Creator Hearts: Investigating the Impact of Positive Signals from YouTube Creators in Shaping Comment Section Behavior

Frederick Choi
fc20@illinois.edu
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Charlotte Lambert
cjl8@illinois.edu
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Vinay Koshy
vkoshy2@illinois.edu
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Sowmya Pratipati
sowmyap2@illinois.edu
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Tue Do
tuedo2@illinois.edu
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Eshwar Chandrasekharan
eshwar@illinois.edu
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Abstract

Much of the research in online moderation focuses on punitive actions. However, emerging research has shown that positive reinforcement is effective at encouraging desirable behavior on online platforms. We extend this research by studying the "creator heart" feature on YouTube, quantifying their primary effects on comments that receive hearts and on videos where hearts have been given out by creators. Overall, creator hearts increased creator agency over feed presentation in YouTube comments sections, and also served as an incentive mechanism to drive user engagement. We find that creator hearts increased the visibility of comments, and increased the amount of positive engagement they received from other users. We also find that the presence of a creator-hearted comment soon after a video is published can incentivize viewers to comment, increasing the total engagement with the video over time. We discuss how creators can use hearts to shape behavior in their communities by highlighting, rewarding, and incentivizing desired behaviors from users. We discuss avenues for extending our study to understanding positive signals from moderators and curators on other platforms.

CCS Concepts

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

Keywords

positive reinforcement, incentives, desirable behavior, moderation

ACM Reference Format:
Frederick Choi, Charlotte Lambert, Vinay Koshy, Sowmya Pratipati, Tue Do, and Eshwar Chandrasekharan. 2025. Creator Hearts: Investigating the Impact of Positive Signals from YouTube Creators in Shaping Comment Section Behavior. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713521

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713521

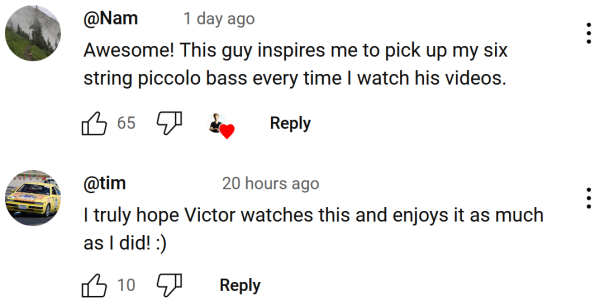


Figure 1: YouTube comments that have received a "heart" from a creator are displayed with a distinctive badge (top).

1 Introduction

Moderation is an essential part of maintaining a healthy and active online social platform. However, many of the techniques employed today focus on the moderator's role as a censor or disciplinarian, responsible for eliminating undesirable content and taking punitive actions [24, 47]. This approach to moderation, while important, leaves several issued unaddressed. Reactive techniques cannot undo harm already done [34, 53, 63], and the burden falls on moderators to quickly identify and remove offending content. On most platforms, effective and timely removal means that the offending content, along with all evidence of the moderator's work, vanish without a trace for most, if not all, users. As a result, these moderators, who are crucial to the success of online communities, ultimately receive little recognition or reward for their labor.

The purpose and practice of online moderation and platform governance goes well beyond just censoring content [26]. Desirable activity on a platform is not merely the absence of undesirable activity [10], and obfuscating or punishing negative behaviors does not necessarily help users understand how to participate in a desirable way. Moderators must take on different roles with many objectives beyond just taking punitive actions and removing undesirable content [55]. Jurgens et al. [34] highlight the approach of articulating affirmative capabilities and promoting desired behaviors instead of trying to articulate and enforce every prohibited

behavior. Along these lines, moderators should establish norms for desirable behavior, and create incentive structures to encourage users to voluntarily adhere to them [37].

In this paper, we focus on the moderator's role as a *curator*, responsible not only for removing undesirable content, but also for encouraging desirable behaviors. We study how moderators can use existing affordances and how platform designers can introduce new features to give moderators the tools they need to guide and reinforce user behavior.

1.1 Creating Incentive Structures to Guide User Behavior

Kiesler et al. [37] propose several techniques which moderators can leverage to scaffold desirable behavior within their communities. They highlight the importance of positive feedback mechanisms both for rewarding behavior and for helping others learn norms by example. Positive feedback can also be studied within the framework of positive reinforcement and the factors which mediate its effects [23, 44, 59–61]. In this framework, positive feedback can be considered a stimulus which, when delivered contingent to some user behavior, reinforces that behavior from the user.

In practice, the effects of positive feedback, particularly positive feedback from moderators, has been difficult to study as clear signals are scarce [38]. However, by studying formal feedback signals built into the platform (e.g., badges, stars), researchers have been able to investigate their impact on user behavior, the kinds of activities users engage in, and the content they produce. Anderson et al. [4] found that badges on StackOverflow were effective as incentives to steer behavior, reflected in how users would change their behavior to receive badges. Lambert et al. [39] found that Reddit users who received high scores and gold made more frequent and higher quality posts. Wang and Diakopoulos [62] saw a reinforcement effect on commenters who had been awarded a New York Times Editors' Pick, exhibiting an increase in both the quality and quantity of their future comments. They also observed a spill-over effect on other authors who engaged with the comment—the quality of their future comments increased slightly as well.

We extend this line of research to study "creator hearts" on YouTube. We will use the terms *creators* and *channels* interchangeably in this paper. Viewers participate in a community around a creator's content through the comments sections on their videos [48]. Creator hearts allow creators to participate in the community and endorse comments of their choosing by giving them a "heart" (illustrated in Figure 1). The unique clarity of creator hearts as signals of endorsement from a creator makes them ideal for studying the effects of positive feedback from users in distinct positions of authority within their community. We can examine the causal effects of creator hearts through a quasi-experimental study on YouTube using observational data. And whereas previous studies of positive feedback were limited in their study populations, studying creator hearts on YouTube allows us to use observational data from a diverse range of creator–audience communities on YouTube.

1.2 Research Questions

We investigate how creator hearts can play a role in creating incentive structures to encourage desired behaviors within YouTube comment sections. To this end, we explore the causal effects of creator hearts—specifically, if and how hearts affect the behavior of the audience community on YouTube—a problem that remains relatively unexplored. We aim to characterize how creator hearts impact user perceptions and incentives, and investigate if it is worth it for creators to experiment with hearts as a form of positive reinforcement. We explore how creator hearts affect users' decisions to engage with the comment, as well as the video it was posted on. Though hearts make comments more likely to be placed at the top of the comments list by design [40], we seek to quantify their actual impact on the visibility of comments as well as their ability to draw more attention to the comments that receive them. Since our focus was not on studying the specific norms of each creator-audience community, we do not analyze the content of the comments themselves. Instead, we study the activities that users choose to engage or not engage in as a result of receiving or observing a creator heart, which materialize into large scale behaviors that generalize across channels. Specifically, we ask the following:

RQ1. What effect does a creator heart have on how the community engages with the comment that receives it?
- (a) How does it affect:
  - (i) the position (rank) of the comment,
  - (ii) the number of likes the comment receives,
  - (iii) and the number of replies the comment receives?
- (b) How do these effects vary with the size of the community?

RQ2. What effect does the presence of creator-hearted comments have on the engagement with the video?
- (a) How does it affect the number of comments the video receives over time?
- (b) How do these effects vary with the timing of when the creator heart is given?

1.3 Summary of Contributions

1.3.1 Methods. We illustrate our data collection and analysis procedure in Figure 2. First, we conducted an initial scan of heart-giving activity across 11.8K channels on YouTube. Then, from a subset of 1K channels, we collected time-series data of comment-level and video-level outcomes for 16.5K videos and 2M comments, and observed the timing of 81.5K creator hearts. We then applied matching and Wilcoxon signed-rank analysis to measure the causal effect of a creator heart on comment-level outcomes (like count, rank, and number of replies) for 11K hearted comments, stratifying by channel sizes based on their subscriber counts. Finally, we measured the causal effect of the presence and timing of a creator heart on audience participation through comments using a series of regressions on datasets ranging in size from 508-682 videos each.

1.3.2 Findings. From our scan of 11.8K channels, we found hearts were given by creators with audiences of all sizes, ranging from 14 to 50M+ subscribers. However, we failed to find any hearts from 44.5% (5,233) of the channels we scanned. Though we cannot know the exact proportions, it is clear that many, though not all, creators on YouTube are regularly giving away hearts.

From our analysis of channels with at least 5K subscribers, we found that comments that received a heart appeared closer to the start of the comments section. Hearted comments also received an increased number of likes from the community. Aggregating

across channels of all sizes, we found that hearts given within the first 5 hours of the video's publishing had a significant effect on engagement with the video. We observed the greatest effect when a creator heart was given within the first hour, associated with a 22% increase in the mean number of comments after 12 hours, and a 27.3% increase after 24 hours.

1.3.3 Implications. Our findings show that creator hearts can be an effective means of highlighting and drawing the community's attention to comments that the creator wants to endorse. Our findings also show that creator hearts are able to incentivize engagement from the audience, and that the earlier a heart is given, the greater its ability to drive engagement. We discuss how creators can take advantage of these effects to regain some control over how their algorithmically-curated comments sections are presented to their audience and to better communicate norms by highlighting examples of desirable behavior. We also discuss creator hearts' ability to incentivize participation, the importance of hearts in creating visible traces of creator engagement within the community, and the implications for similar signals on other platforms in rewarding and incentivizing user behaviors. Though further research is needed that examines the actual content of the comments, we encourage creators to experiment with using hearts to highlight exemplary comments and to reinforce desirable behavior.

2 Background

In this section, we review prior work to situate our research. Specifically, we focus on prior work studying specific interface signals, the shift towards positive forms of feedback, and general research into the YouTube community.

2.1 Approaches to Online Moderation

An ever-growing body of research seeks to explain and evaluate the effectiveness of the various moderation strategies employed by online platforms. At the highest level of governance, platforms have been known to apply sanctions to entire communities by limiting access to them or banning them altogether. Studies of such interventions have shown their effectiveness in reducing undesirable behavior from those individuals involved and across the platform as a whole [17, 18]. Many platforms, including Facebook, Instagram, and YouTube, also hire commercial content moderators who manually review and remove user-generated content that can do serious harm to a platform. However, this work can also do serious harm to the moderator's mental well-being [47].

On platforms such as Reddit, Discord, and StackOverflow which are structured around communities, moderators are themselves usually members of the communities they moderate. These moderators are almost always unpaid volunteers, but are nonetheless vital to the success of online platforms, collectively contributing hundreds of hours worth of moderation work every day carrying out most of the moderation tasks on the platform [41]. Despite all their work, moderators often struggle to keep up as the scale of their communities and the activity within them grow. Researchers are continuing to explore the use and development of tools designed to help moderators manage their ever-increasing workload [16, 19, 32, 36].

