatment group assignment (MO vs. CM, with the Control serving as a reference group), with the fixed effect indicator of whether or not each video was produced pre or post intervention (pre period as March 2023 and post period as Apr and May 2023). Each row represents an outcome measured by the LLM fine-tuned through our content analysis, with the two columns in the same row modeled together as outcomes from the same LPM. Detailed findings and additional parameter estimates of the LPM are provided in the supplementary files.

			Moderator = Engagement>10%	Moderator = Follower>750k	Moderator = Licensed	Moderator = Coaching
Knowledge Construction	Total knowledge construction (binary)	CM X POST X Moderator	**0.10 (0.04)***	**-0.15 (0.06)***	0.03 (0.04)	0.02 (0.08)
		MO X POST X Moderator	0.00 (0.04)	-0.02 (0.06)	0.03 (0.03)	-0.06 (0.04)
	Total knowledge construction (continuous)	CM X POST X Moderator	0.07 (0.11)	-0.20 (0.14)	0.15 (0.10)	-0.20 (0.19)
		MO X POST X Moderator	0.05 (0.09)	-0.17 (0.13)	0.13 (0.08)	-0.21 (0.11)
	Personal reflection	CM X POST X Moderator	0.05 (0.05)	-0.13 (0.07)	0.06 (0.04)	-0.13 (0.09)
		MO X POST X Moderator	0.04 (0.04)	-0.05 (0.06)	0.06 (0.04)	-0.08 (0.05)
	Express agreement	CM X POST X Moderator	**-0.10 (0.05)***	0.06 (0.07)	-0.01 (0.04)	-0.03 (0.08)
		MO X POST X Moderator	**-0.14 (0.04)***	0.03 (0.06)	-0.04 (0.04)	-0.01 (0.05)
	Express disagreement	CM X POST X Moderator	0.03 (0.03)	-0.02 (0.04)	0.01 (0.03)	0.01 (0.05)
		MO X POST X Moderator	0.05 (0.02)	-0.06 (0.04)	**0.05 (0.02)***	**-0.06 (0.03)***
	Ask clarification questions	CM X POST X Moderator	0.01 (0.03)	-0.07 (0.04)	0.03 (0.02)	-0.05 (0.05)
		MO X POST X Moderator	0.02 (0.02)	**-0.10 (0.04)***	**0.04 (0.02)***	**-0.07 (0.03)***
	Reinterpret knowledge	CM X POST X Moderator	0.06 (0.05)	-0.06 (0.07)	0.04 (0.05)	-0.02 (0.09)
		MO X POST X Moderator	0.05 (0.04)	-0.03 (0.06)	0.06 (0.04)	-0.05 (0.05)
	Apply knowledge to new area	CM X POST X Moderator	0.04 (0.05)	-0.07 (0.07)	0.05 (0.04)	-0.02 (0.09)
		MO X POST X Moderator	0.03 (0.04)	-0.03 (0.06)	0.00 (0.04)	0.00 (0.05)
Mental Health Behavioral Intentions	Express appreciation to the creator	CM X POST X Moderator	**-0.13 (0.06)***	0.06 (0.07)	-0.03 (0.05)	0.10 (0.10)
		MO X POST X Moderator	**-0.17 (0.05)***	-0.03 (0.06)	**-0.13 (0.04)***	0.05 (0.06)
	Seek well-being information	CM X POST X Moderator	-0.01 (0.03)	-0.00 (0.04)	0.00 (0.03)	-0.03 (0.05)
		MO X POST X Moderator	-0.02 (0.02)	-0.03 (0.04)	0.01 (0.02)	-0.04 (0.03)
	Disclose mental health problems	CM X POST X Moderator	0.01 (0.03)	-0.04 (0.04)	-0.05 (0.03)	0.04 (0.05)
		MO X POST X Moderator	0.01 (0.03)	-0.07 (0.04)	-0.01 (0.02)	-0.01 (0.03)
	Disclose receiving wb-related prof. service*	CM X POST X Moderator	**0.06 (0.02)***	-0.03 (0.03)	0.01 (0.02)	-0.01 (0.04)
		MO X POST X Moderator	0.02 (0.02)	0.01 (0.03)	0.03 (0.02)	-0.03 (0.02)
	Share well-being-related coping strategy	CM X POST X Moderator	-0.02 (0.03)	-0.06 (0.04)	0.01 (0.03)	-0.07 (0.05)
		MO X POST X Moderator	-0.01 (0.02)	-0.06 (0.03)	0.01 (0.02)	-0.02 (0.03)

Table 2. Moderated Effects of MO and CM Intervention Assignment on Commenting Behavior.

* wb-related prof. service: well-being-related professional service

Parameter estimates are standardized and from a three-level multilevel Linear Probability Models (LPM), with 164,855 comments nested in 1,863 videos that are created by 49 creators (CM group N=20, MO group N=14, Control group N=15). The primary outcomes are 13 dichotomized variables on mental health knowledge construction and mental health behavioral intentions measured by content analysis assisted-LLM. The models identify how the effect of change in the outcomes attributable to exposure to our study's interventions are moderated by four moderators, including engagement rate (larger than 10% or not), follower number (larger than 750k or not), creator is licensed practitioner or not, and creator is offering paid coaching service or not. To achieve the goal, the models created three-way interaction among dichotomous fixed effect indicators of treatment group assignment (MO vs. CM, with the Control serving as a reference group), the fixed effect indicator of whether or not each video was produced pre or post intervention (pre period as March 2023 and post period as Apr and May 2023), and the dichotomized moderator. Each row represents the modeling results of a moderator, and each row corresponds to a study outcome measured by the LLM fine-tuned through our content analysis. Detailed findings and additional parameter estimates of the LPM are provided in the supplementary files.

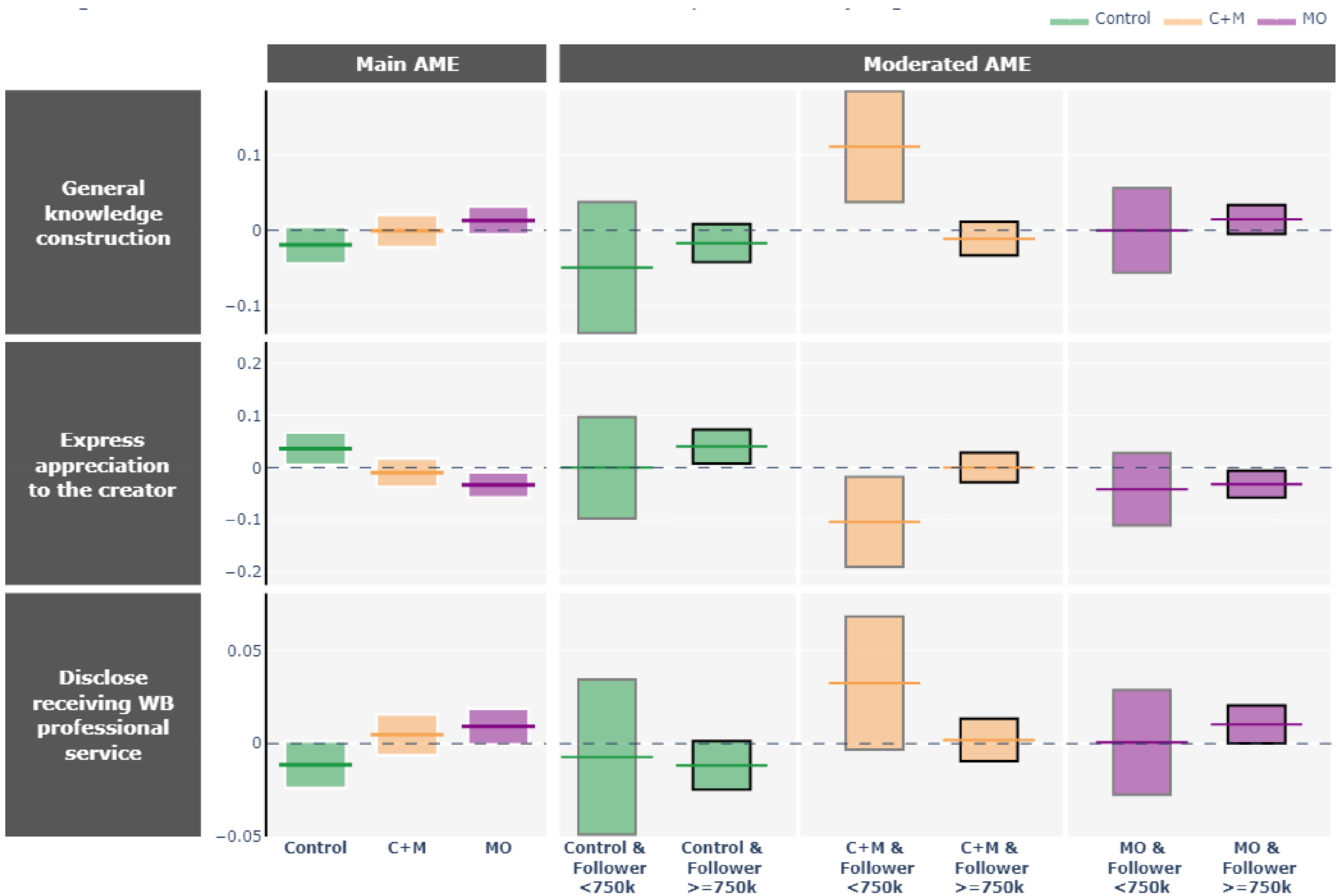


Figure 1. The average marginal effect (AME) and the moderated AME of treatment assignment on the probability of knowledge construction, express appreciation, and disclose receiving well-being related professional service, moderated by whether the creators has a large number of follower (>750k)

Case 2 – Effect of TikTok Mental Health Awareness Month (MHAM) on online MHC

In the second case, we applied the large language models (LLMs) and the analytical strategy developed earlier to assess the impact of TikTok's Mental Health Awareness Month (MHAM) in May 2023 on mental health knowledge construction and mental health behavioral intentions using video comments (TikTok, 2023). We collected videos posted between January 1, 2023, and August 1, 2023, from the US, tagged with #MentalHealthAwareness. The videos were restricted to the US, as TikTok's MHAM is a region-specific event. 139,251 comments from 11,479 videos created by 8,664 users were collected. After removing comments that only consist of single words, only contain emojis, and only tag other users, 88,979 comments from 7,014 videos created by 5,597 users remained in the following analysis. We employed a sharp regression discontinuity design (RDD) to estimate the impact of MHAM on mental health communication (MHC) outcomes observed in the comments. The outcomes were aggregated at the hourly level, with two model specifications: one using May 1, 00:00, as the cutoff point to capture immediate changes in user MHC following the start of MHAM, and the other using May 1–30 as the cutoff to compare data from April to June and assess whether the effects extended beyond May 2023. The results are shown in Table 3, detailing the impact of TikTok's MHAM on video engagement and online MHC. The detailed specifications for the models in Table 3 are documented in Table A5.

Regarding knowledge construction, MHAM led to a significant improvement in general knowledge construction in comments immediately after the start of the event (β = 0.07, Robust CI: 0.04–0.13), with a greater effect persisting through June 2023 (β = 0.18, Robust CI: 0.14–0.26). MHAM also enhanced knowledge construction through personal reflection (β = 0.10, Robust CI: 0.06–0.18), reinterpretation of knowledge (β = 0.12, Robust CI: 0.05–0.22), and application to new areas of knowledge (β = 0.09, Robust CI: 0.03–0.17) right after the event began, with these effects continuing into June 2023 (personal reflection: β = 0.53, Robust CI: 0.45–0.66; reinterpretation: β = 0.09, Robust CI: 0.01–0.20; application: β = 0.08, Robust CI: 0.01–0.17).

Additionally, MHAM was associated with a reduction in expressions of appreciation towards creators in the comments immediately after the event (β = -0.14, Robust CI: -0.28, -0.06), with the effect extending through June 2023 (β = -0.30, Robust CI: -0.45, -0.18). We also observed an increase in the disclosure of mental health issues in comments following the start of MHAM (β = 0.02, Robust CI: 0.0—0.06), with an even larger effect observed in June 2023 (β = 0.24, Robust CI: 0.16–0.37). However, MHAM did not significantly affect the number of comments, shares, views, or likes on the videos.

In summary, our evaluation shows that MHAM improves knowledge construction and the self-disclosure of mental health issues among TikTok users, highlighting its potential to enhance online MHC and provide community support for individuals in need. However, MHAM did not lead to increased engagement or reach of MHC-related content, suggesting the absence of algorithmic amplification for such content during the event.

	Effect of MHA on health communication on May (May 1 00:00 as cutoff)		Effect of MHA on health communication on June (May 1 -- 30 as cutoff)	
	Point estimate	**Robust CI**	**Point estimate**	**Robust CI**
Video engagement				
Like count (log)	0.04	(-0.02, 0.11)	0.02	(-0.04, 0.10)
Share count (log)	0.03	(-0.06, 0.11)	-0.03	(-0.13, 0.06)
View count (log)	0.04	(-0.00, 0.08)	0.01	(-0.03, 0.07)
Comment count (log)	0.03	(-0.06, 0.11)	-0.02	(-0.12, 0.07)
Knowledge construction				
Summary measures				
Total knowledge construction (binary)	-0.03	(-0.08, 0.05)	**0.21**	**(0.16, 0.33)**
Total knowledge construction (continuous)	**0.07**	**(0.04, 0.13)**	**0.18**	**(0.14, 0.26)**
Individual measures				
Personal reflection	**0.10**	**(0.06, 0.18)**	**0.53**	**(0.45, 0.66)**
Express agreement	-0.02	(-0.13, 0.05)	0.00	(-0.09, 0.10)
Express disagreement	-0.01	(-0.04, 0.02)	0.01	(-0.02, 0.05)
Ask clarification questions	-0.01	(-0.04, 0.03)	-0.02	(-0.04, 0.02)
Reinterpret knowledge	**0.12**	**(0.05, 0.22)**	**0.09**	**(0.01, 0.20)**
Apply knowledge to new area	**0.09**	**(0.03, 0.17)**	**0.08**	**(0.01, 0.17)**
Mental health behavioral intentions				
Express appreciation to the creator	**-0.14**	**(-0.28, -0.06)**	**-0.30**	**(-0.45, -0.24)**
Seek wellbeing-related information	0.00	(-0.04, 0.03)	-0.01	(-0.04, 0.03)
Disclose mental health problems	**0.02**	**(0.00, 0.06)**	**0.24**	**(0.16, 0.37)**
Disclose receiving wellbeing-related prof. service	-0.01	(-0.03, 0.01)	**-0.03**	**(-0.06, -0.00)**
Share wellbeing-related coping strategy	-0.01	(-0.05, 0.03)	-0.03	(-0.08, 0.01)

Table 3. The effect of TikTok's Mental Health Awareness Month (MHAM) intervention on video engagement, mental health knowledge construction and mental health behavioral intentions.