Most of the current practices and research in online moderation, including those mentioned above, focus on moderation through censoring undesirable content. While necessary, such measures are insufficient for producing desirable outcomes. For one, since removing content cannot undo harm already done [34, 53, 63], it is important that content is dealt with as quickly as possible—a task that is becoming more difficult as platforms grow. But, on the other hand, platforms often desire that moderators work invisibly, removing offending content before anyone can see it, and leaving no trace their intervention [47].

This lack of visibility of moderator activity is a drawback of removal-focused moderation for several reasons. First, this lack of transparency can be problematic as it can mislead users to believe the norms they see online reflect norms in the real world. For example, this can cause a user to experience identity-based harm when the content online implies that stereotypes and harmful societal norms related to their identity are the norm offline as well [58]. Another issue, especially on platforms that rely on volunteer moderators (e.g., Reddit, Discord, StackOverflow), is that the behind-the-scenes efforts of moderators to monitor activity and investigate user-reported incidents go unseen by their communities. And so, these moderators, who do not receive any compensation from the platform, also end up getting little recognition from their communities for the effort they put in.

Content removals and punitive actions are only one aspect of moderating an online platform [26]. We focus this paper on a more holistic approach that incorporates promoting and encouraging desirable behavior from users on the platform through positive reinforcement and other incentive structures.

2.2 Positive Reinforcement Offline and Online

Positive reinforcement is a concept introduced in the field of psychology by B. F. Skinner [23, 59–61] in which a stimulus is introduced to increase the likelihood that a behavior will reoccur in the future. Specific methods for employing positive reinforcement have been studied in the context of education [5, 46], workplaces [7, 12], and parenting [6, 20], and have been found effective at encouraging desired behavior.

Miltenberger [44] lays out several factors that influence the efficacy of positive reinforcement. Immediacy is one such factor as the sooner the stimulus is administered following the desired behavior, the stronger the reinforcing effect. Contingency is another, as a stimulus that is present only following a particular behavior (i.e., contingent on that behavior) is more likely to reinforce that behavior. Establishing operations such as deprivation makes stimuli more effective as reinforcers, while abolishing operations such as satiation makes stimuli less effective. Magnitude can influence efficacy as more intense stimuli are generally more effective. Finally, the efficacy of reinforcers and the influence of the various factors will vary from person to person.

In the context of social media, the various signals such as likes and warnings can be seen as reinforcing and aversive stimuli, and studied in the context of both reinforcement and punishment. Several researchers have already measured their effects on user behavior and evaluated their efficacy at producing particular behavioral outcomes [4, 27, 30, 33, 45, 62]. Using creator hearts on YouTube as a case study, we aim to further investigate their role in behavior modification positive reinforcement. We aim to study the design of

such signals, their implementation in the interface, their relationship with the rest of the platform's design, and their potential as a positive form of intervention in the context of online moderation.

2.3 Signals, Algorithms, and Incentive Mechanisms in Social Media Interfaces

The design of a platform's interface creates incentive structures that influence user behavior. It is important to study the roles each of the design elements and affordances play in moderating those platforms [8]. Broad design interventions such as reputation systems [3] and gamification [28, 52] have been found to motivate users toward desired actions. More granular design interventions such as an interstitial in the case of quarantined subreddits have also proved effective in altering the course of user behavior [17].

It is also important to consider how a platform's design affects how users learn norms. Establishing clear and salient norms is a crucial step for moderating online communities [37]. Norms can be characterized as injunctive or descriptive, the former referring to stated norms and the latter referring to observed or practiced norms. Making injunctive norms more visible, such as by announcing rules within discussions [42], is shown to increase rule compliance and participation from newcomers. Making examples of desirable behavior more visible is also important as it influences the descriptive norms that users learn. However, the increasing reliance on algorithmic curation poses a barrier as it becomes harder to control what is shown to users. This is especially problematic when algorithmic curation promotes the wrong things (or hides the right things), as highlighting too many examples of negative behaviors can lead users to believe that those behaviors are the norm [37].

The signals a platform implements play an important role in how users learn norms. Formal feedback mechanisms are a way to clearly communicate norms by giving users clear signals about how their content is being received by moderators and the community [37]. Upvotes and likes are common mechanisms for users to give feedback to one another on platforms like Reddit, Facebook, and Twitter. Upvotes (i.e., high score) and gold have been found to positively influence user engagement and behavior on Reddit [39]. Badges, another example of a positive signal, have been shown to reinforce certain behaviors. The New York Times Picks badge was shown to positively correlate with the quality of exposed users' future posts [62]. In the context of Stack Overflow, certain badges can be used to motivate desired behaviors when used strategically [4].

It is important to distinguish feedback signals from regular users and moderators since the influence of these signals can depend on the sender's level of authority. Seering et al. [56] explored the imitation effects of Twitch users seeing positive and negative behaviors from different types of users: moderators, subscribers, turbo users, and regular users. They found that users with more authority were associated with larger imitation effects for positive behavior. This prior work establishes a precedent for authoritative figures having the potential to influence change through example-setting. However, the research does not account for example-setting or feedback from the Twitch creator.

In our work, we investigate whether the trends found in prior work (e.g., [39, 56, 62]) are present in the YouTube context when considering creator hearts as a formal feedback mechanism. Unlike upvotes and gold on Reddit, which are anonymous signals of approval from the broader community (since its not apparent who provided the feedback), creator hearts are clear signals of endorsement from YouTube creators. Creator hearts are contrasted from badges on StackOverflow or the NYT Editor's Pick in that there are no clearly defined criteria for how to earn them, and it remains unclear to users what actions are required to obtain a creator heart. However, we can consider creator hearts as a general signal of approval from an authoritative figure, even if the exact intent or behavior they are approving of is not clear or varies across creators.

Although moderators are typically the figures of authority that enforce community norms on many platforms, creators uniquely establish the norms and values through their content on YouTube. Thus, we believe that creators have the potential for strong influence on their communities and focus our analyses on creator hearts.

2.4 YouTube Comments and Communities

Since its creation, YouTube has allowed viewers to engage with videos by leaving comments visible to the public in a video's comment section. They also serve to bridge creators and viewers to form creator–audience communities [48]. In 2016, YouTube introduced a feature that allows creators to send positive signals to their community by giving "hearts" to comments on their videos [40]. Comments that have received a heart are displayed with a distinctive badge, shown in Figure 1. They are a formal feedback mechanism for the creator to signal their endorsement. Contrasting with other formal feedback mechanisms on the platform such as likes and dislikes, creator hearts provide a way for users to receive a signal of approval directly from the single most authoritative figure in their community. A heart badge is easy to identify, it is visible to all users, and the source of the heart is unambiguous since only creators can give hearts, enabling us to explore its use and effects in practice.

There have been several studies involving YouTube comments, including explorations into what content gets rated highly [57], what knowledge is shared through commenting [22], and why people comment [54]. However, none of these studies touch on how the creator impacts those elements of commenting. Additionally, Rotman et al. [48] sought to understand the sense of community on YouTube. Despite the fact that YouTube itself does not have explicit communities, the authors found that YouTube users largely felt like they were part of specific communities. The structure of the platform and user interactions do not necessarily define the community structure. Instead, YouTube communities seem to be centered largely around content. Considering this emphasis on content-based communities, we seek to quantify the role of creator feedback within creator-audience communities on YouTube. We expand on prior work [13] exploring quantitative impacts of creator hearts by investigating the impact on the comments that receive them, as well as investigating the impact of the timing of when creator hearts are given.

3 Data Collection

To study the effect of creator hearts on comment-level outcomes (RQ1) and video-level outcomes (RQ2), we construct a dataset of hearted comments from a diverse sample of channels across YouTube. In this section, we describe our data collection process.

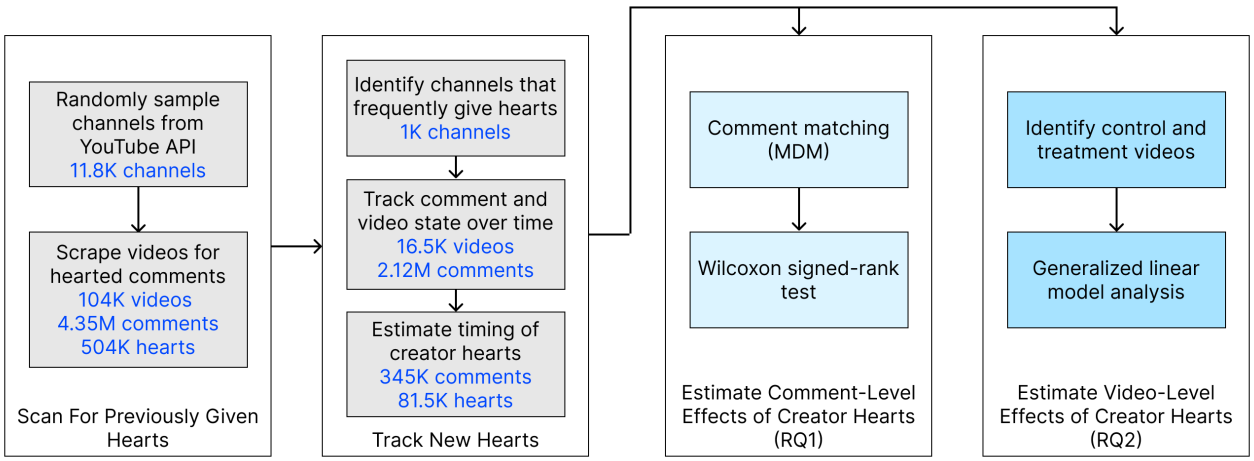


Figure 2: Flowchart depicting the data collection and analyses, with statistics of the data collected at each stage.

This involves generating a stream of randomly sampled YouTube channels, scanning channels for hearts previously given, and tracking channels for new hearts.

We report overview statistics of the data collected in Table 1. In Section 3.4, we provide descriptive statistics of the data and broadly characterize which creators give out hearts and how many they do.

3.1 Constructing a Random Sample of Channels

In the first part of our data collection process, we crawled the Youtube Data API to collect a random sample of videos over time, from which we created a stream of randomly sampled channels to be used for the following stages. Using a script, we queried the YouTube API search endpoint to collect up to 50 of the most recently published videos once every 30 minutes. This initial sampling occurred continuously between February and April 2023.

One limitation of our technique is that our sampling pool is biased towards channels that post more frequently, as their videos would have had more chances to appear in one of our queries. Another limitation stems from the YouTube API itself. Although the API allows us to query for a fixed number of videos published within a specified time frame, it is not clear how the YouTube API selects which videos to return out of the thousands that match. Although the documentation indicates that the ordering of results can be specified to return the most recently published videos first, observation of the returned videos contradicts this. We acknowledge that this introduces unknown biases into our sample. However, since the videos span dozens of topics and the subscriber counts and average comment counts of the channels in our sample span several orders of magnitude (see Section 3.4), we believe our results are still generalizable to a broad range of YouTube creators.

3.2 Scanning for Previously Given Hearts

In the second part of our data collection process, we scanned for hearted comments on videos from sampled channels. We used this dataset to get a preliminary understanding of the overall presence of heart-giving behavior and trends across creators on YouTube.

Because the YouTube API does not provide an easy way to identify creator-hearted comments. we instead relied on scraping YouTube's interface. We used Selenium in Python to automate this process, but identifying hearted comments remained slow. Loading the page for a video would take several seconds, even with images disabled, and each group of 10-20 comments would take an additional second or so to load in. To avoid undue strain on the YouTube servers, as well as on our own network bandwidth, we limited our scraper to operate sequentially on only one video at a time.

With this limitation in mind, we narrowed our search for hearted comments to the top 100 comments (sorted by "Top comments") on the 10 most recent videos for each channel. We performed this procedure starting in March and through April of 2023 while we continued to discover new channels. In total, we collected data for 11.8K channels, and identified 501K hearted comments out of the 4.31M comments we had scraped from 102K videos.

3.3 Tracking New Hearts

To study the effects of a creator heart on comment-level (RQ1) and video-level outcomes (RQ2), we needed to track the comments on each video over time and observe *when* comments receive hearts from creators. We used the YouTube API to collect snapshots of comment- and video-level outcomes, and we used our scraper to track which comments had received a heart or not. To increase the likelihood of observing a heart being given, we selected a subset of 1K channels of the 11.8K channels we had previously observed based on how frequently and consistently they had given hearts previously. We then subscribed to the RSS feeds for these channels so that we could start monitoring new videos as quickly as possible. We decided to cycle through up to 20 videos at a time so that consecutive snapshots would be no longer than 5-10 minutes apart on average. After monitoring a video for 4 days, we would

Table 1: Overview statistics of creator hearts data after each stage of data collection. The initial identification of channels was carried out over February-April of 2023, and the tracking was carried out over August-December of 2023.

	Totals after scanning for previously given hearts	Totals after tracking creator hearts
Channels	11,750	1,002
Videos	102,450	16,514
All Comments	4,564,895	2,124,053
Scraped Comments	4,308,795	344,937
Hearts	500,620	81,503

retire the video, and its spot in the cycle would be replaced when we received notification that a new video was published. This process ran between August and December of 2023. A total of 2.12M comments from 16.5K videos were tracked via the API, of which 345K comments were tracked via the scraper. This yielded a total of 82K hearted comments.

3.4 Descriptive Statistics

From the initial scan of 11.8K channels, we found hearts from 55.5% (6517) of channels. Some of the channels giving out hearts had as few as 14 subscribers, where as others had more than 50M. These numbers demonstrate that giving out hearts is widely practiced by channels of all sizes, though this behavior is hardly universal.

The distribution of the average number of hearted comments we found across each channels' videos (Figure 3) shows that creators who do give out hearts tend to give out several per video. In fact, from channels that give out hearts, we observed an average of 6.75 and a median of 2.56 hearted comments per video.

We found the most likes from creators with between 10K and 1M subscribers and who receive 100-1K comments on average per video (Figure 4). One explanation comes from the fact that there cannot be more hearts than comments, thus creators with few subscribers and who receive few comments would be limited in how many hearts they can give. On the other extreme, channels with many subscribers and who receive many comments may have given hearts to comments that we did not reach while scraping. That being said, the typical viewer of those videos are unlikely to see those hearts either, and may get the impression that larger channels are less engaged with their community.

4 RQ1: Effects on Engagement with Hearted Comment

After data collection, we address RQ1, drawing upon the potential outcomes framework [49] to examine the causal effect of a creator heart on comment-level outcomes. We employ a matching-based approach to estimate the causal effects of the treatment (i.e., creator heart) compared to the *counterfactual*—what would have happened

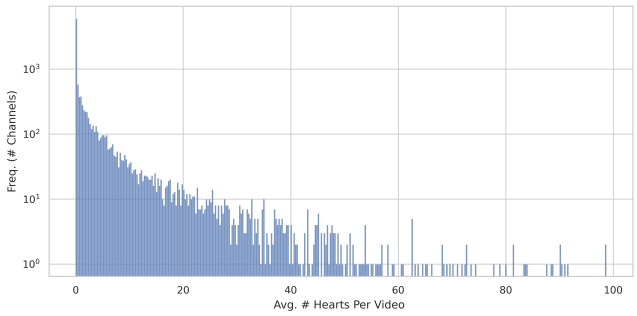


Figure 3: Distribution (across channels) of the average number of creator-hearted comments (of the top 100 comments) per video. We observed at least one heart from 55.5% (6,517) of the channels in our sample, and an average 6.75 hearts per video from these channels.

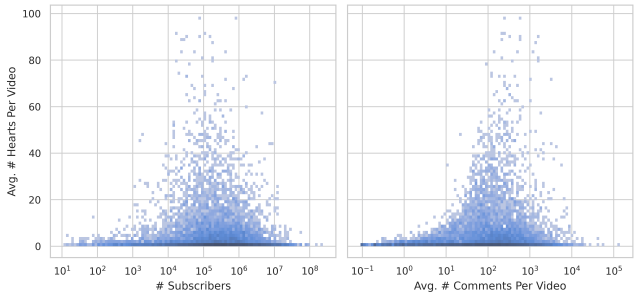


Figure 4: Distributions of the number of hearts observed from each channel, averaged across their videos.

if the treatment was not administered—by comparing hearted comments (i.e., treated group) against a similar set of non-hearted comments (i.e., control group). This section will first describe the use of Mahalanobis Distance Matching (MDM) [50] to match up hearted comments with non-hearted comments that have already received similar amounts of community engagement and have similar potential for future engagement. Then, we detail the results of performing a Wilcoxon signed-rank test to determine what effect creator hearts have on engagement and visibility outcomes.

4.1 Matching Comments for Causal Inference

In order to explore whether hearted comments have different outcomes, we first identify a set of control comments and treated comments. The treatment group consists of all comments in our dataset that received a creator heart. Specifically, the treatment group contains the snapshots of treated comments roughly when they were hearted by the creator. The control group contains all snapshots of all comments in the dataset that did not receive a heart from the creator at any point. Since the control group is a collection of snapshots, most unique comments appear multiple times in the control group at different points in their lifespan. On average, the control group contains roughly 250 snapshots of a comment.

Table 2: This table reports two measures of match quality for our treatment-control matches. First, we report the standardized mean differences (SMDs) between each of our covariates between the treatment and control groups. Then we report the results of a Mann-Whitney U-test on each covariate between the treatment and control groups.

Covariate	SMD	Mann-Whitney U-Test	
		U	p-value
Age	0.012	62104846	0.238
Rank	0.011	62155772	0.282
Like Count	<0.001	62589306	0.817
Num. Replies	0.002	62739207	0.867
Video Age	0.005	62372322	0.531
Subscriber Count	<0.001	62563231	0.817

To perform matching between our set of treated comments and control comments, we use Mahalanobis Distance Matching (MDM) [50], which has been used in prior social computing research for causal inference [18]. Specifically, we choose pairs of treated and control comments which minimize the Mahalanobis distance based on six covariates:

(1) **Comment age:** seconds since the comment was published
(2) **Rank:** rank of the comment within the comment section (lower rank comments are closer to the top of the feed)
(3) **Like count:** number of likes on the comment (not including creator hearts)
(4) **Video age:** seconds between video publication and comment publication
(5) **Number of replies:** the number of replies on the comment
(6) **Number of subscribers:** the number of subscribers to the channel that posted the video

Intuitively, the goal of the matching procedure is to pair treated comments with untreated comments that would have undergone the same trajectory as their treated match, had they been hearted by the creator. To do this, we try to ensure that matched comments:

(1) Have received similar amounts of community engagement prior to the intervention,
(2) Have similar "potential" for future engagement.

To achieve (1), we match on pre-treatment number of replies, like count, and rank. By controlling for pre-treatment engagement, an unobserved factor can only introduce spurious correlation between the treatment and the outcomes measured if it influences both the likelihood of receiving a creator heart and post-treatment engagement, while also having a significantly weaker influence on pre-treatment engagement. We discuss limitations to this approach in Section 6.5.3, but this helps mitigate bias from qualitative features whose influences are difficult to control for directly but which do not change over time, such as characteristics of the creator, comment author, and the content of the comment itself.

To achieve (2), we match on video and comment age, as well as the number of subscribers the channel associated with the video has. Age variables are useful proxies for potential engagement since newer comments, and comments posted early in the video's lifespan have likely been seen by a smaller proportion of a video's audience. Similarly subscriber counts are a rough measure of the size of a YouTube channel's community, giving us some insight into the number of people who might potentially come across a comment.

We log-transformed each of the covariates apart from the reply count to encourage the distributions to be more normal. Using these six identified covariates, we use nearest neighbor matching with replacement based on the measured Mahalanobis distance to match each treated comment with one control comment, resulting in 11,196 matches. Out of these matches, there were 5,265 unique control comments. This is partially because we sampled with replacement and otherwise because our control sample contained multiple snapshots of the same comment at different points in time. That means that the same comment may be a good match for one treated comment in one snapshot, and a good match for a different treated comment at a later snapshot.

We note that our matches are not representative of all channels on YouTube, largely because the treated comments dictate the matches. Since the smallest channel in our dataset in which the creator hearted at least one comment has more than 5K subscribers, the matching process discarded any comments posted in significantly smaller channels. Thus, we are unable to make claims about how creator hearts affect engagement in small channels with fewer than 5K subscribers. In addition, since we sampled from channels which give hearts more frequently, it is unclear how or if our claims would extend to creators who rarely or never give hearts.

To measure the quality of these matches, we use two methods. First, we compute the standardized mean difference (SMD) for our six covariates across the two groups. Prior work considers the control and treatment groups to be appropriately balanced if all covariates have SMDs less than 0.25 [35, 51]. As shown in Table 2, all of our covariates satisfy this condition, thus we consider our matches to be well-balanced. In Appendix A we provide the full distributions of standardized differences between treatment and control pairs for each covariate.