Given the sensitivity of the treatment effect estimator in regression discontinuity (RD) designs to both bandwidth and polynomial order, we have implemented several measures to ensure the robustness of our estimates (Gelman & Imbens, 2019). All models were estimated using the rdrobust package in R, which applies triangular kernel weights and conducts a data-driven search to select the optimal bandwidth. We fitted models with polynomial orders ranging from 1 to 4 using *rdrobust* proposed by (Calonico et al., 2015) and applied the *RDMSE* methods proposed by (Pei et al., 2020) to compute the asymptotic mean squared error (AMSE) for each local polynomial regression discontinuity. The polynomial order that minimized the AMSE was selected as the final model, and the results reported here are based on this selection. The *rdrobust* package is specifically designed to incorporate recent advances in bandwidth selection within the RD literature, while the *RDMSE* package builds on *rdrobust* to ensure a valid selection of polynomial order. Detailed specifications for the models in Table 3, including bandwidth selection, polynomial orders, and the effective sample sizes used in the point and bias-correction estimators, are documented in Table A5.

Case 3 – Effect of TikTok Mental Health Awareness Month institution and creator spotlights on user MHC

During TikTok's Mental Health Awareness Month (MHAM) in 2023, the platform partnered with seven organizations that support mental wellbeing (TikTok, 2023). TikTok also launched the Mental Health Media Education Fund, donating “over $2 million in ad credits” to these partner organizations. Additionally, TikTok spotlighted the work of 10 creators who “use TikTok to educate the community on #MentalHealthAwareness and made significant impact both on and off the platform over the past year.” In total, 17 institutions and creators were spotlighted on the MHAM webpage.

In the third case, we applied the LLMs and the analytical strategy developed earlier to assess the impact of TikTok's MHAM spotlighted institutions and creators on knowledge construction and mental health behavioral intentions based on video comments. Using the TikTok Research Tools API (TikTok, n.d.), we requested videos created by the content creators listed on TikTok’s MHAM webpage between January 1 and August 31, 2023, resulting in 169,628 comments from 1,489 videos by the spotlighted institutions and creators. After removing comments consisting of only single words, emojis, or tags of other users, 139,473 comments from 1,321 videos by the collaborating creators remained for analysis. Notably, the sample in Case 3 is distinct from that in Case 2 because 1) Case 2 focused solely on videos posted from the US, while Case 3 included all 17 creators collaborating with TikTok, and 2) videos in Case 3 were not required to feature the #MentalHealthAwareness hashtag. We once again applied a sharp regression discontinuity design (RDD), following the strategy in Case 2, to study the effect of TikTok’s MHAM creator spotlights on video engagement and online MHC, with results presented in Table 4. The detailed specifications for the models in Table 4 are documented in Table A6.

Regarding knowledge construction, we found that MHAM spotlights significantly reduced general knowledge construction in the comments ($\beta = -0.09$, Robust 95% CI: -0.19, -0.02), including a decrease in knowledge construction through clarification-seeking questions ($\beta = -0.07$, Robust 95% CI: -0.22, -0.02) and applying knowledge to new contexts ($\beta = -0.16$, Robust 95% CI: -0.32, -0.04) after May 2023 compared to before.

In terms of mental health behavioral intentions, the MHAM spotlights led to a decrease in wellbeing-related information-seeking behavior ($\beta = -0.07$, Robust 95% CI: -0.16, -0.00) and reduced the self-disclosure of mental health issues ($\beta = -0.03$, Robust 95% CI: -0.07, -0.01) in comments following the start of MHAM. Similar to the findings in Case 2, MHAM spotlights did not significantly increase comment, share, view, or like counts for videos created by the spotlighted institutions or creators.

In summary, our evaluation unexpectedly revealed that the MHAM spotlighted institutions and creators reduced knowledge construction, wellbeing-related information-seeking, and self-disclosure of mental health issues among TikTok users. These findings will be further explored in the discussion section.

	Effect of MHA on health communication on May (May 1 00:00 as cutoff)		Effect of MHA on health communication on June (May 1 -- 30 as cutoff)	
	Point estimate	**Robust CI**	**Point estimate**	**Robust CI**
Video engagement				
Like count (log)	0.04	(-0.05, 0.13)	0.06	(-0.06, 0.21)
Share count (log)	0.02	(-0.13, 0.16)	0.09	(-0.10, 0.29)
View count (log)	0.05	(-0.05, 0.12)	0.06	(-0.04, 0.16)
Comment count (log)	0.08	(-0.03, 0.22)	0.08	(-0.10, 0.24)
Knowledge construction				
Summary measures				
Total knowledge construction (binary)	0.06	(-0.01, 0.13)	0.01	(-0.11, 0.09)
Total knowledge construction (continuous)	-0.02	(-0.10, 0.04)	**-0.09**	**(-0.19, -0.02)**
Individual measures				
Personal reflection	-0.03	(-0.18, 0.09)	-0.05	(-0.27, 0.05)
Express agreement	-0.05	(-0.20, 0.09)	-0.06	(-0.12, 0.20)
Express disagreement	0.03	(-0.08, 0.10)	0.01	(-0.14, 0.05)
Ask clarification questions	-0.07	(-0.16, 0.02)	**-0.07**	**(-0.22, -0.02)**
Reinterpret knowledge	0.07	(-0.02, 0.19)	0.08	(-0.16, 0.10)
Apply knowledge to new area	0.02	(-0.09, 0.12)	**-0.16**	**(-0.32, -0.04)**
Mental health behavioral intentions				
Express appreciation to the creator	-0.08	(-0.21, 0.00)	-0.10	(-0.03, 0.22)
Seek wellbeing-related information	**-0.07**	**(-0.16, -0.00)**	-0.08	(-0.18, 0.02)
Disclose mental health problems	**-0.03**	**(-0.07, -0.01)**	-0.04	(-0.08, 0.00)
Disclose receiving well-being-related prof. service	-0.01	(-0.03, 0.02)	0.00	(-0.02, 0.04)
Share wellbeing-related coping strategy	-0.02	(-0.08, 0.02)	-0.03	(-0.05, 0.06)

Table 4. The effect of TikTok's Mental Health Awareness Month (MHAM) creator spotlight on video engagement, mental health knowledge construction, and mental health behavioral intentions of TikTok audience.

Given the sensitivity of the treatment effect estimator in regression discontinuity (RD) designs to both bandwidth and polynomial order, we have implemented several measures to ensure the robustness of our estimates (Gelman & Imbens, 2019). All models were estimated using the rdrobust package in R, which applies triangular kernel weights and conducts a data-driven search to select the optimal bandwidth. We fitted models with polynomial orders ranging from 1 to 4 using *rdrobust* proposed by (Calonico et al., 2015) and applied the *RDMSE* methods proposed by (Pei et al., 2020) to compute the asymptotic mean squared error (AMSE) for each local polynomial regression discontinuity. The polynomial order that minimized the AMSE was selected as the final model, and the results reported here are based on this selection. The *rdrobust* package is specifically designed to incorporate recent advances in bandwidth selection within the RD literature, while the *RDMSE* package builds on *rdrobust* to ensure a valid selection of polynomial order. Detailed specifications for the models in Table 4, including bandwidth selection, polynomial orders, and the effective sample sizes used in the point and bias-correction estimators, are documented in Table A5.

Discussion

Creator training and MHAM initiatives boost audiences' mental health knowledge construction, but MHAM institution and creator spotlights do not

Our analysis concludes that the use of asynchronous toolkits on evidence-based mental health communication in creator training notably improves the mental health knowledge construction among video commenters (Case 1). The asynchronous materials/toolkits provided to creators increased the amount of evidence-based mental health content (EBMHC) they made on TikTok. Here we show that this enhancement is also transferred to an increased level of mental health knowledge construction among video viewers (Motta et al., 2024). To our knowledge, this study is the first to effectively evaluate the effect of creator training programs on user behaviors, a significant result considering TikTok's vast user base and the total audience mental health content on the platform can command. For instance, videos posted by creators in the CM+MO groups from March to May 2023 garnered approximately 2 million views. A 3% increase in mental health knowledge construction among the millions of video audiences can be translated into a substantial boost in mental health literacy across the population. Furthermore, we observed that TikTok's MHAM event also boosted the likelihood of knowledge construction among users, with effects persisting to June 2024 (Case 2). This study is among the first to empirically evaluate the impact of a social media platform's initiative on user outcomes. However, a closer examination of the mental health communication in comments to videos from the 7 institutions and 10 individual creators spotlighted by TikTok revealed a decrease in the mental health knowledge construction process (Case 3). The next section of the discussion will explore how the selection of content creators spotlighted by TikTok's MHAM might have influenced this outcome.

Creator training and MHAM enhance the amount of substantive mental health conversation on TikTok, but MHAM creator spotlights do not

Our findings further indicate that the creator training programs (both CM and MO) generally reduce expressions of appreciation towards the video audience, a trend also noted in the MHAM cases. These findings are in contrast to our expectation, which anticipated increased expressions of appreciation in the treatment group. In a post-hoc analysis, we found that comments showing expression to authors are often short in length and contain brief phrases like 'thank you' or emoji 'heart.' This reduction may indicate that comments on videos from treated creators involve more substantive mental health discussions. Further supporting this observation, the MO treatment was found to increase commenters' disclosures of receiving wellbeing-related professional help. Additionally, there was a significant increase in disclosures of self-mental health challenges following the MHAM event. A key intervention during MHAM was the encouragement of the use of the #MentalHealthAwareness hashtag, which our findings suggest made videos more likely to address personal mental health issues post-event, indicating the effectiveness of MHAM and its potential to reduce stigma and foster online discussions about mental health. One caveat in our study is that, when annotating self-disclosures of mental health conditions, we did not distinguish between self-diagnosed mental health issues and those diagnosed by professionals (Yıldırım, 2023). In future work, we plan to differentiate these two measures, as they convey distinct implications: self-diagnosis can be problematic and potentially mislead others, whereas sharing a professionally diagnosed mental health condition may represent a courageous step towards self-expression and destigmatization on the platform.

How can designers of creator training or engagement programs ensure their interventions improve online mental health communication?

Insight 1: Work with the appropriate group of content creators

Our study reveals that creator training programs most significantly enhance video commenter's mental health knowledge construction among creators who have high engagement rates, fewer than 750k followers, who are licensed mental health providers, or who do not offer paid coaching services. This insight suggests that future training programs should target creators with smaller audiences or higher engagement rates, because such creators are more likely to apply and disseminate evidence-based mental health communication to their followers.

Furthermore, our findings indicate that TikTok's MHAM creator spotlights actually reduces knowledge construction in video comments, again highlighting the targeting collaborators of content creators campaigns. Note that the creators spotlighted by MHAM tend to be either creators with large number of followers, or mental health related institutions. Our analysis suggests that to improve TikTok users' mental health communication, platforms and institutions may want to consider adjusting creator selection criteria. The finding points to the potential benefits of prioritizing emerging creators over established ones if the goal is to bolster the health of the mental health discourse online. Platforms like TikTok and institutions aiming to train content creators might reconsider their target group of collaborators to achieve this.

Insight 2: Separate the toolkits learning and conference training sections in creator program design

Our results indicate that the effectiveness of creator training programs on video audiences' knowledge construction varies depending on the delivery method. In our previous study, creators provided with simple, asynchronous training material toolkits were more likely to incorporate evidence-based mental health content (EBMHC) into their videos, while adding a synchronous virtual conference component did not significantly enhance EBMHC production (Xu et al., 2024). This trend persisted when examining the training's impact on user-level mental health knowledge construction: the provision of simple, asynchronous toolkits enhanced knowledge construction among video commenters (the MO group), whereas the addition of a synchronous virtual conference did not produce a significant increase in the outcome (the CM group).

This finding underscores the need to reconsider the design of future training delivery to mental health content creators. For creators in the CM group, who both received the toolkits and participated in the virtual conference, the toolkit learning process occurs in a rich virtual conference environment with health professional-led briefings and networking opportunities. The time committed to attending and engaging with multiple virtual conference sessions is time that is not available for creators to make content. That is, the toolkits represent a more efficient catalyst of evidence insertion in creator content. Our findings suggest that future training programs for creators prioritize the delivery of easy-to-navigate evidence toolkits directly to creators via email or text.

However, this does not suggest that synchronous training is ineffective or unnecessary. Our experience suggests that virtual creator briefings hosted by well-known creators are powerful tools for connecting creators with our toolkits. In addition, a qualitative study exploring creators' experiences at an in-person creator summit revealed significant benefits beyond knowledge acquisition from asynchronous toolkits (Xu et al., 2024). For instance, creators not only learned about conducting evidence-based health communication online but also built new connections and drew inspiration for content generation. Future creator training programs should consider how best to integrate toolkits, virtual creator briefings, and in-person creator summits. For example, designers of virtual and in-person components might consider setting aside dedicated time for toolkit review and discussion, enabling onsite content creation, and dedicating time for networking. Future studies should also aim to quantify the potential impacts of virtual or in-person components on these additional dimensions.

The low prevalence of mental health misinformation in user comments on TikTok mental health videos in this study requires cautious interpretation and rigorous future analyses

Another finding from this study, which requires cautious interpretation, is the low prevalence of mental health misinformation among the 4,000+ comments from 54 randomly selected videos we annotated (Figure A2). This result does not necessarily indicate that the prevalence of mental health misinformation in TikTok user comments is universally low. Previous research found that 26.1% of 1,534 statements from 144 threads in two Italian-speaking Facebook groups contained medically inaccurate mental health information (Bizzotto et al., 2023). Similarly, Nguyen et al. identified mental health misinformation in 13.2% of 650 YouTube videos and 36% of 100 BitChute videos (V. C. Nguyen et al., 2024).