Second, we perform Mann-Whitney U-tests on each of the covariates used for matching to assess the quality of the matching, similar to the approach taken by Chandrasekharan et al. [18]. Table 2 reports the U statistics and p-values for each of these tests, showing that, for each test, we cannot reject the null hypothesis. This implies that there is no significant difference between these six covariates when comparing treatment and control comments.

4.2 Results

With pairs of treated and control comments identified, we then compared the pairs on three different outcomes. To do this comparison, we first identified the age of each comment at the time of its last snapshot in our dataset, which we refer to as its *death*. Within a pair, we look at the snapshots of each comment at the minimum death age between the two comments. This allows us to compare the outcomes of each matched pair after roughly the same amount of time has passed since the age at the time they were matched.

We focus on three outcomes measured at those final snapshots: number of likes, rank, and number of replies. In Appendix B we check the normality of the differences for each outcome between the treatment and control groups to determine whether a t-test is appropriate. We find that the differences are not normal when

Table 3: T-statistic and p-value for each of the Wilcoxon signed-rank tests we performed (one for each outcome and channel-size pair). Small channels have up to 100K subscribers, medium channels have between 100K and 1M subscribers, and large channels have more than 1M subscribers. The bottom three rows indicate the percentage of matches in which the corresponding outcome for a treated comment is greater than, equal to, or less than its matched control comment, with significant results in bold.

Outcome	# Likes			Rank			# Replies		
	small	medium	large	small	medium	large	small	medium	large
T	334780	2287101	184370	1080190	8452634	697309	37254	234862	7279
p-value	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00
% Greater	**40.48 %**	**42.28** %	**34.58** %	41.61 %	40.87 %	39.34 %	6.31 %	6.04 %	3.63 %
% Equal	39.97 %	40.51 %	48.93 %	10.68 %	10.68 %	10.83 %	79.49 %	82.36 %	89.12 %
% Less	19.55 %	17.21 %	16.49 %	47.71 %	**48.44** %	**49.83** %	**14.21** %	**11.6** %	**7.25** %

considering the raw data, nor after log-transforming the data. Thus, we proceed with a Wilcoxon signed-rank test.

As shown in Table 3, we conducted two-sided Wilcoxon signed-rank tests for each outcome and spearated by channel size. Specifically, we consider *small* channels to be those with less than 100K subscribers ($n = 2379$), medium channels to have between 100K and 1M subscribers ($n = 6804$), and large channels to have more than 1M subscribers ($n = 2013$). With these splits, we carried out 9 statistical tests, which generally reveal that there were significant differences between each of the outcomes for treated comments when compared to their paired control comment. This indicates an observed effect of creator hearts on the trajectory of a comment. To determine the direction of this effect, we calculate the percentage of pairs in which the observed outcome for the treated comment was greater than, equal to, or less than the outcome for its control match. These percentages are reported in Table 3.

We observe that treated comments tend to end up with more likes than their control counterparts, with the effect being slightly less apparent in large channels. Evidently, receiving a creator heart encourages other community-members to like the comment. We also see that the final rank of a hearted comment is lower (i.e., in a higher position of the comments section) than that of its matched control comment nearly 50% of the time. YouTube itself says that giving a heart to a comment may give the comment a featured position in a preview of the comments section [1], but we are able to show that receiving a creator heart boosts the exposure of a comment in the comment section itself. However, this effect is not significant in smaller channels at a threshold of $p < 0.05$, likely due to the smaller number of comments their videos typically receive. Finally, we see that hearted comments more often end up with fewer replies than those that did not receive the same feedback, an effect that is stronger in small and medium channels. This is somewhat counter-intuitive to the expectation that creator hearts encourage engagement, however we note that the most common outcome is that both treated and control comments in a pair have equal numbers of replies at their death. We can attribute this to the sparse-ness of replies in our dataset.

5 RQ2: Effects on Engagement with Video

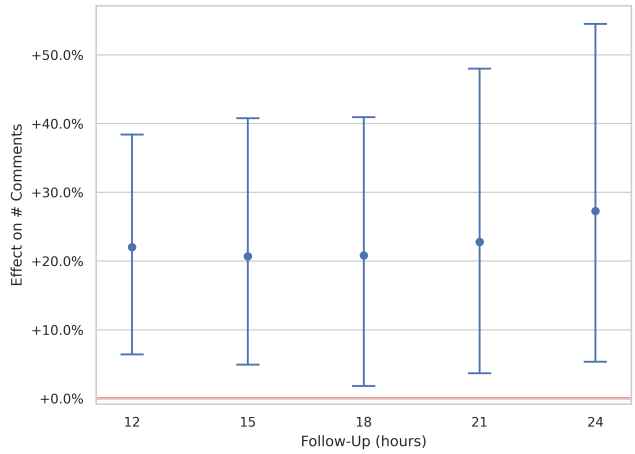
In this section, we examine the causal effect of a creator heart on video-level outcomes. First, we describe the outcomes we are measuring, the potential confounds, and the general approach we take in modeling the effect of a creator heart. Then, we describe our data preparation and how we identify suitable treatment and control videos to compare outcomes. This is followed by a formal definition of the model we used to analyze the effect of a creator heart. We then present our results from measuring the effect of creator hearts on video-level outcomes using Bayesian techniques.

5.1 Method

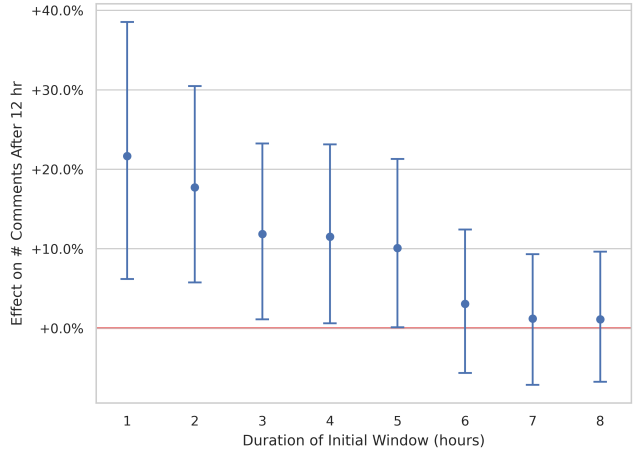
For video-level effects of a creator's heart, the outcome we are interested in is the audience's participation through comments, which we measure as the total number of comments on a video. We want to see if the total number of comments a video receives over its lifetime is affected by whether or not the creator gave a heart to any of the comments within the first few hours immediately after the video is published. We begin by observing the comment section for w_{init} hours starting from when the video is published, and check for the presence of a hearted comment by the end of that initial observation window. Since we cannot wait an infinitely long time to count the "true" total, we instead measure the number of comments at the end of a follow-up window of w_{foll} hours.

The number of comments in the follow-up window varies greatly from video to video, but is highly correlated with the number of comments that were posted in the initial observation window. This is because both counts are influenced by many of the same factors, such as the composition of the channel's typical audience, the performance of the video in YouTube's recommender system, and qualities of the video itself which may encourage more or less participation in the comments. We use the number of comments observed in the initial window as a proxy to control for these confounds that can affect the final comment count. The number of subscribers is another useful covariate to control for varying audience sizes when comparing final counts across different channels. With a regression, we can then see if any of the remaining variation can be explained by the presence of a hearted comment in the initial window.

5.1.1 Data Preparation. To make our study more robust, we repeat our analysis with varying initial and follow-up window durations. This has the added benefit of letting us see how the effect of the creator heart changes over time. We prepared a total of 12 datasets: 5 with the initial window fixed and varying follow-up windows ($(w_{\text{init}}, w_{\text{foll}}) \in \{1\} \times \{12, 15, 18, 21, 24\}$), and an additional 7



(a) 1 hour initial window, varying follow-ups.



(b) Follow-up after 12 hours, varying initial windows.

Figure 5: Effect of the presence of a creator heart within the first few hours of the video publishing. Effects are interpreted as a percent change to the expected number of comments after the follow-up. The markers indicate the posterior means, while the lines indicate the 95% high-density intervals (see Table 4 in Appendix C for exact values). (a) As the follow-up increases from 12 to 24 hours after the initial window of one hour, the high-density intervals remain greater than zero, indicating that the effect remains significant over time. (b) Meanwhile, as the initial window increases from 1 to 8 hours, the downward trend indicates that the effect of the presence of a creator heart decreases the later it is given.

cases with the follow-up window fixed and varying initial windows ($(w_{\text{init}}, w_{\text{foll}}) \in \{2, 3, ..., 8\} \times \{12\}$).

Within each dataset, videos that had at least one hearted comment within the initial window were included as a "treated" sample. Meanwhile, videos where all comments posted within the initial window were confirmed (by scraping) to not have been hearted in the initial window were included as a "control" sample. Videos where there were unscraped comments posted within the initial window are not included as we cannot reliably determine whether a creator heart was present. We also excluded videos for which we did not have a snapshot of the comment count after the follow-up window since we cannot measure the outcomes. As a result, each dataset consisted of slightly different sets of videos and contained between 508 and 682 videos each.

5.1.2 Model Definition. Now, we formalize the model that we used to measure the effect of creator hearts on comment counts within each dataset. Let $c_{\text{init}}[i]$ and $c_{\text{foll}}[i]$ be the number of comments on video i at the end of the initial and follow-up windows respectively. Let $I_{\heartsuit}[i]$ be an indicator variable that is equal to one if we observed at least one hearted comment within the initial window of video i and zero otherwise. Let $s[i]$ be the number of subscribers of the channel that posted video i. We utilize a generalized linear model with coefficients $\boldsymbol{\beta} = \{\beta_0, \beta_{\heartsuit}, \beta_c, \beta_s\}$ to measure the effect hearted comments ($\beta_{\heartsuit}$) on the number of comments in the follow-up ($c_{\text{foll}}[i]$). In terms of random variables, we model $\log c_{\text{foll}}[i]$ as being drawn from the following distribution:

$$\log c_{\text{foll}}[i] \sim \text{Normal}(\beta_0 + \beta_{\heartsuit} I_{\heartsuit}[i] + \beta_c \log c_{\text{init}}[i] + \beta_s \log s[i], \sigma^2) \quad (1)$$

Here, σ^2 is interpreted as the remaining unexplained variance in $\log c_{\text{foll}}[i]$ after regression over the covariates. We regress on the logarithms of comment counts and subscriber counts to encourage normality and to avoid excess leverage from extreme values. This also allows the model to better capture the non-linear relationship between the initial and follow-up counts. The coefficient $\beta_{\heartsuit}$ can be interpreted as follows: assuming all other factors are identical, a video with a creator heart in the initial window has the effect of increasing the expected number of comments after the follow-up by a factor of $\exp(\beta_{\heartsuit})$. We report effect sizes as the percent increase in the expected number of comments: $100 \times (\exp(\beta_{\heartsuit}) - 1)$.