Our findings differ from prior studies in several key ways. Firstly, we focused on comments on TikTok, while other studies analyzed video content or group chats, highlighting differences in both the type of content and platforms used. Identifying misinformation in brief comments, as opposed to more extended video content or group conversations, poses a more difficult challenge. Secondly, Nguyen et al. used a more rigorous coding schema for mental health misinformation, defining it as “medically relevant claims about mental health that are partially or wholly false,” developed through expert-driven discussions (V. C. Nguyen et al., 2024). In contrast, our approach relied on annotators' recognition of common examples of misinformation, leveraging their expertise as public health workers or graduate students. Future research should focus on developing a more rigorous codebook for annotating mental health misinformation in TikTok video comments.

Limitations of the current study

Our study has several limitations. First, the experiment providing training to content creators faced challenges including the selected sample being an incomplete representation of mental health content creators on TikTok; our sample may also disproportionately include creators particularly motivated to integrate evidence-based mental health content (EBMHC) into their work. These limitations and potential remedies were discussed in a previous paper.

Second, our analysis of mental health knowledge construction and related behavioral intentions among user comments represents just one of the many ways that the interventions evaluated here—either the asynchronous toolkits or the synchronous virtual conference—can impact users.

Future research could explore other outcomes, such as improvements in social support or self-diagnosis of mental health problems, and should develop a more rigorous and systematic scheme to assess the level of mental health misinformation in the comments.

Third, our study examined the impact of the training conducted in April 2023 on knowledge construction among video viewer comments from March to May 2023. We have not yet determined if these audience effects are sustained over the long term. Future research should aim to understand the extent to which these audience effects persist.

Forth, in Cases 2 and 3, when examining the effect of TikTok's Mental Health Awareness Month, we lacked precise information on the exact timing of the MHAM interventions (known only to have launched in May 2023), which may lead to mis-specifying the cut-off points in our regression discontinuity design analysis. However, we contend that our current estimation, using 00:00 on May 1, 2023, likely underestimates the true effect, assuming the MHAM interventions were launched at any time after this point.

Fifth, our study examines TikTok users' mental health knowledge construction and behavioral intentions using comments; however, not all TikTok users comment on videos. Consequently, our results may disproportionately represent users who are more likely to comment. Future studies should explore alternative methods to obtain a more representative sample of TikTok or social media users, for example, by combining population-representative surveys with social media data donation packages (van Driel et al., 2022).

Finally, our study was solely focused on TikTok. As such, we cannot empirically generalize our findings to other platforms. However, given that many social media platforms now prioritize features that offer algorithmically personalized recommendations (such as TikTok's "For You" page) and focus on short video content (Bhandari & Bimo, 2022; Wong, 2022), we hypothesize that similar findings could be observed on other platforms like Instagram or YouTube. Future studies, in collaboration with other institutions involved in creator training and collaboration programs could test these hypotheses using the framework established by our series of research (Motta et al., 2024).

Conclusion

The current study devised a method to analyze the impact of content creator training programs on the perceptions and behaviors of video commenters on social media platforms like TikTok. Utilizing LLM-assisted content analysis of user comments, we fine-tuned LLMs that exhibited satisfying performance in identifying levels of knowledge construction and mental health behavioral intentions. The LLMs employed in our analysis can be adapted for future research.

With this approach, we found that asynchronous toolkits on evidence-based mental health communication significantly improved mental health knowledge construction among video commenters. In contrast, simultaneously adding a synchronous virtual conference to the asynchronous toolkits did not significantly enhance this effect. The impact was more pronounced among creators who have higher engagement rates, fewer than 750k followers, are licensed mental health providers, and do not offer paid coaching. These findings suggest that future creator training initiatives could prioritize working with micro- or mid-tier creators, such as those with less than 750k followers, to enhance mental health knowledge construction among users. Moreover, future training programs might consider separating toolkit learning and conference training in their program design to maximize the effectiveness of the training.

Additionally, we observed that TikTok's Mental Health Awareness Month (MHAM) initiatives in May 2023 enhanced audiences' mental health knowledge construction and encouraged more self-disclosure of mental health challenges. However, this effect was not evident among the creators and institutions TikTok spotlighted during MHAM. These findings indicate that promoting hashtags like #MentalHealthAwareness in online health awareness campaigns can foster substantive discussions on mental health and enhance knowledge construction. Nonetheless, platforms or other organizations aiming to boost mental health knowledge on social media should reassess their criteria for selecting content creators for collaborations.

Method

Study overview

The current study pursues two primary objectives: firstly, to develop LLMs as text classifiers for evaluating online mental health communication using user comments on TikTok videos. These classifiers focus on key outcomes including knowledge construction, mental health misinformation, and mental health behavioral intentions (refer to “Outcome Measurement”). Secondly, to apply the developed LLMs to assess the effects of three online mental health communication interventions, including a training program for mental health content creators (MHCCs) by <<institute name dropped for peer review>> (refer to "Case 1: Creator Training Program"), the TikTok Mental Health Awareness Month (MHAM) held in May 2023 (refer to "Case 2: MHAM Program"), and the MHAM spotlight (refer to "Case 3: MHAM Spotlight"), on user-level outcomes including mental health knowledge construction and mental health behavioral intentions.

This study was exempted from institutional review by Harvard Longwood Campus Institutional Review Board on the Use of Human Subjects and did not require informed consent as it involved no human subjects data. All procedures complied with applicable guidelines and regulations.

Outcome measurement

Given our hypothesis that TikTok serves as a platform for informal learning and community support in mental health, we have structured our study to assess three outcome categories: knowledge construction, mental health misinformation, and mental health behavioral intentions, with specifics provided in Table A1. For knowledge construction, we utilized the Interaction Analysis Model (IAM) (C. Gunawardena et al., 2016; C. N. Gunawardena et al., 1997), a framework proven effective in exploring informal learning in social media interactions (Haythornthwaite et al., 2018; H. Nguyen & Diederich, 2023). IAM outlines learning progresses as users exchange ideas, confront dissonance, negotiate meanings, evaluate emerging understandings, and apply new knowledge to novel contexts, with each stage growing in complexity. Based on IAM and prior research, we crafted a measure for knowledge construction in TikTok video comments for mental health communication across six dimensions: personal reflection, expressing agreement, expressing disagreement, asking for clarifications, reinterpreting knowledge, and applying knowledge to new areas, with each dimension quantified as a binary variable. From this, we derived two aggregated measures: a general knowledge construction score which is the sum of the binary scores across dimensions; and a dichotomized general knowledge construction variable, assigned a value of 1 if the score exceeds one, indicating engagement in at least one dimension of knowledge construction.

Regarding mental health misinformation, we assessed the presence of mental health misinformation, corrections of misinformation, discriminatory or stigmatizing content, and corrections of discriminatory content for each comment. To ensure a high-quality evaluation of mental health misinformation assessment, we rely on: 1) research assistants with graduate-level expertise in mental health to leverage their knowledge for identifying misinformation (refer to: "Case 1–Content analysis"), 2) a compiled list of common mental health misinformation used during coder training, and 3) training on utilizing fact-checking websites to verify questionable content during analysis (NIH, 2023).

For mental health behavioral intentions, outcomes measured for each comment include seeking for well-being information, sharing well-being related coping strategies, disclosing wellbeing related professional help received, self-disclosing of mental health issues, and encouraging others for professional help-seeking. Definitions and examples for these measures are detailed in Table A1.

Case 1: Creator Training Program

In the first case study, our team built on a previous intervention by <<institute name dropped for peer review>> that identified a group of mental health content creators (MHCCs) and conducted a randomized control trial. The inclusion criteria of the MHCCs are: aged 18 or over, English-language mental health content, have at least 10,000 followers across TikTok or Instagram social media platforms, posted videos on the platform at least 4 times per month from December 2022 - February 2023, and have been active on the site since February 2022.The interventions were held on April 2023, including a series of virtual summit and an online content creation toolkit around evidence-based mental health communication (EBMHC). Details of the recruitment process and intervention are documented in our previous work (Motta et al., 2024). The trial comprised three groups: one provided with an asynchronous online content creation toolkit ("Materials Only" Condition, or "MO", N=17), another receiving the toolkit plus a synchronous virtual summit component ("Conference Plus Materials" Condition, or "CM", N=25), and a control group (N=20).

Video comment data collection

To further evaluate the intervention's impact on user behavior, we obtained all videos and related comments from 62 participating Mental Health Content Creators (MHCCs) via the TikTok research API in August 2023, using Python version 3.9. From this group, we successfully retrieved

3,465 videos posted by 58 MHCCs in March, April, and May 2023. Additionally, we collected user information, such as the number of followers, for these MHCCs through the same API. Four creators were excluded from the analysis because they either did not produce videos during the study period or had deleted their videos at the time of the request. Videos without comments were also excluded from the study. We accessed comments from 2,858 videos using the TikTok research API. Note that the TikTok research API imposes data filters, excluding, for instance, “public data from users under 18 and data from Canada.” For videos not captured by the API, we manually collected comments, ultimately compiling comments from 2,901 videos. To align with our research focus on online mental health communication, we applied a filter to exclude videos not relevant to mental health discussions. This filter was based on the content analysis by research assistants in our previous study, that determined whether each video is relevant to mental health communication (Motta et al., 2024). Applying the germen-to-mental health filter resulted in a final sample of 1,882 videos and 188,169 comments from 49 MHCCs (MO: N=15, CM: N=20, Control: N=15). The flow chart in Figure A1 presents the process of video and comments collection.

Content analysis

We conducted a content analysis of comments from a sample of TikTok videos to explore knowledge construction, mental health misinformation, and features of health communication. The codebook for the outcome measures is detailed in Table A1. In addition to the outcome measures, each comment was evaluated to determine if it contained only emojis, was a single word (e.g., "wow"), solely tagged other users (e.g., "@auuuu"), or exclusively expressed appreciation (e.g., "thank you").

Given the hierarchical nature of the creator-video-comment data, we employed clustered randomized sampling to select comments for content analysis, resulting in 54 videos and 4,221 comments. Three research assistants from <<institute name dropped for peer review>> were trained and served as coders for the project. After an initial code training session, which contained a series of iterative pilot coding and feedback sessions, each assistant was assigned to code approximately 9% of all sampled comments (376 comments from 5 randomly selected videos). In the content analysis, each assistant first watched the TikTok video, then started coding the comments. The RAs are blinded on whether the videos and comments are from creators in the treatment or control group in the experimental study. Inter-coder reliability (ICR) was assessed for these comments. The ICR results on the “triple assigned” comments highlighted three outcome variables with agreement scores below 0.70: knowledge construction through personal reflection, expressing agreement, and knowledge reinterpretation. Based on the ICR results, a second training session was held to focus on improving coder agreement in the three outcomes. Coders subsequently re-coded a subset of 120 comments where discrepancies had been noted for the three outcomes. Following the two-round training sessions, ICR exceeded 80% agreement for all variables in the filtering session, misinformation, and communication features categories, and surpassed 70% for most knowledge construction variables—except for expressing agreement and knowledge reinterpretation, which both achieved a 68% agreement rate (Table A2). Upon completing the two training sessions, coders commenced the coding of video comments, with each coder handling approximately one-third of the selected comments from the training set. The distributions of the study outcomes labeled in the content analysis are presented in Figure A2.

Text Augmentation

Using the true labels derived from this content analysis, we initially performed character, word-level, and LLM-based text augmentation to ensure a sufficiently large and balanced sample for model fine-tuning. We employed the textattack library to conduct character and word-level text augmentation (Morris et al., 2020). Specifically, at the character level, two substituted sentences were generated using transformations including WordSwapRandomCharacterDeletion, WordSwapRandomCharacterInsertion, WordSwapRandomCharacterSubstitution. At the word level, five augmented sentences were generated using transformations including WordInsertionRandomSynonym, WordSwapChangeLocation, WordSwapChangeName, WordSwapChangeNumber, WordInnerSwapRandom, WordSwapEmbedding, and WordSwapQWERTY. Both sets of transformations were subject to two constraints: Repeat Modification and Stopword Modification. For LLM-based text augmentation, five augmented sentences were generated by applying the model “gpt-4-turbo” in the Chat Completions API from OpenAI with temperature set as 0.5 (OpenAI et al., 2024). The prompts we used are presented in Appendix Section 1. All text augmentation analyses were conducted in Python 3.9.

We apply cosine similarity to evaluate the performance of the augmented text. The original and augmented texts were embedded using the sentence_transformer library with the "bert-base-nli-mean-tokens" model, and the cosine similarity between the embeddings of the original and augmented texts was calculated. Summary statistics for the cosine similarity are presented in Table A3. The overall mean and median cosine similarity between the original and all augmented texts were 0.85 and 0.90, respectively (SD = 0.16). Specifically, the mean and median cosine similarity between the original text and character-level augmented texts were 0.88 and 0.91, respectively (SD = 0.10); for word-level augmented texts, the mean and median were 0.92 and 0.94 (SD = 0.08); and for LLM-based augmented texts, they were 0.76 and 0.83 (SD = 0.20). Overall, the results indicate that the augmented texts closely resemble the original texts, ensuring that the texts used in the downstream fine-tuning of the LLM are representative of the original content.

LLM Fine-Tuning

Subsequently, we fine-tuned LLMs for text classification tasks to identify and categorize outcomes related to knowledge construction and mental health behavioral intentions. As shown in Figure A2, fewer than 10 comments were labeled as mental health misinformation, corrections to mental health information, mental health discrimination, or corrections addressing mental health-related discrimination. Due to the insufficient number of ground-truth labels for these four variables, we excluded them from the downstream analyses. In total, we fine-tuned 13 LLMs to facilitate text classification. All 13 outcomes were treated as binary classification problems. These included 2 outcomes related to filtering questions (i.e., whether a comment tagged another user or contained only emojis), 6 outcomes related to knowledge construction, and 5 outcomes focused on mental health behavioral intentions (refer to: "Outcome Measurement"). Fine-tuning was performed on the un-augmented text for the two filtering outcomes and on the augmented text for the remaining

11 outcomes. We ensured a balanced class distribution to avoid issues arising from unbalanced samples when using the augmented text by maintaining a roughly equal ratio of comments from both classes.

For each outcome, we fine-tuned the following models: BERT (bert-base-uncased) (Devlin et al., 2019), RoBERTa (roberta-base) (Liu et al., 2019), MentalBERT (mental/mental-bert-base-uncased) (Ji et al., 2021), and MentalRoBERTa (mental/mental-roberta-base) (Ji et al., 2021). The models were trained with a batch size of 32, for 20 epochs, and a learning rate of 2e-5, following the recommendations from the BERT developers (Devlin et al., 2019). Among the models, the BERT-based LLM consistently achieved the highest accuracy for each outcome, leading us to use the fine-tuned BERT models for downstream analyses and predictions. The performance of the BERT-based LLMs for each of the 12 outcomes, as well as the specific epoch at which each model achieved its highest accuracy, is detailed in Table A4.