5.1.3 Bayesian Inference. We use Bayesian techniques to implement the model above [43]. Instead of estimating a single maximum likelihood value for each parameter, this approach produces distributions for each of the parameters, with likelihoods assigned to a range of possible values for each parameter. In our results, we report several key statistics from the resulting posterior distributions: the mean, standard deviation, and the 95% credible interval.

When we report the 95% credible interval for the effect of a creator heart ($\beta_{\heartsuit}$), it should be interpreted to mean that there is a 95% probability that the "true" effect of a creator's heart lies in that interval [29]. The effect of a creator heart should be interpreted as significant (at a significance level of 0.05) when the value of 0 (corresponding to no effect) falls outside the 95% credible interval [29].

The Bayesian approach also requires that we place prior distributions on the parameters, which represent our prior beliefs on the values of each parameter. We selected our priors to minimize our assumptions, placing flat priors on the coefficients $\boldsymbol{\beta}$, and Jeffreys' prior [31] on the variance σ^2: $p(\sigma^2) \propto \sigma^{-2}$.

We ran our regressions using the NUTS sampler in PyMC [2] to sample posterior distributions for each parameter. Posterior distributions for $\beta_{\heartsuit}$ are summarized in Table 4, and posterior distributions for all parameters are summarized in Table 5.

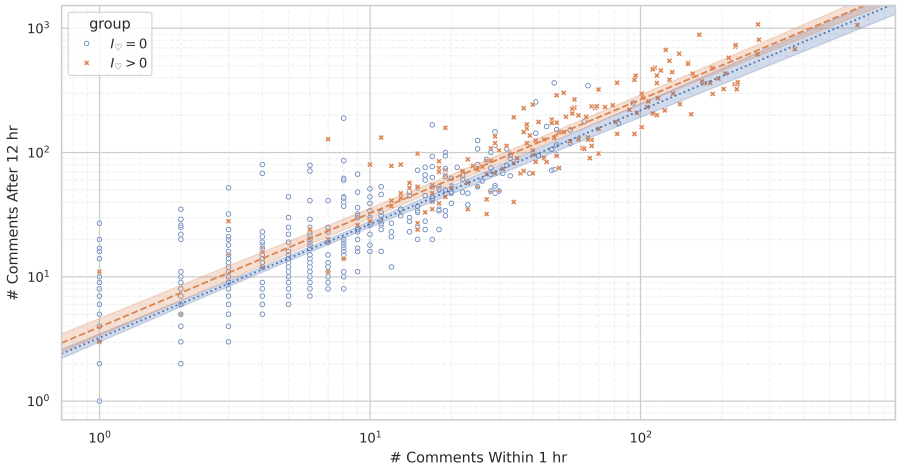


Figure 6: Plot of comment counts from the initial and follow-up windows ($w_{\text{init}} = 1$, $w_{\text{foll}} = 12$) and posterior means for treatment (dashed line) and control (dotted line) videos. The control group (n=494) consists of videos where no comments were hearted by the creator within the first hour after the video was published ($I_\heartsuit = 0$), and the treatment group (n=179) consists of videos where at least one comment was given a heart within the first hour ($I_\heartsuit = 1$). The difference in posterior means illustrates a mean 22.1% increase (HDI 6.3%-38.7%) in the number of comments after 13 hours (1 hour initial window + 12 hour follow-up window) from the presence of a creator heart within the first hour.

5.2 Results

Starting with the windows $w_{\text{init}} = 1$ and $w_{\text{foll}} = 12$, the results show that the presence of a creator heart within the first hour is associated with a mean 22.0% increase (CI 6.4%-38.4%) in the number of comments 12 hours later (Figure 6). Results from holding the initial window constant while varying the follow-up window from 15 to 24 hours indicates that the effect remains significant over time, with a mean of up to 27.3% increase (CI 5.3%-54.5%) in the total number of comments after 24 hours (Figure 5a). Together, these results suggest that the presence of a creator heart within the first hour after a video is published has a significant, lasting effect, on the total number of comments a video receives.

To investigate the dependence on when the heart was given, we ran similar regressions with the initial observation period ranging from 1 hour in duration to 8 hours, and following up 12 hours after the end of the initial observation window. The results are illustrated in Figure 5b. The downward trend in effect size as the initial observation window increases in duration indicates that the effect of the presence of a creator heart quickly falls off the later it is given, and we were unable to measure a significant effect from an initial window of 6 hours or more. One explanation for the decreased effect size is that the activity in the comments section as a whole might start to level off, diminishing the possible impact a heart can have. Another explanation is that hearts that are given earlier have had more time to be seen by a greater portion of the audience and influence their decision to leave a comment. In either case, the evidence indicates that a creator heart has the greatest impact when it is given shortly after a video is published.

However, due limitations of the scraper, it was easier to confirm that at least one hearted comment exists (i.e., belongs in the treatment group) than it was confirm that none of the comments had received a heart (i.e., belongs in the control group). Since we dropped videos we could not reliably assign a group to, there is a correlation between initial comment count and the group that has been assigned: The treatment videos in our dataset skew towards having a higher initial comment count and the control videos skew towards a lower initial count. It is thus possible that the increase in comments in after the follow-up window actually reflect a more complex relationship between the initial and follow-up counts than our model can predict. Future studies can refine the model by accounting for how many views the video had during each window. This controls for how many users had the opportunity to comment, and arrives at a more direct measure of how creator hearts influenced a user's choice to comment or not.

6 Discussion

Based on our findings and prior research, we discuss the ways creator hearts may be useful for shaping user behavior in creator-audience communities, as well as broader implications for understanding the role of positive feedback in shaping user behavior. We discuss properties of the design of creator hearts which might be used to inform the design of similar signals on other platforms. We also discuss our findings of how creator hearts increase curator agency over the presentation of feeds and how they can be used to incentivize participation, with implications for the role creator hearts can play in shaping user behavior and how to use them most effectively. Finally, we discuss the use of creator hearts and positive feedback signals for behavior change within the positive reinforcement framework, and highlight some limitations of our study and avenues for future work.

6.1 Improving signals from curators on community-oriented platforms

We preface our discussion of specific findings by highlighting four key properties of the design of creator hearts which generalize to the design of signals on other platforms, and which align with prior work on mechanisms for guiding user behavior.

(1) *Creator hearts are a clear and visible signal of positive feedback.* Prior work has shown that such formal feedback mechanisms (e.g., likes, creator hearts) are important tools for shaping user behavior, and can be more effective than informal feedback mechanisms (e.g., free-form replies) [37].
(2) *Creator hearts provide a clear indication that the positive feedback came from a central figure in the community.* It is difficult to fake the sender of this signal since only the creator can give creator hearts, making it a reliable cue for identifying positive feedback from the creator with their distinct level of authority within the community [21]. Having a reliable cue of the level of authority of the sender of the positive feedback is important since users with more authority have a greater influence on other users' behaviors [56].
(3) *Creator hearts give creators influence over the ranking and presentation of comments.* As we discuss further in Section 6.2, influence over these factors provide creators the ability to improve the visibility of "desired" or endorsed content, making it more likely for users to see and engage with them. Drawing attention to desired behavior is an important part of a holistic approach to shaping user behavior by helping users learn the appropriate behavioral norms [37, 62].
(4) *Creator hearts are accessible to moderators.* They require no monetary cost to use, and require minimal effort from the creator to give out, especially in comparison to other moderator actions such as replying. This accessibility is important, as moderators have cited a lack of affordances and resources as barriers to adopting more positive approaches to shaping behavior in their communities [38].

Of course, as we will discuss in later sections, the use of creator hearts for behavior change has its limitations, and its design has room for improvement. But these properties of the design of creator hearts points us to ways the design of signals from moderators on other community-oriented platforms can be refined to expand the range of tools which moderators can use to shape their communities. For example, on Reddit, moderators have expressed their desire to shape the behavior in their community through positive actions [38], but many have cited a lack of features and resources as barriers to this approach. The closest features are gilds and awards, but both cost real money, making it difficult for moderators to use regularly. Both can also be given by regular users with no distinction of who the endorsement was given by [39], diluting the moderators' influence through such features. In contrast, creator hearts are exclusive to creators, and it is clear to commenters that they are given by the creator. Creator hearts are also free to use, require less effort than replying, and have fewer restrictions than pinning, thus lowering the barrier to regular use. In fact, the barrier is low enough that we found that many creators were already using hearts, even if they are not necessarily using them intentionally to shape their communities. And our findings have shown how such features increase creator agency over how feeds are presented for their community. Thus, platforms like Reddit (and others like Discord and Twitch) may be able to refine the design of signals from moderators to increase moderator agency over feed presentation and to improve their accessibility to moderators, thus expanding the range of tools moderators can use to shape their communities.

6.2 Increasing Creator Agency over the Presentation of Feeds

Our findings show that creator hearts increase creator agency by giving creators a way to influence the presentation of the comments feed for their audience. We found that receiving creator hearts moves comments closer to the top of the comment section, increasing their visibility. We also found that a creator heart draws more attention to the comment, evidenced by the increase in the number of likes received from the community. This gives creators the ability to curate their comment sections by directing users' attention to particular comments they want to endorse.

This ability to influence the presentation of the feed and direct user attention to particular comments is important as a tool for shaping user behavior. Creators may be able to leverage these effects to highlight and direct their audiences' attention to particular comments which exemplify desired behavior. Unlike removals, bans, and other punitive actions which hide or discourage undesirable behavior, this approach influences user behavior by shaping the norms they learn. Establishing clear and salient norms is essential for moderating online communities, and making examples of normative behavior visible to everyone in the community is one way to encourage normative behavior [37]. And, in fact, prior work has shown that highlighting comments is an effective tool for shaping community behavior by increasing the visibility of and drawing attention to exemplary comments which demonstrate desired behavior [62]. It is clear from our findings that creator hearts give creators influence over critical factors in setting norms through examples, i.e., visibility and attention. For researchers, this makes YouTube a fruitful setting to try to replicate prior work (Wang and Diakopoulos [62]) by tracking users and the content of their comments to see if the same effect can be observed within the diverse, creator-centric communities of YouTube.