Finally, we applied the fine-tuned LLMs to label all comments for Case 1, Case 2, and Case 3 in the study. A total of 23,314 comments (12.39%) were excluded from the downstream analysis because they consisted of single words, contained only emojis, or solely tagged other users, as shown in Figure A1. The distribution of the outcomes in Case 1 predicted by the fine-tuned LLMs are presented in Figure A7.

Analytical Strategy

We aimed to assess whether comments on videos produced by MHCCs across the MO, CM, and control groups were more likely to demonstrate mental health knowledge construction and include mental health behavioral intentions. To achieve this, we employed three-level multilevel Linear Probability Models (LPM) with random effects at the influencer and video levels to account for potential asymmetries. The analysis included 164,855 comments nested within 1,863 videos created by 49 influencers. These models assessed the impact of our study interventions on the outcomes by interacting dichotomous fixed effect indicators of treatment group assignment (MO vs. CM, with the control group as the reference) with the fixed effect indicator of whether each video was produced pre- or post-intervention (March 2023 as the pre-period, and April and May 2023 as the post-period). It is important to note that, because the interventions were implemented in April 2023, including April in the analysis may lead to an underestimation of the treatment effect size. To ensure the robustness of our findings, we also ran an alternative set of models with the same model specification but excluding comments to videos produced in April 2023. The results from these models were consistent with those presented, as detailed in the supplementary materials. The results are presented in Table 1.

To further explore how the effects of the toolkit and conference on EBMHC influenced user comments, we introduced four moderators: engagement rate (greater than 10% or not), follower count (greater than 750k or not), whether the creator is a licensed practitioner, and whether the creator provides coaching. We incorporated three-way interactions between dichotomous fixed effect indicators of treatment group assignment (MO vs. CM, with the control group as the reference), the fixed effect indicator of video production period (pre- vs. post-intervention), and the dichotomized moderator variables. The results of these analyses are shown in Table 2.

Standardized coefficient estimates are reported to facilitate cross-study comparisons, while unstandardized coefficients are available in the supplementary files. The analyses were conducted using the lme4 package in R version 4.2.1 (Bates et al., 2024).

Case 2: TikTok Mental Health Awareness Month (MHAM) Program

In the second instance, we utilized fine-tuned large language models (LLMs) to evaluate the effects of TikTok's Mental Health Awareness Month (MHAM) initiatives on the principal outcomes observed in the comments (TikTok, 2023). May 2023 was designated as Mental Health Awareness Month on TikTok, during which the platform featured several events aimed at promoting mental well-being, combating stigma, and offering support to its community. Specifically, during MHAM, TikTok collaborated with seven institutional content creators, including the Alliance for Eating Disorders (@alliancefored), to launch the Mental Health Media Education Fund and donate over $2 million in ad credits to organizations dedicated to supporting mental health. Additionally, TikTok introduced a #MentalHealthAwareness hub that provided updated, supportive mental health content throughout May and spotlighted 10 creators who have used the platform to educate the community on mental health awareness, significantly impacting the discourse.

To assess how MHAM influenced mental health communication via video comments, we gathered videos posted from January 1, 2023, to August 1, 2023, within the US, tagged with #MentalHealthAwareness, using TikTok's research API. This collection was region-specific to the US, reflecting the localized nature of TikTok's MHAM. From an initial pool of 139,251 comments across 11,479 videos by 8,664 users, we refined the dataset by excluding comments that were single words, contained only emojis, or merely tagged other users. This left 88,979 comments from 7,014 videos by 5,597 users for further analysis, to which we applied the fine-tuned LLMs as in the first case. The outcomes from Case 2, as predicted by the fine-tuned LLMs, are depicted in Figure A8.

We employed a sharp regression discontinuity design (RDD) to estimate MHAM's impact on mental health communication outcomes in the comments. Outcomes were aggregated hourly, with two model specifications: one capturing immediate post-MHAM changes using May 1, 00:00, as the cutoff, and another comparing the period from May 1–30 with data from April to June, to evaluate sustained effects.

Given the sensitivity of our treatment effect estimators in regression discontinuity designs to both bandwidth and polynomial order, we took multiple measures to ensure robustness (Gelman & Imbens, 2019). All models were estimated using the rdrobust package in R, which utilizes triangular kernel weights and a data-driven approach for optimal bandwidth selection. We tested polynomial orders from 1 to 4, with the rdrobust package, as proposed by (Calonico et al., 2015), and employed the RDMSE methods by (Pei et al., 2020), to compute the asymptotic mean squared error for each local polynomial regression discontinuity. The computation was conducted in STATA. The polynomial order that minimized the AMSE was selected for the final model. The results, based on this selection, are reported in Table 3, which details the impact of TikTok’s MHAM on video engagement and online mental health communication. Specifications for these models, including bandwidth selection, polynomial orders, and the effective sample sizes for point and bias-correction estimators, are documented in Table A5. The RD estimation plots for knowledge construction are presented in Figure A9.

Case 3: MHAM Spotlights

In the third case, we utilized the fine-tuned large language models (LLMs) to assess the impact of TikTok's MHAM spotlight on the principal outcomes observed in the comments. As noted above, a total of 17 institutions and creators—comprising seven institutional and ten individual creators—were spotlighted by TikTok during MHAM in 2023. To examine the influence of these institution and creator spotlights on mental health communication through video comments, we used the TikTok Research Tools API to access videos created by the institutions and creators listed on TikTok's 2023 MHAM webpage from January 1 to August 31, 2023. This resulted in the collection of 169,628 comments from 1,489 videos. After filtering out comments that consisted of only single words, emojis, or tags of other users, 139,473 comments from 1,321 videos remained for further analysis. We applied the fine-tuned LLMs to annotate the retrieved video comments, as in Case 2. The distribution of the outcomes from Case 3, as predicted by the fine-tuned LLMs, are illustrated in Figure A10.

Adopting the analytical strategy from Case 2, we once again implemented a sharp regression discontinuity design (RDD) to estimate the impact of TikTok's MHAM institution and creator spotlights on mental health communication outcomes in the comments. The model selection and specifications were identical to those employed in Case 2. The results, based on this model selection, are presented in Table 4, which details the impact of TikTok's MHAM spotlights on video engagement and online mental health communication. Details of these models, including bandwidth selection, polynomial orders, and the effective sample sizes for point and bias-correction estimators, are documented in Table A6. The RD estimation plots for knowledge construction are shown in Figure A11.

Reference

Alex Montero, Grace Sparks, Marley Presiado, & Liz Hamel. (2024, June 13). *KFF Health Misinformation Tracking Poll: Health and Election Issues on TikTok*. KFF. https://www.kff.org/health-misinformation-and-trust/poll-finding/kff-healthmisinformation-tracking-poll-health-and-election-issues-on-tiktok/

Basch, C. H., Donelle, L., Fera, J., & Jaime, C. (2022). Deconstructing TikTok Videos on Mental Health: Cross-sectional, Descriptive Content Analysis. *JMIR Formative Research*, *6*(5), e38340. https://doi.org/10.2196/38340

Bates, D., Maechler, M., Bolker [aut, B., cre, Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., simulate.formula), P. N. K. (shared copyright on, Tanaka, E., & Jagan, M. (2024). *lme4: Linear Mixed-Effects Models using "Eigen" and S4* (Version 1.1-35.5) [Computer software]. https://cran.r-project.org/web/packages/lme4/index.html

Bhandari, A., & Bimo, S. (2022). Why's Everyone on TikTok Now? The Algorithmized Self and the Future of Self-Making on Social Media. *Social Media + Society*, *8*(1), 20563051221086241. https://doi.org/10.1177/20563051221086241

Bizzotto, N., Schulz, P. J., & Bruijn, G.-J. de. (2023). The "Loci" of Misinformation and Its Correction in Peer- and Expert-Led Online Communities for Mental Health: Content Analysis. *Journal of Medical Internet Research*, *25*(1), e44656. https://doi.org/10.2196/44656

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2015). rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs. *The R Journal*, *7*(1), 38–51.

Chen, J., & Wang, Y. (2021). Social Media Use for Health Purposes: Systematic Review. *Journal of Medical Internet Research*, *23*(5), e17917. https://doi.org/10.2196/17917

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Emily A. Vogels, Risa Gelles-Watnick, & Navid Massarat. (2022, August 10). *Teens, Social Media and Technology 2022*. Pew Research Center. https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/

Gelman, A., & Imbens, G. (2019). Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, *37*(3), 447–456. https://doi.org/10.1080/07350015.2017.1366909

Gunawardena, C., Flor, N., Gomez, D., & Sanchez, D. (2016). Analyzing Social Construction of Knowledge Online by Employing Interaction Analysis, Learning Analytics, and Social Network Analysis. *The Quarterly Review of Distance Education*, *17*(3), 35–60.

Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a Global Online Debate and the Development of an Interaction Analysis Model for Examining Social Construction of Knowledge in Computer Conferencing. *Journal of Educational Computing Research*, *17*(4), 397–431. https://doi.org/10.2190/7MQV-X9UJ-C7Q3-NRAG

Haythornthwaite, C., Kumar, P., Gruzd, A., Gilbert, S., Esteve del Valle, M., & Paulin, D. (2018). Learning in the wild: Coding for learning and practice on Reddit. *Learning, Media and Technology*, *43*(3), 219–235. https://doi.org/10.1080/17439884.2018.1498356

Jacobs, A., Pan, Y.-C., & Ho, Y.-C. (2022). More than just engaging? TikTok as an effective learning tool. *UK Academy for Information Systems Conference Proceedings 2022*. https://aisel.aisnet.org/ukais2022/3

Janz, N. K., & Becker, M. H. (1984). The Health Belief Model: A decade later. *Health Education Quarterly*, *11*(1), 1–47. https://doi.org/10.1177/109019818401100101

Jeffrey Gottfried. (2024, January 31). *Americans' Social Media Use*. Pew Research Center.

https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/#:~:text=About%20three%2Dquarters%20of%20adults,for%20each%20site%20and%20app

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2021). *MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare* (arXiv:2110.15621). arXiv. https://doi.org/10.48550/arXiv.2110.15621

Limbu, Y. B., Gautam, R. K., & Pham, L. (2022). The Health Belief Model Applied to COVID-19 Vaccine Hesitancy: A Systematic Review. *Vaccines*, *10*(6), 973. https://doi.org/10.3390/vaccines10060973

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Milton, A., Ajmani, L., DeVito, M. A., & Chancellor, S. (2023). "I See Me Here": Mental Health Content, Community, and Algorithmic Curation on TikTok. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3544548.3581489

Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP* (arXiv:2005.05909). arXiv. https://doi.org/10.48550/arXiv.2005.05909

Motta, M., Liu, Y., & Yarnell, A. (2024). "Influencing the influencers:" a field experimental approach to promoting effective mental health communication on TikTok. *Scientific Reports*, *14*(1), 5864. https://doi.org/10.1038/s41598-024-56578-1

Nguyen, H., & Diederich, M. (2023). Facilitating knowledge construction in informal learning: A study of TikTok scientific, educational videos. *Computers & Education*, *205*, 104896. https://doi.org/10.1016/j.compedu.2023.104896

Nguyen, V. C., Jain, M., Chauhan, A., Soled, H. J., Lesmes, S. A., Li, Z., Birnbaum, M. L., Tang,

S. X., Kumar, S., & De Choudhury, M. (2024). *Supporters and Skeptics: LLM-based Analysis of Engagement with Mental Health (Mis)Information Content on Video-sharing Platforms* (arXiv:2407.02662). arXiv. https://doi.org/10.48550/arXiv.2407.02662

NIH. (2023, January 12). *How To Find Reliable Health Information Online*. National Institute on Aging. https://www.nia.nih.gov/health/healthy-aging/how-find-reliable-health-information-online

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Pei, Z., Lee, D. S., Card, D., & Weber, A. (2020). *Local Polynomial Order in Regression Discontinuity Designs* (Working Paper 27424). National Bureau of Economic Research. https://doi.org/10.3386/w27424

Pourazad, N., Stocchi, L., & Narsey, S. (2023). A Comparison of Social Media Influencers' KPI Patterns across Platforms: Exploring Differences in Followers and Engagement On Facebook, Instagram, YouTube, TikTok, and Twitter. *Journal of Advertising Research*, *63*(2), 139–159. https://doi.org/10.2501/JAR-2023-008

Shahbaznezhad, H., Dolan, R., & Rashidirad, M. (2021). The Role of Social Media Content Format and Platform in Users' Engagement Behavior. *Journal of Interactive Marketing*, *53*, 47–65. https://doi.org/10.1016/j.intmar.2020.05.001

Starvaggi, I., Dierckman, C., & Lorenzo-Luaces, L. (2024). Mental health misinformation on social media: Review and future directions. *Current Opinion in Psychology*, *56*, 101738. https://doi.org/10.1016/j.copsyc.2023.101738

TikTok. (n.d.). *Research API | TikTok for Developers*. Retrieved April 15, 2024, from https://developers.tiktok.com/products/research-api/

TikTok. (2023, May 15). *Supporting our Community this #MentalHealthAwareness Month and*

Year Round. https://newsroom.tiktok.com/en-us/mentalhealthawarenessmonth-2023

USC. (2024). *Mental Health Storytelling Initiative*. https://annenberg.usc.edu/research/annenberg-inclusion-initiative/mental-health-storytelling-initiative

van Driel, I. I., Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and Pitfalls of Social Media Data Donations. *Communication Methods and Measures*, *16*(4), 266–282. https://doi.org/10.1080/19312458.2022.2109608

Wang, X., & Cohen, R. A. (2023). *Health Information Technology Use Among Adults: United States, July–December 2022*. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Wong, K. L. (2022). *A new way of winning customers on social media: An examination of storytelling factors of brands' Instagram and Tiktok short videos*. https://theses.lib.polyu.edu.hk/handle/200/12222

Xu, M. M., Scherer, E., Robbins, R., Liu, Y., Motta, M., & Yarnell, A. (2024). *'Influencing the influencers': Exploring the impact of an in-person summit on TikTokers' mental health communication habits and beliefs*. OSF. https://doi.org/10.31235/osf.io/v83ht

Yıldırım, S. (2023). The Challenge of Self-diagnosis on Mental Health Through Social Media: A Qualitative Study. In G. Battineni, M. Mittal, & N. Chintalapudi (Eds.), *Computational Methods in Psychiatry* (pp. 197–213). Springer Nature. https://doi.org/10.1007/978-981-99-6637-0_10

YouTube. (n.d.). *THE-IQ Creator Program*. https://health.youtube/features-application/#:~:text=THE%2DIQ%20Creator%20Program%20supports,YouTube%20%E2%80%93%20with%20messages%20that%20resonate.