6.2.1 Limitations of creator hearts and tensions between curator agency and algorithmic curation. It is important to acknowledge that while hearts increase creators' agency over how their comments sections are presented to their audiences, our findings also show that the influence of creator hearts is still limited, underscoring a tension between the algorithmic curation of comments sections and the content that creators may want to highlight. For example, even if the YouTube algorithm continues to boost creator hearted comments in general, it is possible that the same algorithm might actively deprioritize a creator-hearted comment due to other attributes of the comment (e.g., few user likes).

This is a limitation to positive feedback interventions in a social media landscape that is dominated by algorithmically-sorted feeds [14, 15], with implications for platform designers who must weigh curator agency over algorithmic curation. On YouTube, one way forward is to afford creators more fine-grained control over how feeds are sorted within their channels. For example, work by

Bernstein et al. [11] explores the possibility that higher-level values (e.g. "pro-democracy", "anti-partisanship") could be embedded into feed-sorting algorithms. This can be complemented with ways for creators to "spot-fix" or override the sorting of individual comments to highlight exemplary behavior, such as by increasing the number of "pinned" comments. Giving creators this level of control over the presentation of feeds could lead to more powerful norm-setting tools relative to the softer-touch approach of creator hearts, and these tools can be extended to curators on other platforms as well.

6.3 Incentivizing Participation with Hearts

We initially hypothesized that creator hearts would also incentivize users to participate more in the comments section. We found this to be the case—after adjusting for a number of confounders, we estimate that videos where creator likes are deployed will see a 20% increase in the number of comments observed after a 12-hour follow-up period. This finding echoes results from a previous study on The New York Times which found that receiving an Editor's Pick temporarily increased the recipients commenting rate on future articles [62]. Our work provides further insight, as our findings suggest that this increase in participation applies to those who witness a heart being given, not just the one who received it.

We also found that that the earlier a creator gives a heart, the greater the impact on engagement with the video. This provides further evidence that creator hearts incentivize witnesses to participate: The earlier a heart is given, the more potential commenters would have witnessed the creator heart having been given, and thus we would see a greater effect on the number of comments. Combined, our findings provide evidence that creator hearts may be able to play a role in incentivizing participation from users. However, as we will discuss in section 6.5.1, further research is warranted to examine the role the they may play in incentivizing specific behaviors beyond the decision to participate or not.

6.3.1 Uncovering mechanisms of behavior change through creator hearts. Our work points to a clear direction for future research: The precise mechanisms behind the aforementioned effects are unclear, but understanding the mechanisms is essential to predicting the conditions under which positive moderation interventions will be effective for changing user behavior.

For example, if the results are largely explained by the indirect effects of creator hearts and the resulting attention from the community, then it is plausible that similar feedback mechanisms would be effective on many different platforms. However, if the results are instead largely explained by the creator's own social capital and the relationship between creators and their audience, then we can reasonably assume positive feedback interventions would be similarly effective on a platform like Twitch, but less effective on a platform like Reddit where community moderators are not always known by users. It is also possible that different mechanisms dominate in different communities within the same platform. The differential effects of positive feedback through various mechanisms warrant further study, the results of which will be instrumental to the design and integration of more effective incentive structures into the interfaces of social platforms.

6.4 Towards Positive Reinforcement

Creator hearts and this study are a special case of the more broadly applicable theory of positive reinforcement. In this section, we will discuss how the framework of positive reinforcement can be applied to study behavior change through positive feedback, and how creator hearts on YouTube are well-suited for studying the factors that mediate the effectiveness of positive reinforcement.

6.4.1 Studying positive reinforcement in social media through creator hearts. As we have discussed in Section 2, a holistic approach to moderation goes beyond considering only punitive actions, and it is important for moderators to have ways to shape behavior in their communities through positive actions. Positive reinforcement is a valuable framework for investigating how moderators can influence behavior using positive feedback signals.

According to Miltenberger [44], positive reinforcement occurs when a positive stimulus, called a *reinforcer*, is introduced following a specific behavior, having the effect of strengthening said behavior. In this work, we studied creator hearts as a reinforcer to strengthen the behavior of users participating through comments. A creator heart is one instance of a positive feedback signal, and this idea of positive reinforcement through positive feedback signals is broadly applicable to many platforms and online communities. For example, Lambert et al. [39] found that positive feedback from the community through gilds and upvotes on Reddit increased the frequency of posting and the quality of their posts. Wang and Diakopoulos [62] had similar findings with Editor's Picks in The New York Times comment section when commenters received positive feedback from an authoritative figure.

A few characteristics of creator hearts make it well-suited for studying how moderators can shape the behavior of their community through positive feedback signal. First, not many platforms have as clear a signal of endorsement as creator hearts, which enable creators to provide positive feedback in response to user behavior, distinct from the feedback which a typical user can provide. This distinction is important as it gives creators control over how positive feedback is administered, and it helps us answer the more general question of how moderators should administer positive feedback for it to be most effective. Second, creator hearts also enable researchers to study the effects of positive feedback from authoritative figures across diverse communities. Understanding how effects vary across communities can lead to more robust insights. With this study, we have laid groundwork for future research toward understanding the application of the positive reinforcement framework in online settings through the study of creator hearts.

6.5 Limitations and Future Work

Next, we summarize the limitations of our work and highlight directions for future work to address these limitations. These future directions can benefit both platform designers in designing effective positive feedback signals as well as moderators in formulating effective strategies for the application of positive feedback.

6.5.1 Examining what exact behaviors are being reinforced. Our findings show how creator hearts can be used to increase participation within a video's comment section. However, a limitation of our study is that we only measured the impact of positive feedback

on participation in a broad sense. Future research should examine the content of comments to investigate the specific behaviors and modes of participation which are being reinforced [25].

6.5.2 Effects on bystanders vs. direct recipients. Our findings highlight the strength of *vicarious reinforcement* in influencing user behavior, as we had seen increased participation from those who witnessed the creator heart being given. Vicarious reinforcement is related to positive reinforcement, with the main difference being in vicarious reinforcement, behaviors are reinforced in those who witness (i.e., bystanders) the reinforcement being administered to others [9]. Future work should examine the effects of directly receiving a creator heart by tracking the receivers' behavior across a creator's later videos, and study potential differences in effects of positive feedback on direct recipients vs. bystanders.

6.5.3 Controlling for algorithmic effects and comment content. Our findings suggest a causal relationship between a creator heart being given and the engagement a comment or video receives. However, as discussed in Section 4.1, an external factor might bias our results if it influences the probability that a comment receives a heart as well as post-treatment engagement outcomes, but has a weaker influence on pre-treatment outcomes. One way this might happen is if an external factor varies over time. One example of such a factor is the influence YouTube's video recommendation algorithm. The algorithm may interact with the content of comments in influencing the likelihood of receiving a creator heart as well as the amount of engagement from the community. There may be similar confounding interactions with the topics of the videos as well. Future work that controls for the effects of the algorithm, video topics, and the content of the comments themselves would be necessary to strengthen the causal analysis conducted in this work.

6.5.4 Examining other factors that influence the strength of positive reinforcement. Our findings demonstrate how both the source of the positive feedback and the timing with which it was given (i.e., *immediacy*) are significant factors which influence its effectiveness as a positive reinforcement tool. Future research is required to examine additional factors of positive reinforcement such as *satiation*, i.e., how the effects of positive feedback on online platforms diminish with repeated application. In addition, investigating how characteristics of the recipient, such as their tenure on the platform or their own social capital, mediate the effects of positive reinforcement can produce valuable insight into the underlying mechanisms which make positive feedback, like creator hearts, effective reinforcers.

7 Conclusion

Moderation involves more than just punitive actions; as we move towards supporting moderators in their broader role as curators, there is a need to study more ways to positively reinforce desirable behavior. In this paper, we quantified the causal effects of *creator hearts* on comments that receive hearts, and on videos where hearts have been given. We found that creator hearts increase the visibility of comments and increase the number of likes a comment receives from the community. We found an increase in the number of comments on videos where a creator heart was present shortly after it is published. We conclude that giving hearts is an easy option for creators who want to highlight exemplary behavior and incentivize participation. Our work as well as future research on similar signals of formal feedback in other platforms will be instrumental to the design of interfaces that better support moderators and their active roles in shaping online communities.

References

[1] 2024. Post & interact with comments - Computer - YouTube Help. https://support.google.com/youtube/answer/6000964

[2] Oriol Abril-Pla, Virgile Andreani, Colin Carroll, Larry Dong, Christopher J. Fonnesbeck, Maxim Kochurov, Ravin Kumar, Junpeng Lao, Christian C. Luhmann, Osvaldo A. Martin, Michael Osthege, Ricardo Vieira, Thomas Wiecki, and Robert Zinkov. 2023. PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science* 9 (Sept. 2023), e1516. doi:10.7717/peerj-cs.1516

[3] B. Thomas Adler and Luca De Alfaro. 2007. A content-driven reputation system for the wikipedia. In *Proceedings of the 16th international conference on World Wide Web*. ACM, Banff Alberta Canada, 261–270. doi:10.1145/1242572.1242608

[4] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. Steering user behavior with badges. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, Rio de Janeiro Brazil, 95–106. doi:10.1145/2488388.2488398

[5] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2014. Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web*. 687–698.

[6] Shaljan Areepattamannil. 2010. Parenting practices, parenting style, and children's school achievement. *Psychological Studies* 55 (2010), 283–289. Publisher: Springer.

[7] Michael B Armstrong and Richard N Landers. 2018. Gamification of employee training and development. *International Journal of Training and Development* 22, 2 (2018), 162–169. Publisher: Wiley Online Library.

[8] Tanvi Bajpai, Drshika Asher, Anwesa Goswami, and Eshwar Chandrasekharan. 2022. Harmonizing the Cacophony with MIC: An Affordance-aware Framework for Platform Moderation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–22.

[9] Albert Bandura. 1977. Social learning theory. *Englewood Cliffs* (1977). https://www.academia.edu/download/90990222/Kelompok_4_Materi_Gestalt_Bandura.pdf

[10] Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the Web Conference 2021*. 1134–1145.

[11] Michael Bernstein, Angèle Christin, Jeffrey Hancock, Tatsunori Hashimoto, Chenyan Jia, Michelle Lam, Nicole Meister, Nathaniel Persily, Tiziano Piccardi, Martin Saveski, et al. 2023. Embedding Societal Values into Social Media Algorithms. *Journal of Online Trust and Safety* 2, 1 (2023).

[12] Christiane Bradler, Robert Dur, Susanne Neckermann, and Arjan Non. 2016. Employee recognition and performance: A field experiment. *Management Science* 62, 11 (2016), 3085–3099. Publisher: INFORMS.

[13] Unji Byun, Moonkyoung Jang, and Hyunmi Baek. 2023. The effect of YouTube comment interaction on video engagement: focusing on interactivity centralization and creators' interactivity. *Online Information Review* 47, 6 (2023), 1083–1097.

[14] Jackie Chan, Aditi Atreyasa, Stevie Chancellor, and Eshwar Chandrasekharan. 2022. Community Resilience: Quantifying the Disruptive Effects of Sudden Spikes in Activity within Online Communities. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–8.

[15] Jackie Chan, Charlotte Lambert, Frederick Choi, Stevie Chancellor, and Eshwar Chandrasekharan. 2024. Understanding Community Resilience: Quantifying the Effects of Sudden Popularity via Algorithmic Curation. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 227–240.

[16] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.

[17] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–26.

[18] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 1. 1–22. doi:10.1145/3134666

[19] Frederick Choi, Tanvi Bajpai, Sowmya Pratipati, and Eshwar Chandrasekharan. 2023. ConvEx: A Visual Conversation Exploration System for Discord Moderators. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–30.

[20] Don Dinkmeyer and Gary D McKay. 1989. *The parent's handbook: Systematic training for effective parenting.* ERIC.

[21] Judith Donath. 2007. Signals in Social Supernets. *Journal of Computer-Mediated Communication* 13, 1 (Oct. 2007), 231–251. doi:10.1111/j.1083-6101.2007.00394.x

[22] Ilana Dubovi and Iris Tabak. 2020. An empirical analysis of knowledge co-construction in YouTube comments. *Computers & Education* 156 (Oct. 2020), 103939. doi:10.1016/j.compedu.2020.103939

[23] C. B. Ferster and B. F. Skinner. 1957. *Schedules of reinforcement.* Appleton-Century-Crofts, East Norwalk. doi:10.1037/10627-000

[24] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

[25] Agam Goyal, Charlotte Lambert, and Eshwar Chandrasekharan. 2024. Uncovering the Internet's Hidden Values: An Empirical Study of Desirable Behavior Using Highly-Upvoted Content on Reddit. *arXiv preprint arXiv:2410.13036* (2024).

[26] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015).

[27] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration.* ACM, Mountain View California, 163–172. doi:10.1145/2038558.2038585

[28] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does gamification work?–a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences.* IEEE, 3025–3034.

[29] Luiz Hespanhol, Caio Sain Vallio, Lucíola Menezes Costa, and Bruno T Saragiotto. 2019. Understanding and interpreting confidence and credible intervals around effect estimates. *Brazilian journal of physical therapy* 23, 4 (2019), 290–301.

[30] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized Social Signals: Computationally-Derived Social Signals from Account Histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* ACM, Honolulu HI USA, 1–12. doi:10.1145/3313831.3376383

[31] Harold Jeffreys. 1946. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 186, 1007 (1946), 453–461.

[32] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.

[33] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. 1–27. doi:10.1145/3359252

[34] David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* 3658–3666.

[35] Emre Kiciman, Scott Counts, and Melissa Gasser. 2018. Using Longitudinal Social Media Analysis to Understand the Effects of Early College Alcohol Use. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (June 2018). doi:10.1609/icwsm.v12i1.15012 Number: 1.

[36] Charles Kiene, Jialun Aaron Jiang, and Benjamin Mako Hill. 2019. Technological frames and user innovation: Exploring technological change in community moderation teams. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[37] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design* 1 (2012), 4–2.

[38] Charlotte Lambert, Frederick Choi, and Eshwar Chandrasekharan. 2024. " Positive reinforcement helps breed positive behavior": Moderator Perspectives on Encouraging Desirable Behavior. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–33.

[39] Charlotte Lambert, Koustuv Saha, and Eshwar Chandrasekharan. 2025. Does Positive Reinforcement Work?: A Quasi-Experimental Study of the Effects of Positive Feedback on Reddit. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, Yokohoma, Japan. doi:10.1145/3706598.3713830

[40] Courtney Lessard. 2016. New tools to shape conversations in your comments section. https://blog.youtube/news-and-events/new-tools-to-shape-conversations-in/

[41] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. Measuring the monetary value of online volunteer work. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 596–606.

[42] J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.

[43] Richard McElreath. 2018. *Statistical rethinking: A Bayesian course with examples in R and Stan.* Chapman and Hall/CRC.

[44] Raymond G Miltenberger. 2016. *Behavior modification: Principles and procedures.* Cengage Learning.

[45] Maria Papoutsoglou, Georgia M Kapitsaki, and Lefteris Angelis. 2020. Modeling the effect of the badges gamification mechanism on personality traits of Stack Overflow users. *Simulation Modelling Practice and Theory* 105 (2020), 102157.

[46] Chris Perryer, Nicole Amanda Celestine, Brenda Scott-Ladd, and Catherine Leighton. 2016. Enhancing workplace motivation through gamification: Transferrable lessons from pedagogy. *The International Journal of Management Education* 14, 3 (2016), 327–335. Publisher: Elsevier.

[47] Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. (2016).

[48] Dana Rotman, Jennifer Golbeck, and Jennifer Preece. 2009. The community is where the rapport is – on sense and structure in the youtube community. In *Proceedings of the fourth international conference on Communities and technologies (C&T '09).* Association for Computing Machinery, New York, NY, USA, 41–50. doi:10.1145/1556460.1556467

[49] Donald B Rubin. 2005. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *J. Amer. Statist. Assoc.* 100, 469 (March 2005), 322–331. doi:10.1198/016214504000001880 Publisher: ASA Website _eprint: https://doi.org/10.1198/016214504000001880.

[50] Donald B Rubin and Elizabeth A Stuart. 2006. Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. (2006).

[51] Koustuv Saha and Amit Sharma. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 590–601. doi:10.1609/icwsm.v14i1.7326

[52] Michael Sailer, Jan Ulrich Hense, Sarah Katharina Mayr, and Heinz Mandl. 2017. How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in human behavior* 69 (2017), 371–380. Publisher: Elsevier.

[53] Sarita Schoenebeck, Cliff Lampe, and Penny Trieu. 2023. Online Harassment: Assessing Harms and Remedies. *Social Media+ Society* 9, 1 (2023), 20563051231157297.

[54] Peter Schultes, Verena Dorner, and Franz Lehner. 2013. Leave a comment! An in-depth analysis of user comments on YouTube. (2013).

[55] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2022. Metaphors in moderation. *New Media & Society* 24, 3 (2022), 621–640.

[56] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.* ACM, Portland Oregon USA, 111–125. doi:10.1145/2998181.2998277

[57] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. 2010. How useful are your comments? analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web (WWW '10).* Association for Computing Machinery, New York, NY, USA, 891–900. doi:10.1145/1772690.1772781

[58] Ellen Simpson and Bryan Semaan. 2021. For You, or For" You"? Everyday LGBTQ+ Encounters with TikTok. *Proceedings of the ACM on human-computer interaction* 4, CSCW3 (2021), 1–34.

[59] B. F. Skinner. 1963. Operant behavior. *American Psychologist* 18, 8 (1963), 503–515. doi:10.1037/h0045185 Place: US Publisher: American Psychological Association.

[60] Burrhus Frederic Skinner. 1965. *Science and human behavior.* Number 92904. Simon and Schuster.

[61] Burrhus Frederic Skinner. 1989. *Recent issues in the analysis of behavior.* Prentice Hall.

[62] Yixue Wang and Nicholas Diakopoulos. 2021. Highlighting High-quality Content as a Moderation Strategy: The Role of *New York Times* Picks in Comment Quality and Engagement. In *ACM Transactions on Social Computing*, Vol. 4. 1–24. doi:10.1145/3484245

[63] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, support, safety, retribution, transformation: A restorative justice approach to understanding adolescents' needs for addressing online harm. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–15.

A Full Distributions of Matching Covariates

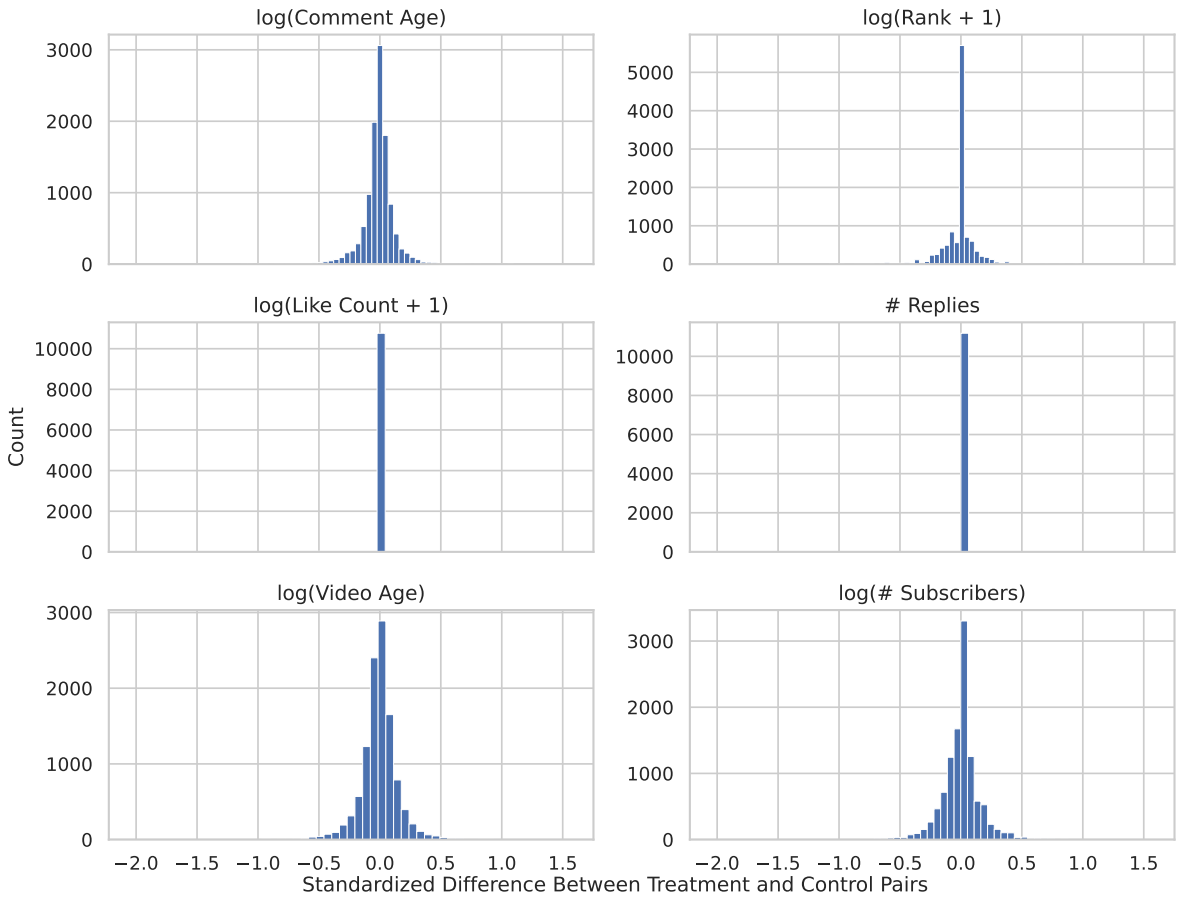


Figure 7: Standardized differences between matched pairs of treatment and control units for each covariate. All plots are roughly centered around zero and relatively concentrated, indicating good match quality.