Appendix

Measurement	Definition	Example TikTok video comments
Knowledge construction		
Personal reflection	The comment contain statements of the commenter's personal experience, personal thoughts, or examples	"Omg! I woke last night with the most uncomfortable sensation of not being myself. I was a completely separate observer of this person running in auto"
Express agreement	The comment contain statements expressing agreement with the video content	"This is so true - we are living in our own illusion! Now to commit the movie right in front of me… ❤️ 🔮 "
Express disagreement	The comment contain statements expressing disagreement with the video content	"ooh I agree with much of what you say except for fact that the reason why people escape to the mind movie is because they'r tryescape real-world prob"
Ask clarification questions	The comment ask for more clarification or information regarding the video content, or express confusion	"I still do not understand the reason for physical illusion or purpose"
Reinterpret knowledge	The commenter uses the information in the video to clarify or challenge their own knowledge or experience. Comments in this category meet two criteria: 1) The comment should provide an interpretation or analysis of the opinions or ideas presented within the content of the video. 2) There must be a discernible link between these ideas or opinions and the knowledge or experience	"THIS ANALOGY 🤧 my instinct was to always fight my feelings instead of surrendering, learning to let go and feel the rain - thank u for this!!!"
Apply knowledge to new area	The commenter applies knowledge gained from the video to develop new ideas or opinions in another area. Comments that fall into this category should: 1) provide an interpretation or analysis of the opinions or ideas presented within the content of the video. 2) develop new ideas or opinions based on the interpretation in 1).	(In a video of resisting emotion) "We can get so caught up in trying to "feel right" that it becomes a compulsion. There is no "feeling right" there's existing. And existing is experiencing emotions, feelings, thoughts. And not judging ourselves for them. Rather embracing that that is what makes us human."
Mental health misinformation		

Contain mental health misinformation	The comment contain misinformation on mental health or well-being.	“People with mental health issues can't work or lead normal lives”
Contain MH misinformation correction	The comment states the misinformation and contains correction of the misinformation about mental health or well-being.	“it is incorrect to say ADHD is rare. Believing incorrect information isn’t a good idea. Even if it is tempting. 🤷\u200d♀️ 🤦\u200d♀”
Mental health discrimination	The comment contains discrimination or stigma on mental health or well-being-related topics as a self-experience or general narrative. Comments that fall into this category may contain language that discriminates against people with mental health problems, or language expressing the stigma of mental health problems faced by individuals. Stigma related to race, gender, or other factors is not included in this category.	“People with depression are just not strong enough. I feel so bad about my weakness.”
Mental health discrimination correction	The comment indicates the commenter is coping with or correcting the discrimination or stigma related to mental health or well-being. Comments that fall into this category may contain language that counters discrimination or stigma against people with mental health problems, or disclose that they are countering such discrimination or stigma in daily life. Stigma related to race, gender, or other factors is not included in this category.	“I have BPD and I’m a cop and honestly I find it makes me better at my job. So like fuck the stigma. ❤️”
Mental health behavioral intentions		
Disclose self-mental health problem	The comments contain disclosure of having mental health related problems.	“Troubled by ED for years and years and i do not see it coming away.”
Seek for wellbeing-related information	The commenter seeks information about well-being, wellness, therapy, or any other well-being related content.	“My therapist does not get my, should I change to a new one?”
Share wellbeing-related coping strategy	The comment contains coping strategies for emotion or well-being-related issues, such as mindfulness, positive thinking, or other personal solutions.	“i try to accomplish at least one thing just one thing sometimes leads to 2 makes me feel a little better”

Disclose getting wb-related professional service	The comment indicates that the commenter is seeking professional assistance or planning to seek professional assistance, from therapist, life coach, religious leaders, or psychiatrist.	“I’m going to start calling my therapist shorty 😂”
Advocate others for seeking wb-related professional service	The commenter encourages other commenters, or specifically tags others using '@', to seek professional assistance, from therapist, life coach, religious leaders, or psychiatrist.	“@auuuu you should have some therapy sessions for your wife to help support but not fall for the lies and manipulation”

Table A1. The codebook for the outcomes in the current study

	Percent Agreement	Gwett’s AC
By coding sections		
All questions	0.81	0.80
Filtering questions	0.87	0.86
Knowledge construction questions	0.72	0.69
Mental Health questions	0.81	0.80
By coding questions		
Tagging only	0.99	0.99
Express appreciation	0.82	0.78
Emoji only	0.85	0.83
Single-word comment	0.85	0.83
Disclose mental health problem	0.87	0.85
Seek for wellbeing-related information	0.87	0.84
Knowledge_Personal reflection	0.73	0.61
Knowledge_Express agreement	0.68	0.48
Knowledge_Express disagreement	0.80	0.76
Knowledge_Ask clarification questions	0.82	0.78
Knowledge_Reinterpret knowledge	0.68	0.55
Knowledge_Apply knowledge to new area	0.75	0.70
Share wellbeing-related coping strategy	0.77	0.71
Disclose receiving well-being-related professional service	0.80	0.76
Advocate others to seek for mental health professional help	0.82	0.80
Contain mental health misinformation	0.82	0.79
Contain correction of mental health misinformation	0.83	0.77
Contain mental health-related discrimination	0.83	0.77
Contain correction of mental health -related discrimination	0.83	0.79

Table A2. The inter-coder reliability assessment results in the current study

	Overall	Character-level	Word-level	LLM-based
Mean	0.85	0.88	0.92	0.76
Median	0.90	0.91	0.94	0.83
SD	0.16	0.10	0.08	0.20
25th Percentile	0.81	0.84	0.89	0.70
75th Percentile	0.95	0.95	0.97	0.90
5th Percentile	0.50	0.67	0.72	0.29
95th Percentile	0.99	0.99	0.99	0.95

Table A3. The summary statistics of the cosine similarity between the original TikTok comment text and the augmented texts through character-level, word-level, and LLM-based text augmentation

Note: text was embedded using sentence_transformer with “bert-base-nli-mean-tokens” model.

	Epoch of best model	N=0	N=1	Precision	Recall	F1 Score	Accuracy
Knowledge construction							
General	4	4100	4548	0.95	0.95	0.95	0.95
Personal reflection	5	2748	2739	0.95	0.95	0.95	0.95
Express agreement	5	2169	2213	0.94	0.94	0.94	0.94
Express disagreement	10	328	338	0.91	0.90	0.90	0.90
Ask clarification questions	6	400	413	0.93	0.93	0.93	0.93
Reinterpret knowledge	5	2458	2559	0.95	0.95	0.95	0.95
Apply knowledge to new area	10	1794	1822	0.95	0.95	0.95	0.95
Online interaction							
Express appreciation to the creator	7	2087	2095	0.92	0.92	0.92	0.92
Seek for wellbeing-related information	8	410	400	0.95	0.94	0.94	0.94
Mental health-related							
Disclose self-mental health problem	4	298	299	0.94	0.94	0.94	0.94
Disclose receiving wb-related professional service	3	104	94	0.98	0.98	0.98	0.98
Share wellbeing-related coping strategy	10	341	335	0.94	0.94	0.94	0.94
Text-preprocessing							
Emoji only	1	619	32	0.99	0.99	0.99	0.99
Tagging only	2	741	24	0.99	0.99	0.99	0.99

Table A4. Performance of fine-tuned LLMs in classifying comments into study outcomes

	Effect of MHA on health communication on May (May 1 00:00 as cutoff)			Effect of MHA on health communication on June (May 1 -- 30 as fuzzy)		
	Band width	Polynomial order	Effective sample size	Band width	Polynomial order	Effective sample size
Video engagement						
Like count (log)	hl=620.39, hr=702.53, bl=936.32, br=1034.97	p=1, q=2	N_hl=283, N_hr=474, N_bl=435, N_br=660	hl=703.88, hr=703.88, bl=1015.48, br=1018.76	p=1, q=2	N_hl=302, N_hr=407, N_bl=489, N_br=565
Share count (log)	hl=922.31, hr=984.44, bl=1514.15, br=1536.80	p=1, q=2	N_hl=423, N_hr=627, N_bl=768, N_br=959	hl=756.24, hr=744.70, bl=1098.17, br=1080.82	p=1, q=2	N_hl=334, N_hr=430, N_bl=526, N_br=599
View count (log)	hl=513.92, hr=599.55, bl=787.03, br=912.03	p=1, q=2	N_hl=237, N_hr=410, N_bl=357, N_br=593	hl=582.90, hr=592.20, bl=911.36, br=977.55	p=1, q=2	N_hl=263, N_hr=357, N_bl=414, N_br=545
Comment count (log)	hl=790.45, hr=821.02, bl=1218.01, br=1278.62	p=1, q=2	N_hl=351, N_hr=536, N_bl=589, N_br=816	hl=770.83, hr=707.41, bl=1257.94, br=1039.83	p=1, q=2	N_hl=333, N_hr=412, N_bl=619, N_br=581
Knowledge construction						
Total_binary	hl=773.14, hr=805.90, bl=1187.25, br=1234.62	p=1, q=2	N_hl=362, N_hr=567, N_bl=613, N_br=843	hl=392.38, hr=289.91, bl=667.92, br=664.58	p=1, q=2	N_hl=186, N_hr=171, N_bl=299, N_br=417
Total_continuous	hl=412.55, hr=412.55, bl=844.69, br=844.69	p=1, q=2	N_hl=198, N_hr=311, N_bl=399, N_br=594	hl=417.27, hr=233.46, bl=709.83, br=567.03	p=1, q=2	N_hl=203, N_hr=138, N_bl=322, N_br=363
Personal reflection	hl=487.43, hr=708.36, bl=1009.28, br=1135.42	p=1, q=2	N_hl=245, N_hr=512, N_bl=511, N_br=786	hl=397.85, hr=349.96, bl=800.41, br=679.17	p=2, q=3	N_hl=188, N_hr=218, N_bl=381, N_br=429
Express agreement	hl=404.59, hr=436.62, bl=741.65, br=748.27	p=1, q=2	N_hl=191, N_hr=333, N_bl=339, N_br=544	hl=400.18, hr=527.47, bl=808.70, br=822.87	p=1, q=2	N_hl=188, N_hr=334, N_bl=383, N_br=496
Express disagreement	hl=850.02, hr=809.56, bl=1355.03, br=1357.55	p=1, q=2	N_hl=401, N_hr=570, N_bl=722, N_br=921	hl=899.18, hr=783.62, bl=1383.78, br=1313.24	p=1, q=2	N_hl=428, N_hr=478, N_bl=737, N_br=770
Ask clarification questions	hl=852.16, hr=976.78, bl=1344.38, br=1468.50	p=1, q=2	N_hl=401, N_hr=676, N_bl=715, N_br=985	hl=656.06, hr=518.04, bl=1024.54, br=933.68	p=1, q=2	N_hl=299, N_hr=325, N_bl=520, N_br=553
Reinterpret knowledge	hl=489.66, hr=383.59, bl=921.07, br=907.39	p=1, q=2	N_hl=245, N_hr=285, N_bl=443, N_br=638	hl=568.11, hr=429.82, bl=890.03, br=776.00	p=1, q=2	N_hl=264, N_hr=273, N_bl=422, N_br=474
Apply knowledge to new area	hl=491.12, hr=491.12, bl=918.33, br=918.33	p=1, q=2	N_hl=245, N_hr=362, N_bl=440, N_br=648	hl=664.91, hr=559.24, bl=982.75, br=909.62	p=1, q=2	N_hl=299, N_hr=358, N_bl=490, N_br=543

Mental health behavioral intention						
Express appreciation to the creator	hl=398.08, hr=379.11, bl=839.57, br=839.57	p=1, q=2	N_hl=188, N_hr=281, N_bl=394, N_br=589	hl=429.14, hr=268.10, bl=734.12, br=652.75	p=1, q=2	N_hl=207, N_hr=155, N_bl=336, N_br=411
Seek for wellbeing-related information	hl=734.55, hr=785.04, bl=1231.32, br=1279.30	p=1, q=2	N_hl=336, N_hr=556, N_bl=638, N_br=876	hl=609.63, hr=559.95, bl=1103.60, br=929.79	p=1, q=2	N_hl=291, N_hr=358, N_bl=558, N_br=550
Disclose self-mental health problem	hl=205.30, hr=379.83, bl=577.87, br=804.17	p=1, q=2	N_hl=82, N_hr=281, N_bl=270, N_br=566	hl=196.87, hr=214.68, bl=610.06, br=519.54	p=1, q=2	N_hl=78, N_hr=137, N_bl=292, N_br=326
Disclose receiving wb-related professional service	hl=848.62, hr=770.25, bl=1270.73, br=1222.58	p=1, q=2	N_hl=400, N_hr=553, N_bl=667, N_br=837	hl=588.04, hr=490.33, bl=937.82, br=880.71	p=1, q=2	N_hl=277, N_hr=304, N_bl=455, N_br=528
Share wellbeing-related coping strategy	hl=608.34, hr=609.26, bl=926.58, br=953.96	p=1, q=2	N_hl=290, N_hr=453, N_bl=448, N_br=672	hl=596.47, hr=431.43, bl=960.67, br=752.81	p=1, q=2	N_hl=279, N_hr=275, N_bl=474, N_br=462

Table A5. Model specification of regression discontinuity models in Table 3 (Case 2)

	Effect of MHA on health communication on May (May 1 00:00 as cutoff)			Effect of MHA on health communication on June (May 1 -- 30 as fuzzy)		
	Band width	**Polynomial order**	**Effective sample size**	**Band width**	**Polynomial order**	**Effective sample size**
Video engagement						
Like count (log)	hl=850.06, hr=892.48, bl=1527.31, br=1630.52	p=1, q=2	N_hl=133, N_hr=161, N_bl=267, N_br=272	hl=589.63, hr=589.63, bl=910.89, br=797.04	p=1, q=2	N_hl=89, N_hr=75, N_bl=144, N_br=111
Share count (log)	hl=987.65, hr=1288.89, bl=1429.67, br=1730.91	p=1, q=2	N_hl=166, N_hr=213, N_bl=255, N_br=291	hl=583.90, hr=484.58, bl=832.92, br=742.13	p=1, q=2	N_hl=91, N_hr=58, N_bl=130, N_br=105
View count (log)	hl=812.95, hr=836.54, bl=1419.78, br=1564.57	p=1, q=2	N_hl=125, N_hr=157, N_bl=251, N_br=251	hl=547.32, hr=547.32, bl=927.42, br=826.53	p=1, q=2	N_hl=81, N_hr=71, N_bl=152, N_br=114
Comment count (log)	hl=957.06, hr=1316.62, bl=1457.02, br=1865.69	p=1, q=2	N_hl=160, N_hr=212, N_bl=264, N_br=307	hl=641.18, hr=601.23, bl=901.25, br=840.91	p=1, q=2	N_hl=93, N_hr=78, N_bl=145, N_br=120
Knowledge construction						
Total_binary	hl=997.49, hr=1206.10, bl=1473.18, br=1700.79	p=1, q=2	N_hl=174, N_hr=200, N_bl=274, N_br=303	hl=693.17, hr=689.55, bl=1173.20, br=1041.71	p=1, q=2	N_hl=105, N_hr=100, N_bl=223, N_br=171
Total_continuous	hl=624.45, hr=828.75, bl=1090.69, br=1360.16	p=1, q=2	N_hl=93, N_hr=164, N_bl=197, N_br=233	hl=453.48, hr=453.48, bl=877.95, br=786.50	p=1, q=2	N_hl=74, N_hr=52, N_bl=142, N_br=116
Personal reflection	hl=516.95, hr=529.95, bl=969.64, br=1024.68	p=1, q=2	N_hl=78, N_hr=109, N_bl=168, N_br=177	hl=406.95, hr=650.55, bl=775.59, br=1010.78	p=1, q=2	N_hl=65, N_hr=97, N_bl=119, N_br=164
Express agreement	hl=692.54, hr=773.74, bl=1159.24, br=1237.26	p=1, q=2	N_hl=105, N_hr=150, N_bl=221, N_br=210	hl=574.33, hr=574.33, bl=923.32, br=916.38	p=1, q=2	N_hl=89, N_hr=79, N_bl=155, N_br=151
Express disagreement	hl=954.04, hr=968.09, bl=1405.18, br=1602.25	p=1, q=2	N_hl=163, N_hr=176, N_bl=255, N_br=278	hl=854.92, hr=795.36, bl=1293.13, br=1276.26	p=1, q=2	N_hl=138, N_hr=120, N_bl=247, N_br=205
Ask clarification questions	hl=985.37, hr=1041.72, bl=1413.96, br=1550.23	p=1, q=2	N_hl=170, N_hr=177, N_bl=255, N_br=265	hl=874.44, hr=713.57, bl=1272.16, br=1037.09	p=1, q=2	N_hl=141, N_hr=108, N_bl=244, N_br=171
Reinterpret knowledge	hl=1046.28, hr=1046.28, bl=1574.79, br=1547.78	p=1, q=2	N_hl=185, N_hr=177, N_bl=277, N_br=265	hl=710.12, hr=679.60, bl=1027.01, br=980.84	p=1, q=2	N_hl=110, N_hr=98, N_bl=178, N_br=163
Apply knowledge to new area	hl=942.02, hr=1066.32, bl=1389.28, br=1548.80	p=1, q=2	N_hl=161, N_hr=177, N_bl=255, N_br=265	hl=500.98, hr=486.91, bl=784.10, br=759.98	p=1, q=2	N_hl=76, N_hr=65, N_bl=121, N_br=112
Mental health behavioral intention						

Express appreciation to the creator	hl=905.92, hr=1101.26, bl=1335.48, br=1597.64	p=1, q=2	N_hl=149, N_hr=179, N_bl=255, N_br=276	hl=662.51, hr=671.41, bl=1113.06, br=1113.06	p=1, q=2	N_hl=100, N_hr=97, N_bl=205, N_br=182
Seek for wellbeing-related information	hl=1180.29, hr=1083.74, bl=1783.64, br=1619.77	p=1, q=2	N_hl=226, N_hr=177, N_bl=330, N_br=285	hl=808.77, hr=772.93, bl=1154.21, br=1115.54	p=1, q=2	N_hl=127, N_hr=114, N_bl=220, N_br=182
Disclose self-mental health problem	hl=751.59, hr=804.60, bl=1388.90, br=1480.79	p=2, q=3	N_hl=116, N_hr=159, N_bl=255, N_br=256	hl=448.36, hr=325.67, bl=734.71, br=638.18	p=1, q=2	N_hl=74, N_hr=33, N_bl=111, N_br=96
Disclose receiving wb-related professional service	hl=613.98, hr=715.37, bl=1120.36, br=1264.82	p=1, q=2	N_hl=93, N_hr=133, N_bl=206, N_br=217	hl=617.77, hr=616.82, bl=1091.52, br=1259.02	p=1, q=2	N_hl=93, N_hr=89, N_bl=197, N_br=201
Share wellbeing-related coping strategy	hl=697.10, hr=697.10, bl=1162.18, br=1219.26	p=1, q=2	N_hl=105, N_hr=128, N_bl=222, N_br=205	hl=481.16, hr=506.67, bl=787.47, br=773.90	p=1, q=2	N_hl=74, N_hr=71, N_bl=121, N_br=114

Table A6. Model specification of regression discontinuity models in Table 4 (Case 3)

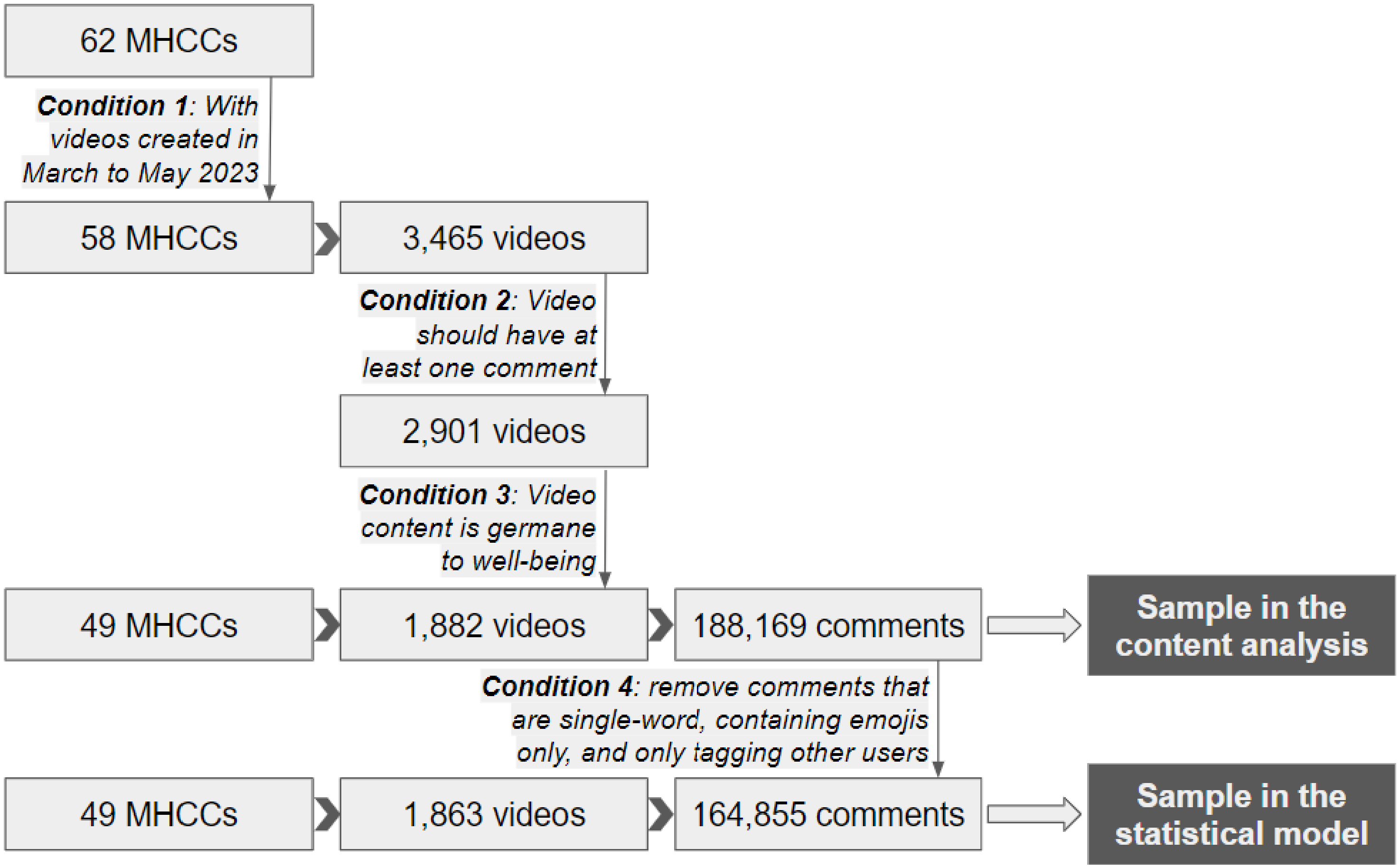


Figure A1. The process of data collection in Case 1

Note: MHCCs – mental health content creators on TikTok.