B Probability Plots to Check Normality

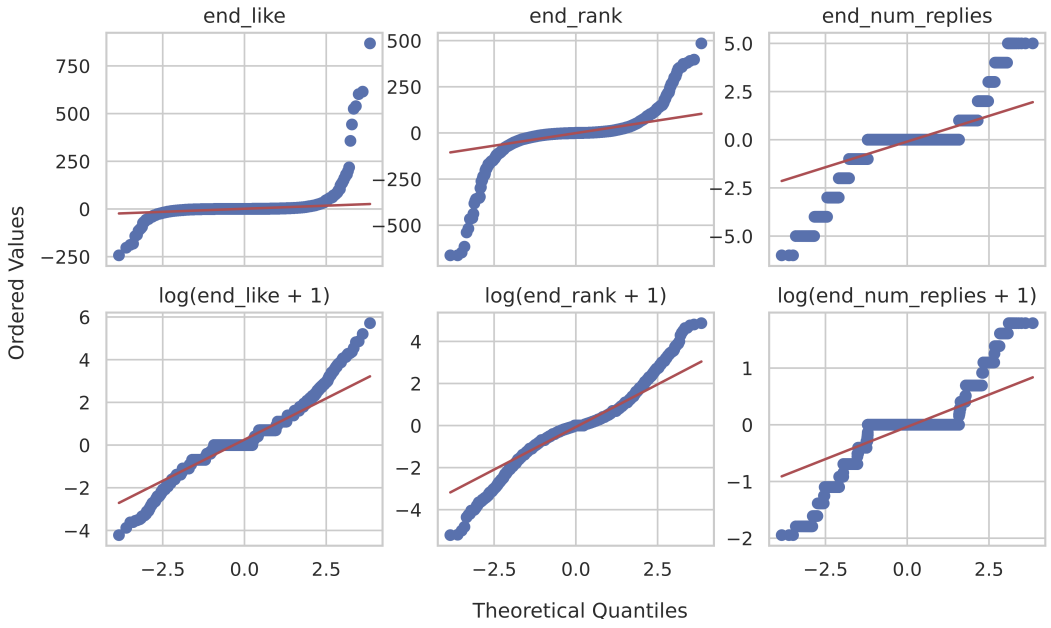


Figure 8: Probability plots of the differences between treatment and control comments for each of our three outcomes against the quantiles of a normal distribution. Best-fit lines are shown in red, demonstrating the non-normality of the data.

C Posterior Distributions of Regressions on Comment Counts

Table 4: Posterior distributions of $\beta_\heartsuit$ for varying initial observation windows and follow-ups. In parenthesis are the effect sizes interpretated as a percent change to the expected number of comments after the follow-up period, computed as $100 \times (\exp \beta_\heartsuit - 1)$. Significant results are in bold. Results are considered significant when 0 (no effect) falls outside the 95% credible interval. Results where 0 falls outside the 99% credible interval are indicated with an additional *.

Window Durations (w_{init}, w_{foll})	Sample Size (Total, $I_\heartsuit = 0$, $I_\heartsuit = 1$)	Mean	Std. Dev.	95% CI
1, 12	672, 493, 179	**0.199 (+22.0%)***	0.067	[0.062, 0.325] ([+6.4%, +38.4%])
1, 15	613, 470, 143	**0.188 (+20.7%)**	0.075	[0.048, 0.342] ([+4.9%, +40.8%])
1, 18	562, 442, 120	**0.189 (+20.8%)**	0.084	[0.018, 0.343] ([+1.8%, +40.9%])
1, 21	535, 427, 108	**0.205 (+22.8%)**	0.091	[0.036, 0.392] ([+3.7%, +48.0%])
1, 24	508, 412, 96	**0.241 (+27.3%)**	0.099	[0.052, 0.435] ([+5.3%, +54.5%])
2, 12	682, 483, 199	**0.163 (+17.7%)***	0.054	[0.056, 0.266] ([+5.8%, +30.5%])
3, 12	672, 478, 194	**0.112 (+11.9%)**	0.051	[0.011, 0.209] ([+1.1%, +23.2%])
4, 12	668, 475, 193	**0.109 (+11.5%)**	0.051	[0.006, 0.208] ([+0.6%, +23.1%])
5, 12	656, 469, 187	**0.096 (+10.1%)**	0.049	[0.001, 0.193] ([+0.1%, +21.3%])
6, 12	641, 464, 177	0.030 (+3.0%)	0.045	[-0.058, 0.117] ([-5.6%, +12.4%])
7, 12	631, 461, 170	0.012 (+1.2%)	0.041	[-0.074, 0.089] ([-7.1%, +9.3%])
8, 12	623, 459, 164	0.011 (+1.1%)	0.041	[-0.070, 0.092] ([-6.8%, +9.6%])

Table 5: **Posterior distributions of the parameters of the model defined in eq. (1) for varying durations of initial and follow-up windows (measured in hours).**

Window Durations (w_{init}, w_{foll})	Sample Size (Total, $I_\heartsuit = 0$, $I_\heartsuit = 1$)	Parameter	Mean	Std. Dev.	95% CI
1, 12	672, 493, 179	β_0	1.749	0.187	[1.375, 2.113]
		$\beta_\heartsuit$	0.199	0.067	[0.062, 0.325]
		β_c	0.916	0.020	[0.874, 0.954]
		β_s	-0.045	0.015	[-0.075, -0.017]
		σ	0.576	0.015	[0.545, 0.606]
1, 15	613, 470, 143	β_0	1.755	0.200	[1.369, 2.148]
		$\beta_\heartsuit$	0.188	0.075	[0.048, 0.342]
		β_c	0.911	0.022	[0.872, 0.954]
		β_s	-0.042	0.016	[-0.072, -0.012]
		σ	0.598	0.017	[0.566, 0.635]
1, 18	562, 442, 120	β_0	1.858	0.216	[1.435, 2.266]
		$\beta_\heartsuit$	0.189	0.084	[0.018, 0.343]
		β_c	0.908	0.023	[0.861, 0.954]
		β_s	-0.046	0.017	[-0.080, -0.015]
		σ	0.613	0.018	[0.577, 0.648]
1, 21	535, 427, 108	β_0	1.977	0.225	[1.547, 2.425]
		$\beta_\heartsuit$	0.205	0.091	[0.036, 0.392]
		β_c	0.904	0.025	[0.855, 0.953]
		β_s	-0.053	0.018	[-0.088, -0.019]
		σ	0.628	0.019	[0.591, 0.666]
1, 24	508, 412, 96	β_0	2.073	0.248	[1.564, 2.517]
		$\beta_\heartsuit$	0.241	0.099	[0.052, 0.435]
		β_c	0.901	0.026	[0.850, 0.952]
		β_s	-0.057	0.019	[-0.093, -0.019]
		σ	0.651	0.021	[0.611, 0.693]
2, 12	682, 483, 199	β_0	1.072	0.148	[0.772, 1.356]
		$\beta_\heartsuit$	0.163	0.054	[0.056, 0.266]
		β_c	0.924	0.015	[0.895, 0.954]
		β_s	-0.017	0.011	[-0.040, 0.004]
		σ	0.480	0.013	[0.456, 0.507]
3, 12	672, 478, 194	β_0	0.697	0.136	[0.434, 0.963]
		$\beta_\heartsuit$	0.112	0.051	[0.011, 0.209]
		β_c	0.942	0.014	[0.916, 0.969]
		β_s	-0.003	0.010	[-0.023, 0.018]
		σ	0.426	0.012	[0.404, 0.451]
4, 12	668, 475, 193	β_0	0.688	0.137	[0.421, 0.954]
		$\beta_\heartsuit$	0.109	0.051	[0.006, 0.208]
		β_c	0.936	0.013	[0.910, 0.962]
		β_s	-0.008	0.011	[-0.028, 0.013]
		σ	0.425	0.012	[0.402, 0.449]
5, 12	656, 469, 187	β_0	0.611	0.133	[0.368, 0.889]
		$\beta_\heartsuit$	0.096	0.049	[0.001, 0.193]
		β_c	0.933	0.013	[0.907, 0.957]
		β_s	-0.006	0.010	[-0.027, 0.013]

Continued on next page

Table 5: **Posterior distributions of the parameters of the model defined in eq. (1) for varying durations of initial and follow-up windows (measured in hours).** (Continued)

		σ	0.405	0.011	[0.382, 0.426]
6, 12	641, 464, 177	β_0	0.519	0.121	[0.274, 0.741]
		$\beta_\heartsuit$	0.030	0.045	[-0.058, 0.117]
		β_c	0.960	0.012	[0.937, 0.983]
		β_s	-0.007	0.009	[-0.025, 0.011]
		σ	0.372	0.010	[0.352, 0.392]
7, 12	631, 461, 170	β_0	0.460	0.111	[0.234, 0.672]
		$\beta_\heartsuit$	0.012	0.041	[-0.074, 0.089]
		β_c	0.969	0.011	[0.949, 0.991]
		β_s	-0.007	0.009	[-0.023, 0.010]
		σ	0.337	0.010	[0.317, 0.354]
8, 12	623, 459, 164	β_0	0.487	0.113	[0.251, 0.698]
		$\beta_\heartsuit$	0.011	0.041	[-0.070, 0.092]
		β_c	0.966	0.011	[0.946, 0.988]
		β_s	-0.010	0.009	[-0.026, 0.009]
		σ	0.333	0.010	[0.314, 0.352]