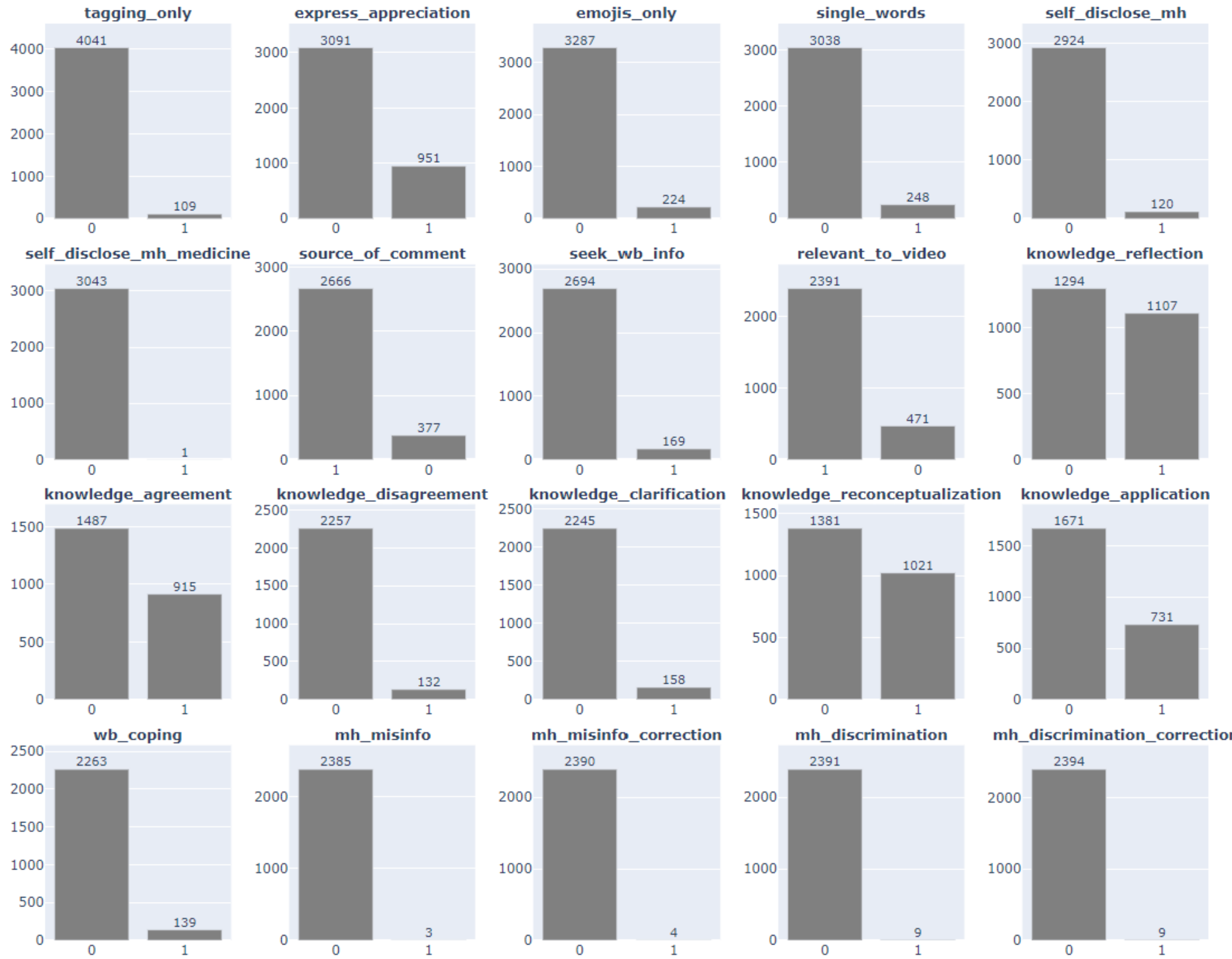


Figure A2. The distributions of the study outcomes from the content analyses

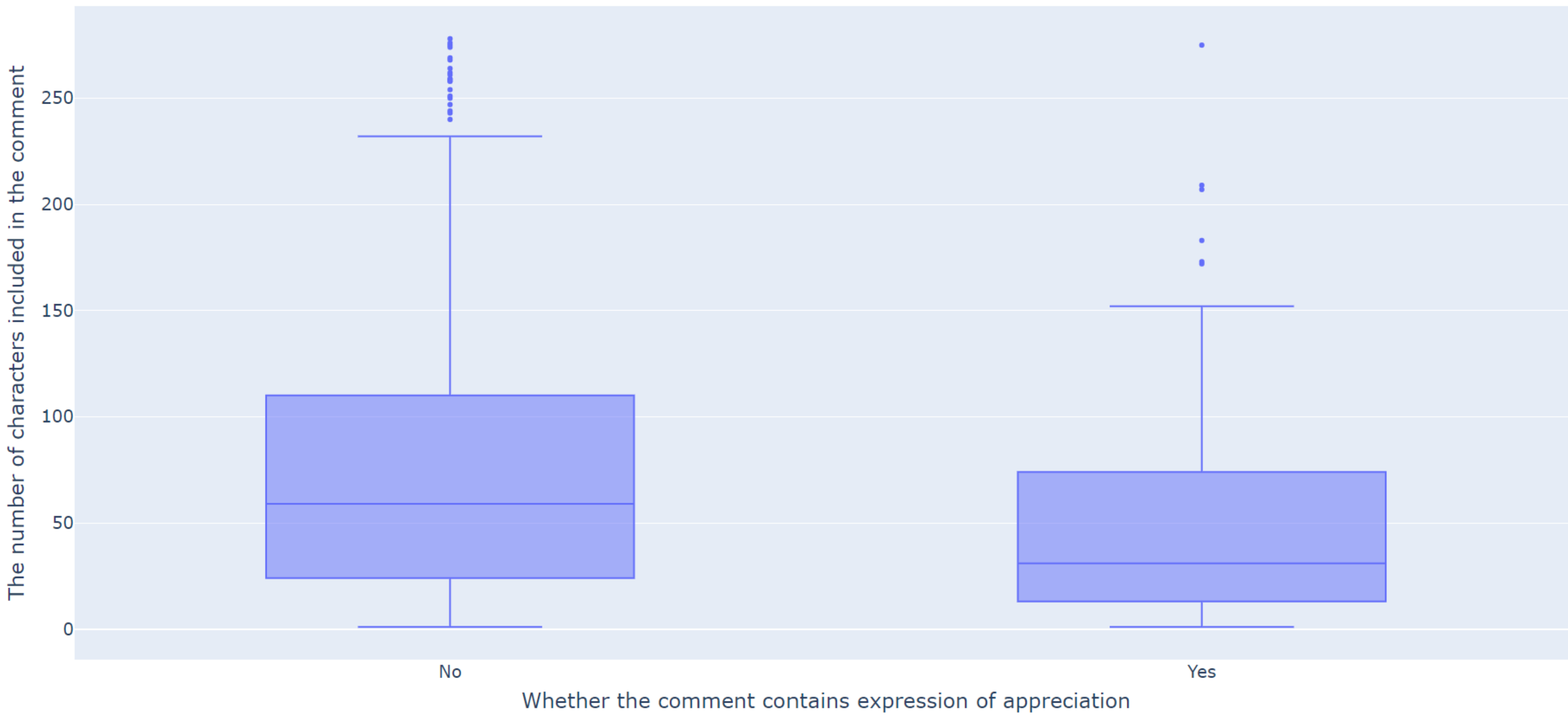


Figure A3. The distribution of character counts per comment based on expressions of appreciation

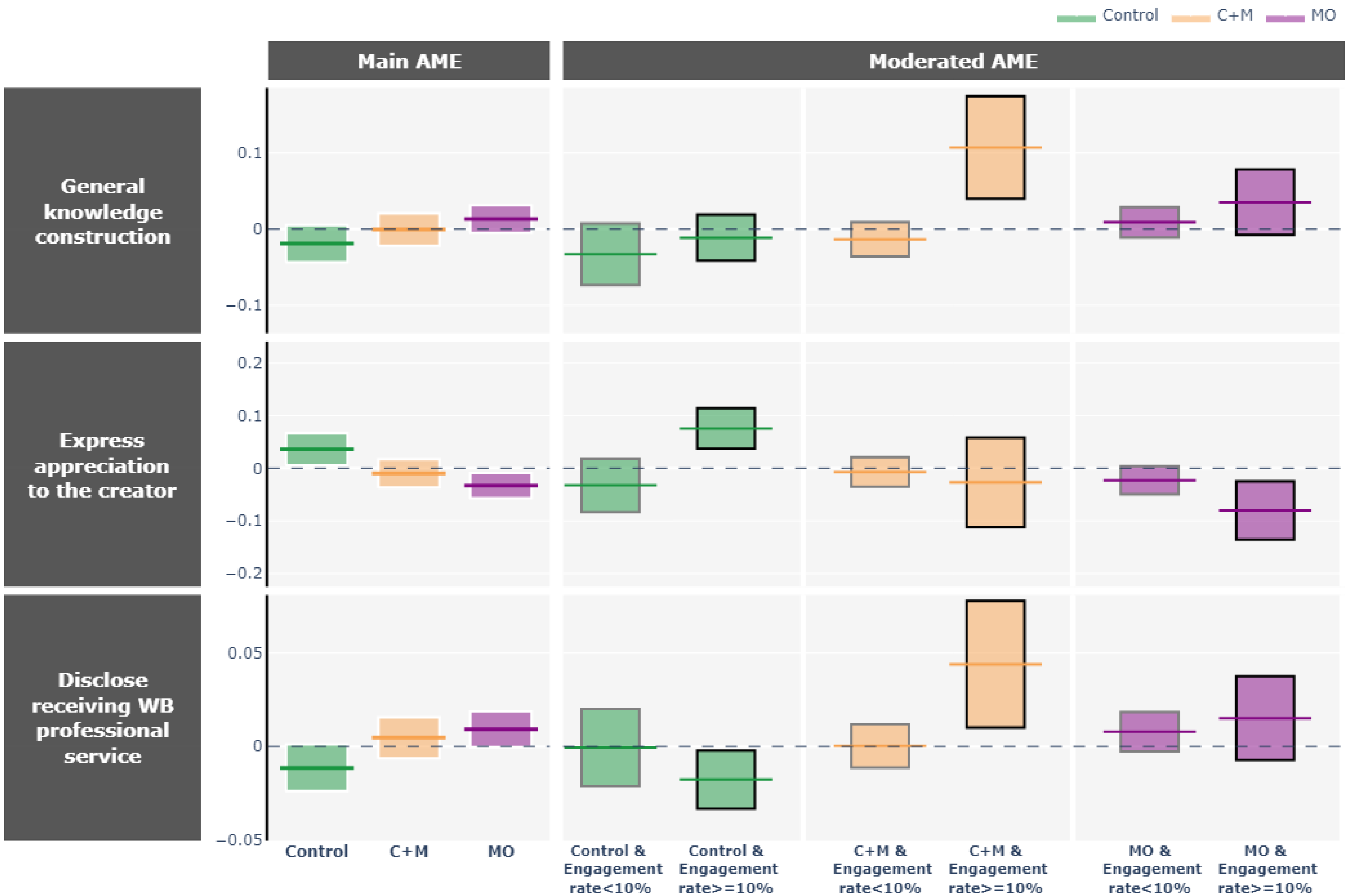


Figure A4. The average marginal effect (AME) and the moderated AME of treatment assignment on the probability of knowledge construction, express appreciation, and disclose receiving well-being related professional service, moderated by whether the creators have an engagement rate larger than 10%

Figure A5. The average marginal effect (AME) and the moderated AME of treatment assignment on the probability of knowledge construction, express appreciation, and disclose receiving well-being related professional service, moderated by whether the creators are licensed professional or not

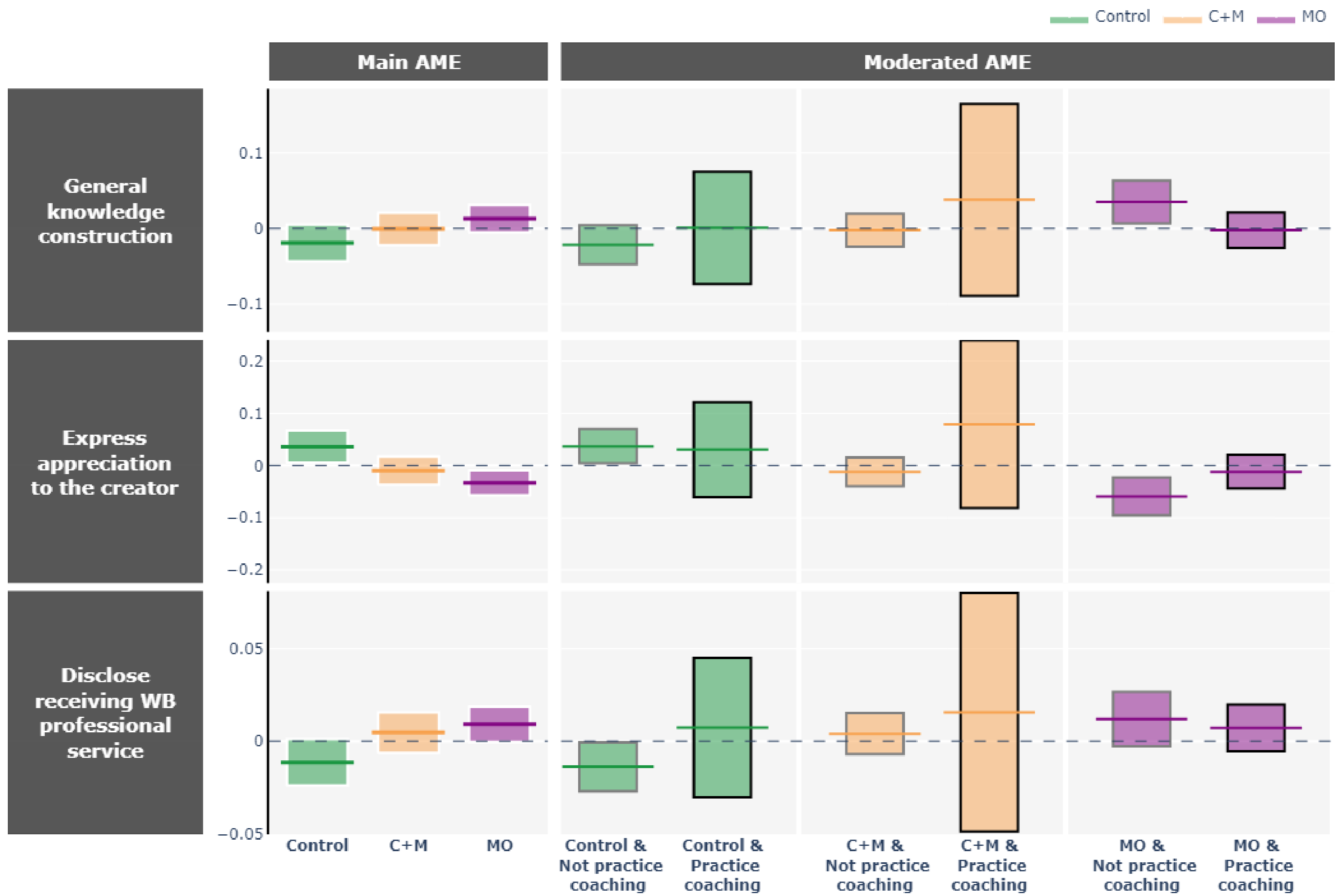


Figure A6. The average marginal effect (AME) and the moderated AME of treatment assignment on the probability of knowledge construction, express appreciation, and disclose receiving well-being related professional service, moderated by whether the creators offering paid coaching service or not

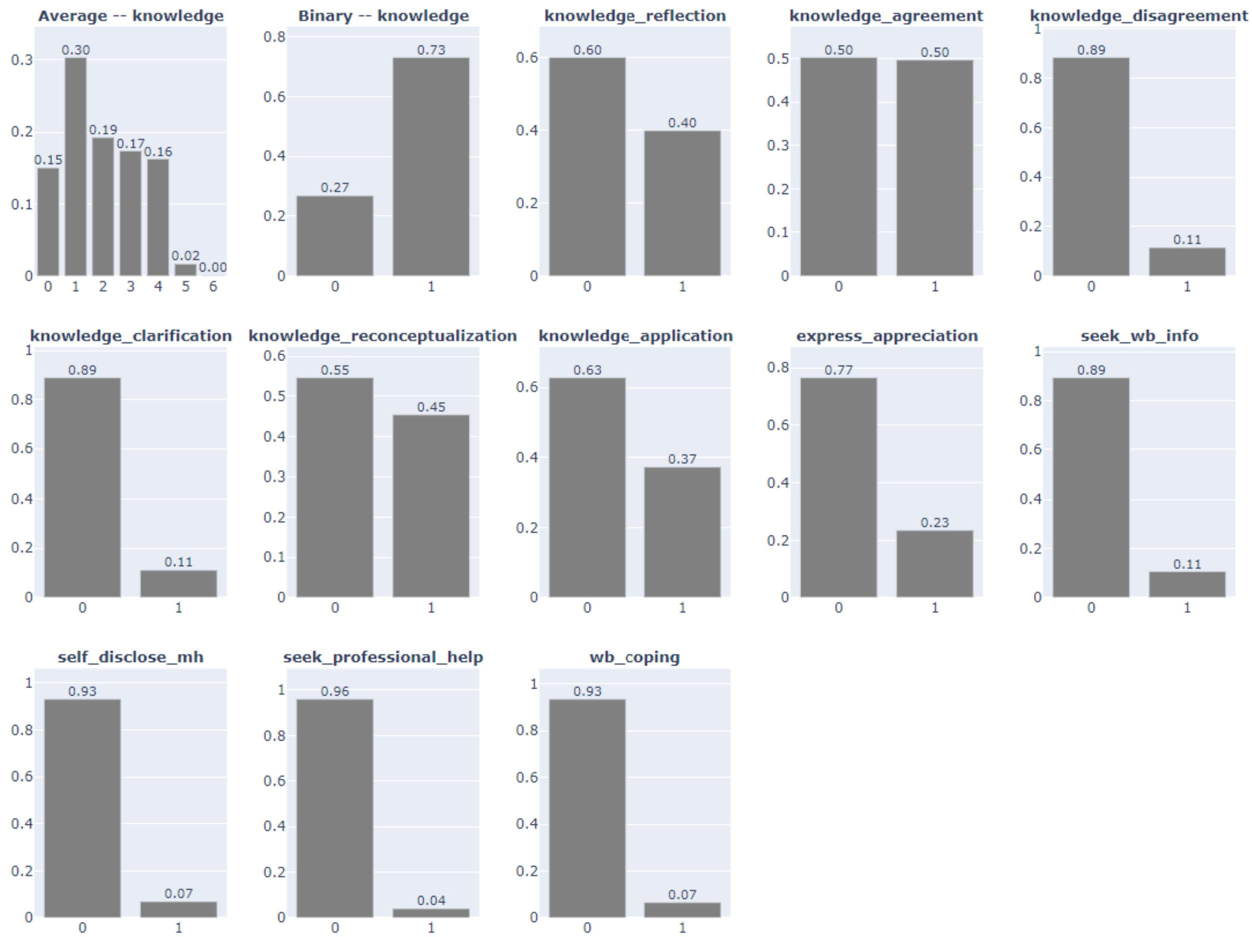


Figure A7. The distributions of the study outcomes among all TikTok comments included in Case 1.

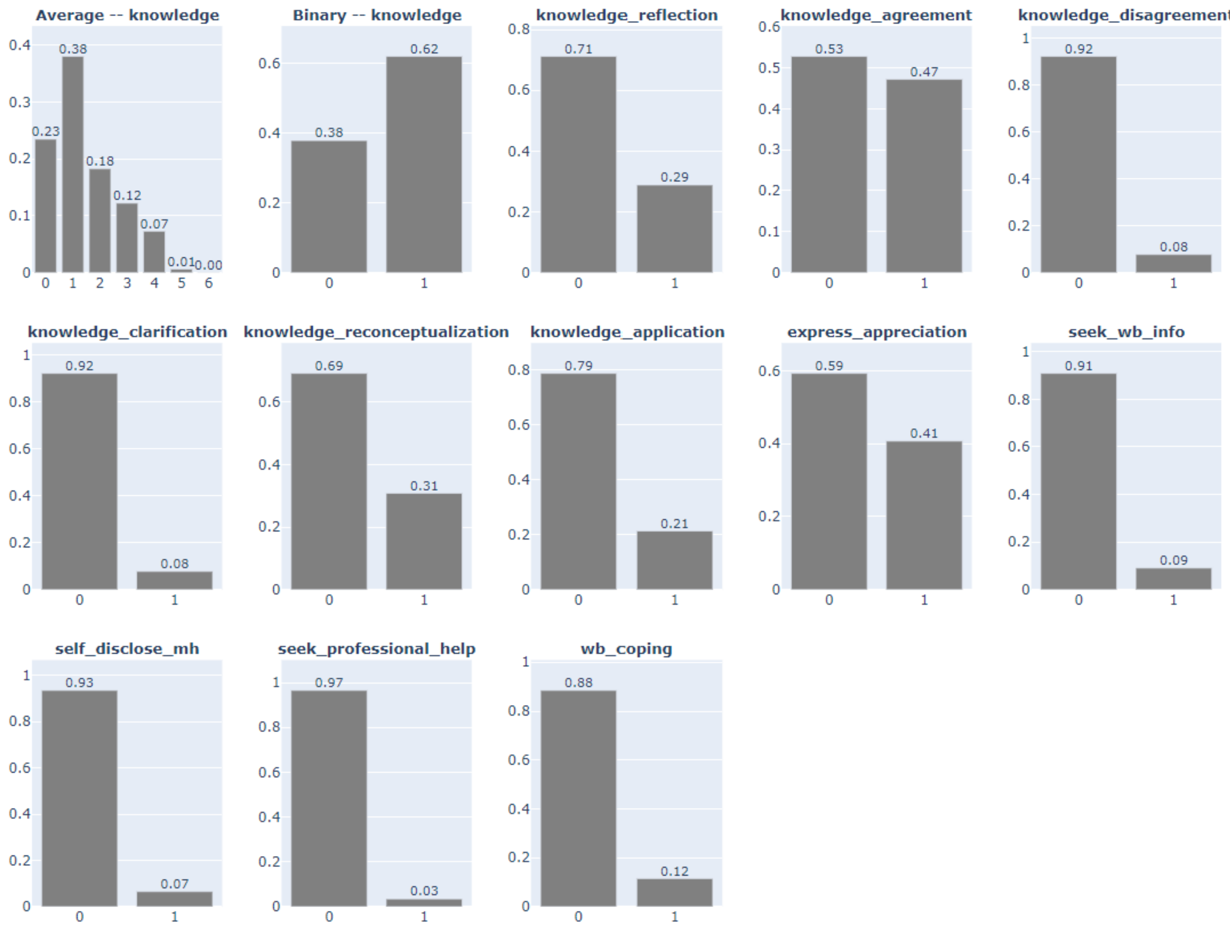
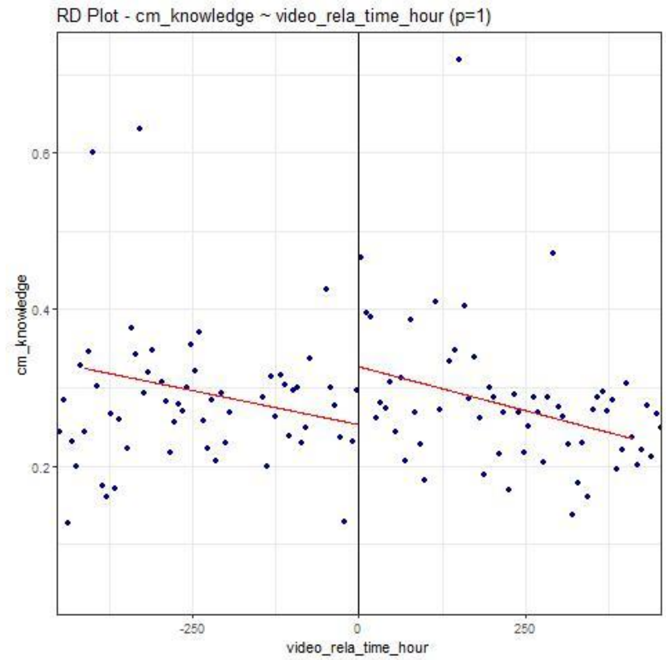


Figure A8. The distributions of the study outcomes among all TikTok comments included in Case 2.

(A) Keep data from May 2023



(B) Drop data from May 2023

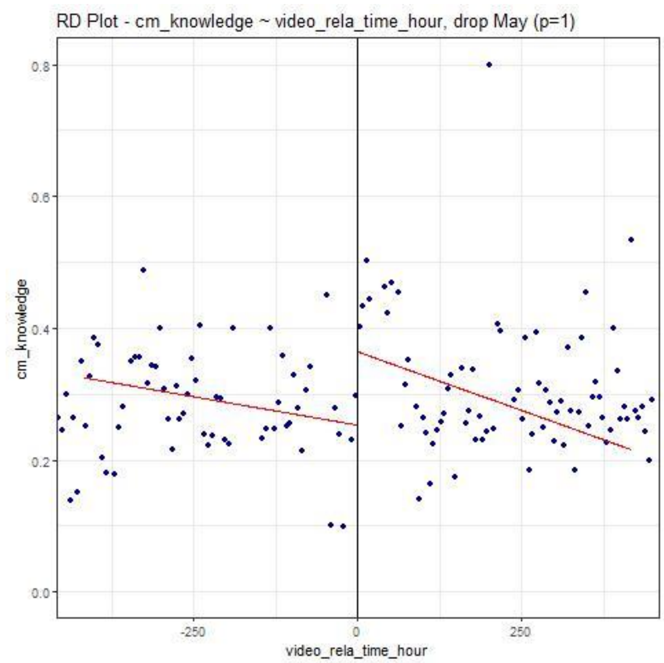


Figure A9. Regression discontinuity plot for the outcome of total knowledge construction (continuous variable) in Case 2

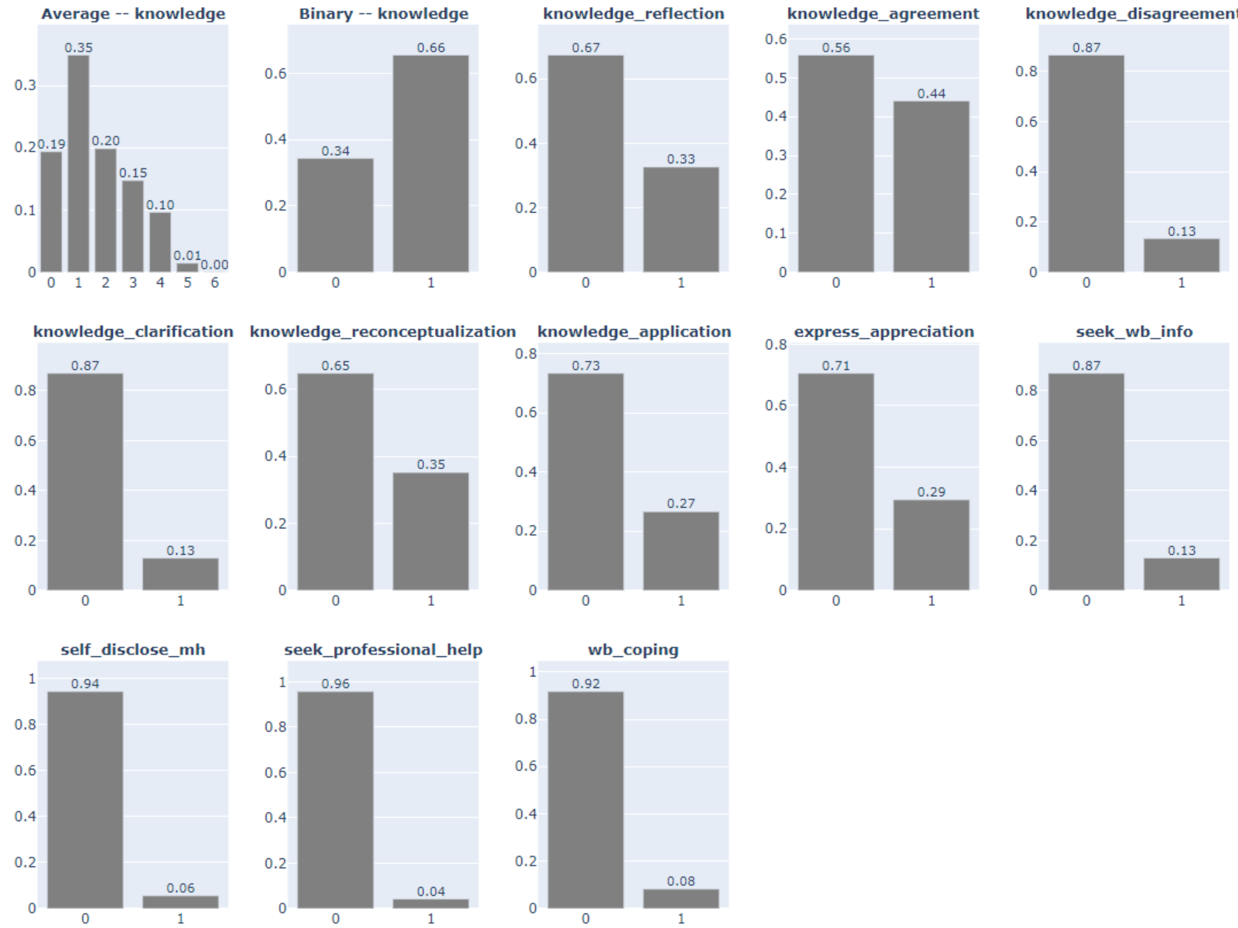
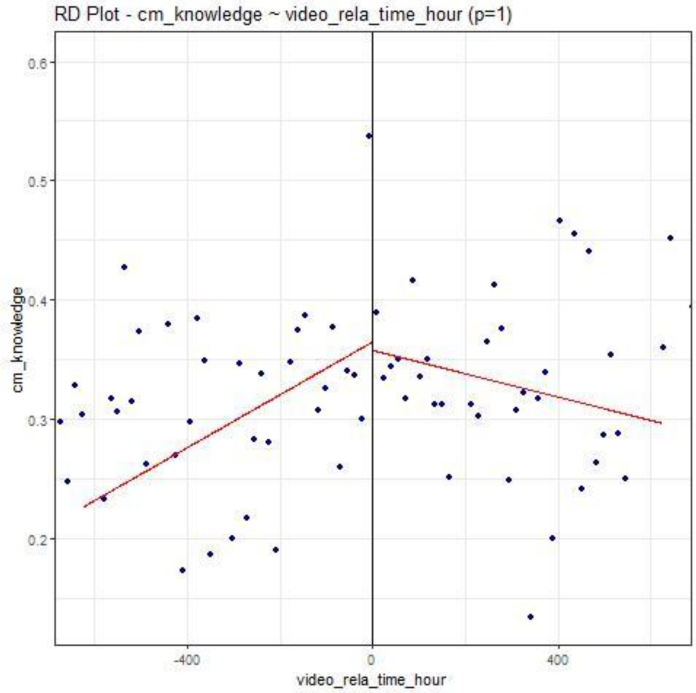


Figure A10. The distributions of the study outcomes among all TikTok comments included in Case 3.

(A) Keep May 2023



(B) Drop May 2023

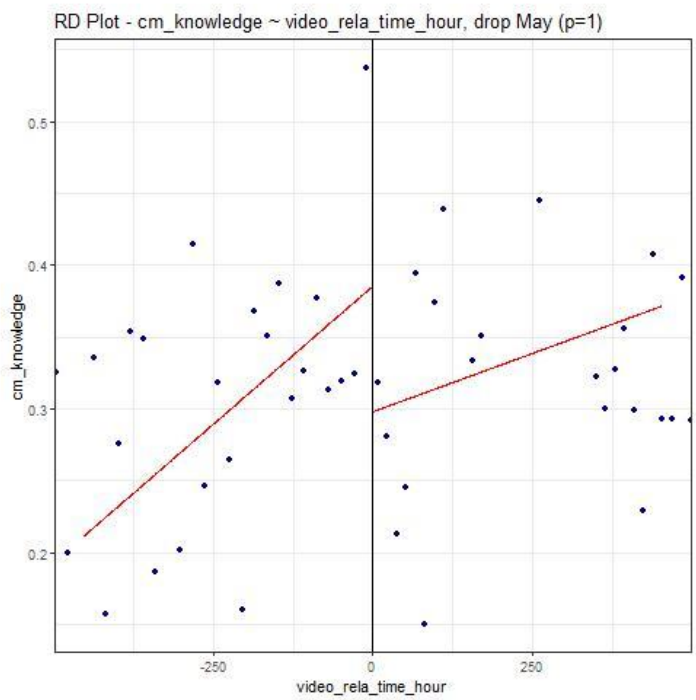


Figure A11. Regression discontinuity plot for the outcome of total knowledge construction (continuous variable) in Case 3

Section 1: Prompts applied in LLM-based text augmentation

```
messages=[
    {
        "role": "system",
        "content": "You are a helpful research assistant that rephrases text. "
    },
    {
        "role": "user",
        "content": "The texts you are going to rephrase are user comments on TikTok videos that are related to self-care, well-being, and mental health. \n"
                   "I will give you a sample, please give me 5 rephrased comments in the format of numbered bullet points.\n"
                   "Here is the comment:\n"
                   f"'{[comment_to_aug]}'"

    },
```